# How Does Mind Matter?

## Solving the Content Causation Problem

Gerard O'Brien

The primary purpose of this paper is to develop a solution to one version of the problem of mental causation. The version under examination is the content causation problem: that of explaining how the specifically representational properties of mental phenomena can be causally efficacious of behaviour. I contend that the apparent insolubility of the content causation problem is a legacy of the dyadic conception of representation, which has conditioned philosophical intuitions, but provides little guidance about the relational character of mental content. I argue that a triadic conception of representation yields a more illuminating account of mental content and, in so doing, reveals a candidate solution to the content causation problem. This solution requires the rehabilitation of an approach to mental content determination that is unpopular in contemporary philosophy. But this approach, I conclude, seems mandatory if we are to explain why mental content matters.

**Keywords**
Content determination | Mental causation | Mental representation | Resemblance

Author

Gerard O'Brien
gerard.obrien@adelaide.edu.au
University of Adelaide
Adelaide, Australia

Commentator

Anne-Kathrin Koch
Anne-Kathrin.Koch@gmx.de
Johannes Gutenberg-Universität
Mainz, Germany



## 1 Introduction: The content causation problem

Philosophy delights in those aspects of the world that initially seem obvious and natural, but which on reflection turn out to be deeply mysterious. The mental causation of behaviour is one such phenomenon. Nothing could be more obvious than that our minds matter—that our beliefs and desires, and our perceptions and thoughts ultimately have a causal impact on our behaviour. And yet it has proved notoriously difficult to explain just how this could be the case.

The problem of mental causation has morphed and fragmented over the years. In its original guise, it was the problem of how a non-physical mental substance or property could causally interact with the physical brain. The obvious solution to this version of the problem was to adopt a thorough-going materialism of some kind, with the consequence that mental phenomena are identified with properties of the brain from which they inherit their causal efficacy.

With the advent of functionalism in the later years of the last century, this "obvious" solution ran into difficulties. If mental phenomena are multiply-realizable, as the orthodox

construal of this metaphysical position seems to imply, then mental properties can't be identified with properties of the brain after all; and since the latter do all the causal work insofar as behaviour is concerned, the problem of mental causation re-asserts itself in a different form (Kim 1992; Crane 1995). This version of the problem of mental causation, which seems to generalise beyond the realm of the mental to all multiply-realizable phenomena, is still keenly debated in philosophy (Kim 2000, 2005; Hohwy 2008).

There is yet another rendering of the problem, however, that revolves around the causal efficacy of the specifically *representational* properties of mental phenomena. This third version typically arises in the philosophy of mind from the conjunction of three widely accepted theses about mental phenomena and their physical realization in the brain:

The content causation problem

1. Mental phenomena are causally efficacious of behaviour in virtue of their representational contents.
2. The representational contents of mental phenomena are not determined by the intrinsic properties of the brain.
3. The brain is causally efficacious of behaviour in virtue of its intrinsic properties.

The first of these theses is a fundamental tenet of both folk psychology and the computational theory of mind that has been constructed on its foundations. It is simply common sense that our perceptions and thoughts are about various aspects of the world in which we are embedded. It is also commonsense that mental phenomena causally interact with other mental phenomena and bodily behaviour in a fashion determined by their *content*—i.e., how they represent the world as being. Fodor refers to this as the "parallelism between content and causal relations" (1987).

The second thesis is widely accepted because most contemporary philosophers think that the representational properties of mental phenomena are determined at least in part by factors beyond the brain. This is the conclusion drawn from a number of famous thought experiments implicating twin-earth, arthritis, and various species of tree (Putnam 1975; Burge 1979, 1986). But, even more importantly, the second thesis seems to be an entailment of the most popular approach among philosophers for explaining how the representational properties of mental phenomena are determined. This is the conjecture that mental phenomena are contentful in virtue of their causal relations with those aspects of the world they are about (Adams & Aizawa 2010).

The final thesis is consistent with all we know about the brain basis of behavioural causation. While the brain enters into complex causal relations with aspects of the environment via multifarious sensory channels, our best neuroscience informs us that the changes to musculature that constitute our behavioural responses are wholly determined by the intrinsic properties of the brain to which they are causally connected.

In conjunction, these three widely accepted theses form an inconsistent triad. This generates a distinct and narrower version of the problem of mental causation: How can mental phenomena be causally efficacious of behaviour in virtue of their representational contents if these contents are not determined by intrinsic properties of the brain? In what follows, I shall refer to this as the *content causation problem*. This is the version of the problem of mental causation with which I shall be concerned in this paper (see e.g., Kim 2006, pp. 200–202).

There are some philosophers who seek to resolve the content causation problem by rejecting either the first[1] or the third[2] of the theses composing the inconsistent triad. However, the most popular response has been to reject or at

1 This, for example, is one way of construing Dennett's instrumentalism (1978, 1987).

2 There has been a some discussion in the literature about whether the *relational* properties of brain states are implicated in the causation of behaviour. The standard way of defending this claim is by individuating behaviour *broadly*, so as to incorporate factors beyond bodily movements (Burge 1986; Wilson 1994). But many philosophers think this form of individuation does great violence to scientific practice in general and to neuroscience in particular, and hence this way of resolving the problem of content causation is thought to seem very unpromising (Fodor 1987).

least modify the second thesis. This leads to the the *narrow content program*:

> The project of developing an account of mental phenomena according to which (at least the causally relevant component of) their representational properties are determined by intrinsic properties of the brain.

There are a number of different proposals about narrow content in the literature. Two of these have been particularly prominent. One is Fodor's suggestion that narrow contents can be unpacked as "functions from contexts to truth conditions" (1987, Ch. 2). The other is that narrow content is determined by "short-armed functional roles" (Block 1986; Loar 1981, 1982). But these (and other) proposals have been roundly criticised for failing to capture the *relational* character of mental content:

> The main charge has been that narrow content, as construed in these accounts, is not real content. When one thinks of an apple, what one thinks about is not a role or a function, but a fruit. Real content must put the subject in cognitive contact with the external world. […] A water concept, for example, must involve a relation between the thought wherein the concept is deployed and some worldly property or kind, presumably having to do with water. The problem with narrow content, construed as short-armed functional role or as a function from contexts to wide contents, is that it is not clear how it could involve any such relation. (Kriegel 2008, p. 308)

At this point, however, we seem to butt up against a classic paradox. On the one hand, those theories that appear to capture the relational character of mental content (i.e., causal theories) hold that content is not wholly determined by the intrinsic properties of the brain and, hence, imply that it isn't causally efficacious of behaviour. On the other hand, theories with the potential to account for the causal efficacy of mental content (i.e., narrow content theories), fail to capture its relational character. A solution to the content causation problem thus requires something that prima facie appears impossible: an explanation of the *relational* character of mental content that invokes only the *intrinsic* properties of the brain. Little wonder then that many philosophers despair of ever finding a solution to this puzzle.[3]

It is reasonable to hazard, however, that one of the main barriers standing in the way of a more productive treatment of the content causation problem is the radically underdeveloped understanding of mental content with which contemporary philosophy operates. In the foregoing quotation, for example, Kriegel characterises the relational character of content in terms of a subject's "cognitive contact" with the external world; yet he readily admits elsewhere that this notion is "not altogether transparent" (Kriegel 2008, p. 305). This is typical of the literature on this topic, which has become accustomed to describing content using the notoriously vague language of *aboutness*. While this language might capture our commonsense intuitions about mental phenomena, its imprecision may prevent us from discerning the lineaments of candidate solutions to the content causation problem.

This last point, at least, gives us the motivation for intruding yet another discussion into this already crowded philosophical space. The foundational conjecture upon which this paper is based is that the apparent insolubility of the content causation problem issues from an impoverished and unenlightening account of the relational character of mental content. Furthermore, this impoverishment is largely a con-

3 Perhaps the best we can do, according to some of these, is to accept that the representational properties of mental phenomena are causally inert, but to argue that there is enough room between explanation and causation for representational properties to be *explanatorily relevant*—despite their inertness (Baker 1993; Block 1989; Fodor 1986, 1989; Heil & Mele 1991; Jackson & Pettit 1990a, 1990b; LePore & Loewer 1989). A more radical response is to opt out of representation-based explanation altogether, as advocated originally by eliminativists (Churchland 1981; Stich 1983), and more recently by anti-representationalists (Brooks 1991). Finally, note that another radical position currently fashionable in philosophy—the extended-mind hypothesis (Menary 2010)—doesn't represent a solution to the content causation problem, since it signally fails to align mental phenomena with the brain-based causation of behaviour.

sequence of the *dyadic* conception of mental representation that has hitherto conditioned most philosophical thinking in this area. By contrast, a minority of philosophers has argued that mental representation is more properly analysed as a *triadic* relation. Triadicity, I will argue, yields a richer and ultimately more illuminating account of the relational character of mental content. Armed with this alternative treatment, we are in a position to assess the content causation problem anew. On the one hand, this novel viewpoint confirms the worry philosophers have expressed that causal theories of mental content are impossible to reconcile with the brain-based causation of behaviour. On the other hand, and much more positively, the triadic conception reveals a path that, from the perspective of content causation at least, looks more promising. The proposal that we travel down this path will undoubtedly face resistance, since it requires us to rehabilitate an approach to mental content that is unpopular in contemporary philosophy. But this approach, I shall conclude, seems unavoidable if we are to explain how mind matters.

## 2 The triadicity of representation

The bulk of philosophical writing on representation in general and mental representation in particular assumes, either explicitly or implicitly, that representation is a dyadic relation between something that does the represent*ing* and something that is represent*ed*. The task for a theory of representation, from this perspective, is to explain the necessary and sufficient conditions under which this dyadic relation obtains (see e.g., Stich 1992). But such a dyadic conception provides very little guidance about the relational character of representational content. All we have to work with is a mysterious action-at-a-distance phenomenon, whereby one part of the world, in virtue of the obtaining of a certain relation, is about another part.

To fill this gap, philosophers have almost invariably modelled their understanding of content on the semantic properties of the elements that compose our natural languages. Given the towering influence of Tarskian truth-conditional semantics in this field, it is inevitable that the relational character of representational content is usually characterised in terms of *reference* (Kriegel 2008, p. 305). But such an approach, while perhaps appropriate for linguaform representation, sits awkwardly with all manner of the non-linguistic forms of representation with which we are familiar (Haugeland 1991; Fodor 2007; Cummins & Roth 2012). Moreover, it is not obvious we are more enlightened by replacing talk of aboutness with that of reference.

In this context, it is worth observing that over the years a minority of philosophers has expressed dissatisfaction with the dyadic conception of representation. The most salient complaint is that such an approach fails to take into consideration the role that "users" of representation play. The general thought here is that some parts of the world don't represent other parts solely in virtue of some relationship between them; that the former represent the latter only when they are employed by some system to perform this representational function. According to Dennett, for example, physical entities "are by themselves quite inert as information bearers. […] They become information-bearers only when given roles in larger systems" (1982, p. 217). Likewise, Millikan has long observed that a biological approach to representation forces one to consider not just the "production" of representations, but also their "consumption" (1984). And, in a similar vein, Bechtel argues that that since whether something acts as a representation is ultimately determined by its function for some user, it follows that there are "three interrelated components in a representation story: what is represented, the representation, and the user of the representation" (1998, p. 299).

This *triadic* conception of representation is not new, of course, since it forms the basis of Charles Sanders Peirce's theory of semiotics, which was developed in the latter part of the 19$^{th}$ century (Hardwick 1977). Indeed, Peirce's (sometimes obscure) writings embody one of the most comprehensive analyses of representation in all of philosophical literature. Peirce approached this issue principally by investigating those public forms of representation with which

we are all familiar—words, sentences, paintings, photographs, sculptures, maps, and so forth—but he also sought to apply his triadic analysis to the special case of mental representation. This suggests that Peirce's writings might be an appropriate point of departure for exploring what the triadic conception entails about the relational character of representational content.

This strategy is very effectively adopted by von Eckardt when, following Peirce's lead, she analyses representation as a triadic relation involving a "representing vehicle, a represented object, and an interpretation" (von Eckardt 1993, pp. 145-149).[4] As with dyadic stories, the representing vehicle is the physical object (e.g., a spoken or written word, painting, map, sculpture, etc.) that is about something, and the represented object is the object, property, event, relation, or state of affairs that the vehicle is about. It is the addition of the interpretative relatum that sets the triadic account apart:

> A sign [i.e., a representing vehicle] [...] is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which is created I call the *interpretant* of the first sign. (von Eckardt 1993, p. 145)

Interpretation is thus understood as a cognitive effect in the subject for whom the vehicle operates as a representation. But as von Eckardt observes, not any kind of effect will do. This cognitive effect, presumably implicating the production of *mental* representing vehicles, must bring the subject into some appropriate relationship to the original vehicle's represented object (von Eckardt 1993, p. 157). Given this constraint, it is natural to interpret this third relatum in terms of the subject's *thinking* about the object in question. So (non-mental) representation, on the triadic story, is a *functional* kind: it is a process whereby a representing vehicle triggers a thought (or thoughts) in a subject about a represented object.

There are a couple of significant consequences of the triadicity of representation. The first is that, contrary to a dyadic story, representing vehicles aren't about anything independent of interpretations. Words, sentences, paintings, photographs, sculptures, maps, and so forth, considered in isolation from the cognitive impact they have on us, don't represent. This, of course, does some violence to the way that we talk about public representing vehicles—but it is far from catastrophic. The relevant revision is to think of these physical entities as possessing the *capacity* to trigger the necessary cognitive effects in us. The second (and, for our purposes, more important) consequence is that, unlike dyadic accounts in which content is unpacked solely in terms of relations between vehicles and represented objects, the triadic story entails that content is also conditioned by the interpretative relatum. This imposes an additional explanatory requirement on theories of content determination. It is not enough to merely explain how relations between vehicles and objects make it the case that the former are about the latter. These theories must also explain how it is in virtue of these relations that representing vehicles are capable of triggering thoughts in subjects about represented objects.

Once it has been suitably modified for the special, and presumably foundational, case of *mental* representation, the additional explanatory requirement that triadicity imposes on theories of content determination can form the basis of a richer account of the relational character of mental content. Such modification is necessary, of course, because treating the interpretation of mental vehicles solely in terms of a subject thinking about a represented object violates the *naturalism constraint*. This is the requirement that we explain mental representation without recourse to the antecendently representational (see e.g., Cummins 1989, pp. 127–129; Cummins 1996, pp. 3–4; Dretske 1981, p. xi; Fodor 1987, pp. 97–98; Millikan 1984, p. 87; von Eckardt 1993, pp. 234–239).

The relevant modification is fairly obvious, however, and represents a well-trodden path in

4 Von Eckardt actually uses the terms "representation bearer", "representational object", and "interpretant" to describe the three relata implicated in representation. I prefer the terminology I have used here because it is more consistent with the philosophical literature on mental representation.

philosophy. From the perspective of Peirce's triadic analysis, the role of interpretation is to forge a psychologically efficacious connection between the user of a representing vehicle and the vehicle's object. With public forms of representation it is perfectly acceptable to unpack this in terms of the (non-mental) vehicle activating thoughts directed at the object. But if we allow this story to run a little further it will point us in the right direction for the interpretation of mental vehicles too. Thoughts directed at objects modify our behavioural dispositions towards these same objects. This is why public forms of representation are so useful—they enable us to regulate our behaviour towards selective aspects of the world. But this story can be transported into the brain in order to account for the interpretation of mental representing vehicles. Instead of external vehicles triggering thoughts, and these in turn modifying behavioural dispositions, we simply suppose that mental vehicles have the same cognitive and ultimately behavioural effects. This acts to block the threatened regress since, presumably, it is possible to unpack behavioural dispositions without invoking further mental representation.

We are now in a position to deliver on one of the aims enumerated in the introductory section: that of fashioning a more illuminating account of the relational character of mental content. We saw earlier that Kriegel describes this character in terms of the "cognitive contact" between mental phenomena and the worldly aspects they represent, but admits that this notion isn't particularly transparent. Happily, the triadic analysis of mental representation affords a means of explicating what this cognitive contact consists in. Rather than simply employing the vague language of aboutness, the triadic analysis encourages us to understand the relational character of mental content in terms of the capacity of mental phenomena to regulate the behaviour of subjects towards specific aspects of the world. Cognitive contact is thus a relatively straightforward causal capacity. It is the capacity of cognitive creatures, bestowed by their internal states, to respond selectively to elements of the environment in which they are embedded.

This is where things currently stand. A solution to the content causation problem requires something that prima facie appears impossible—namely, an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain. But the paradoxical appearance of content causation, I have suggested, might be a legacy of the dyadic conception of representation that has conditioned philosophical intuitions about content determination, but which provides little guidance about the relational character of mental content. The triadic analysis of representation, I have argued, generates a more enlightening account of this relational character—one pitched in terms of the causal capacities of cognitive creatures to regulate their behaviour towards specific aspects of their environments. From the perspective of this analysis, therefore, a solution to the content causation problem requires a theory of content determination to explain how relations between mental vehicles and their represented objects can endow subjects with the capacity to respond selectively to those very features of the world.

Philosophers seeking to fashion theories of mental content determination over the centuries have famously focused on just two kinds of relations between mental vehicles and their represented objects: "causal" relations and "resemblance" relations (Fodor 1984, pp. 232–233). In the following section I shall engage in an all-too-brief appraisal of the prospects of these two world-mind relations to deliver a solution to the content causation problem.

## 3 World-mind relations and the content causation problem

Causal theories of mental content determination have dominated philosophy for nearly half a century. They hold that representing vehicles are contentful in virtue of being (actually, nomologically, or historically) caused by their represented objects (Devitt & Sterelny 1987; Fodor 1984, 1987, 1990; Stampe 1977, 1986). Perhaps the most well-known causal theory in all of the literature has been developed, through a number of iterations, by Dretske (1981, 1988, 1995).

What makes Dretske's account particularly apposite in the current context is that it has been fashioned, at least in its later iterations, to address explicitly the account of content intruded by the triadic analysis of mental representation (though Dretske doesn't use this terminology). At one point in his discussion, for example, Dretske states that he approves of Armstrong's (1973) description of beliefs as "maps by which we steer", and goes on to observe that "beliefs are representational structures that acquire their meaning, their maplike quality, by actually *using* the information it is their function to carry in steering the system of which they are part" (Dretske 1988, p. 81). This, for Dretske, motivates the very desideratum we extracted from the triadic analysis in the last section:

> It will not be enough merely to have a C [inner state of some cognitive system] that indicates F [i.e., causally covaries with some external condition] cause M [some observable behaviour]. What needs to be done [...] is to show how the existence of one relationship, the relationship underlying C's semantic character, can explain the existence of another relationship, the causal relationship (between C and M) comprising the behaviour in question. (1988, p. 84)

Dretske's response to this problem, famously, is to appeal to teleology. It is only when an inner state, which causally covaries with some bit of the external environment, is "recruited" by the cognitive system (either by an evolutionary design process or through individual development) to cause appropriate behaviour, that the state acquires the *function* of indicating that part of the environment, and thereby comes to *represent* it (Dretske 1988, pp. 84–89).

On the face of it, Dretske's theory seems to represent a promising solution to the content causation problem. From the perspective of the triadic analysis, a solution to this problem requires an explanation of how certain relations between mental vehicles and their objects can dispose cognitive subjects to behave selectively towards those represented objects. Dretske's elegant proposal is that reliable causal relations between inner states and environmental conditions (i.e., when the latter reliably cause the former to be tokened) can endow cognitive systems with these dispositions when the former states are conscripted by design processes to cause behaviour that is in some way relevant to the latter conditions. When this happens, the inner states are elevated to the status of representing vehicles, and their subsequent activity in bringing about behaviour directed towards their represented objects are examples of content causation.

Unfortunately, a closer inspection of Dretske's suggestion reveals a fundamental flaw. Contrary to what he contends, the relations at the core of his proposal are powerless to explain the required behavioural dispositions. Rather than describing this problem in the abstract, let me illustrate it using one of Dretske's favourite examples of a very simple representation-using system:

> A drop in room temperature causes a bimetallic strip in [a thermostat] to bend. Depending on the position of an adjustable contact, the bending strip eventually closes an electrical circuit. Current flows to the furnace and ignition occurs. The thermostat's behaviour, its turning the furnace on, is the bringing about of furnace ignition by events occurring in the thermostat—in this case [...] the closure of a switch by the movement of a temperature-sensitive strip [...].
>
> The bi-metallic string is given a job to do, made part of an electrical switch for the furnace, because of what it indicates about room temperature. Since this is so, it thereby acquires the function of indicating what the temperature is [...]. We can speak of [...] representation here. (Dretske 1988, pp. 86–87)

There is a subtle sleight of hand at work here, however. It is Dretske's contention that the bimetallic strip is recruited (by the manufacturer) to play a causal role in the thermostat because

of what it indicates about ambient temperature. But that's not the full story. Bi-metallic strips have an *additional* property that appeals to the manufacturers of thermostats: their degree of curvature corresponds in an orderly fashion with ambient air temperature, such that it can be configured to complete a circuit when the temperature drops to a pre-set level.

In Dretske's thermostat example, therefore, there are two distinct relations between representing vehicles and represented objects: a systematic correspondence relation (wherein variations in ambient air temperature are mirrored by orderly variations in the bi-metallic strip's shape) and an indication relation (wherein variations in ambient air temperature *cause* variations in the bi-metallic strip).[5] These two relations are not independent of one another, of course, as the former is mediated by the latter. But we can still consider which of these relations is doing the work, insofar as the capacity for the thermostat to control the behaviour of the furnace is concerned. And here the answer is clear: it is the fact that the curvature of the bi-metallic strip systematically mirrors the temperature, and not the causal covariation per se, that explains its capacity to operate the furnace in an appropriate manner. Consider the counterfactuals: curvature correspondence without causal covariation (e.g., where a mere correlation exists) would still generate the appropriate behaviour, but causal covariation without curvature correspondence (e.g., where the bi-metallic strip heats up but maintains its shape) wouldn't. The important point is that while the causal relation plays an important role in mediating the correspondence relation, it is the latter, not the former, that explains the thermostat's capacity to bring about the desired behaviour.[6]

So Dretske's own example fails to satisfy the desideratum that he set for himself: the obtaining of a reliable causal connection between ambient air temperature and the bi-metallic strip doesn't explain the thermostat's capacity to control the behaviour of the furnace. Moreover, this example illustrates a fundamental problem with *all* causal theories of mental content determination: there is a fatal disconnect between world-mind causal relations, on the one hand, and the mind's behavioural dispositions on the other. This disconnect exists because any (actual, nomological, or historical) causal relations that might exist between external conditions and inner vehicles do not explain, in their own right, how a cognitive system inherits the capacity of behaving sensitively to the former. Whether cognitive systems have this capacity is determined by the properties of their inner vehicles in concert with their organizational, architectural, and motoric properties. And while external conditions can cause tokenings of and alterations to inner vehicles, the mere obtaining of such causal relations can't explain how the tokened or altered vehicles are capable of interacting with these multifarious systemic properties such that they bestow the appropriate behavioural dispositions.[7] This is why manufacturers are very choosy about the materials from which they construct thermo-

5 Dretske scholars will cry foul at this point, of course. This is because Dretske claims that while indication is mostly founded on causal relations, it need not be. Indeed, he goes as far as to suggest that indication obtains whenever there is a non-coincidental covariation between vehicle and object (Dretske 1988, pp. 56–57). But this characterisation of indication transforms Dretske's proposal into something close to a resemblance theory (the approach to be examined in the next section), since it privileges systematic correspondence relations over causal relations. Consequently, insofar as Dretske's position is to be understood as a causal theory of content determination (as is widely assumed in the literature), it is essential that indication is interpreted as a relation of causal covariation. I adopt this interpretation in what follows.

6 One would expect to find causal relations mediating systematic correspondence relations between the representing vehicles of biological systems and aspects of the world. But, as Dretske is well aware, this is not always the case. Nature will make do with what works, and some kind of systematic correspondence in the absence of causal commerce will do just as well. This can be illustrated by another of Dretske's favourite examples: the evolutionary recruitment of magnetosomes in anaerobic bacteria to steer them towards deoxygenated water (1986). According to Dretske, evolutionary forces operating on these bacteria have selected magnetosomes because they are indicators of anaerobic water capable of influencing the direction in which the bacteria swim. But as Millikan has pointed out, the connection between the orientation of magnetosomes and anaerobic water is merely correlational, not causal (2004, Ch. 3). Magnetosomes indicate and steer northern hemisphere anaerobic bacteria in the direction of magnetic north, which results in these bacteria swimming into deeper (and hence deoxygenated) water. But there is no causal connection between magnetic north and deoxygenated water. In this case, therefore, magnetosomes have been selected because their alignment systematically corresponds with the direction of anaerobic water, not in virtue of any causal covariation between them.

7 Cummins reaches a similar conclusion, though via a somewhat different route (1996, p. 74).

stats. Engineering a causal covariation relation between ambient air temperature and the innards of a thermostat is easy; engineering these innards such that they possess the requisite causal capacities is a great deal harder.

Ultimately, therefore, Dretske's ingenious attempt to solve the content causation problem doesn't succeed. Dretske holds that the internal states of cognitive systems are elevated to representing roles when they are recruited by design processes to regulate behaviour towards the external conditions they indicate. He takes this to be a case of genuine content causation because he thinks that the causal relations between represented objects and representing vehicles can explain the causal activity in which the vehicles subsequently engage. But Dretske has over-estimated the explanatory power of world–mind causal relations. And he has done so because he has illicitly smuggled into his story a quite distinct form of content determination—one that exploits systematic correspondence relations between representing vehicles and their represented objects. Such systematic correspondences are, of course, a species of resemblance relation. The failure of Dretske's proposal is thus instructive, since it suggests that this alternative world–mind relation offers some prospect of a solution to the content causation problem.

Resemblance theories of content determination hold that representing vehicles are contentful in virtue of resembling their represented objects. The most obvious and straightforward application of this idea can be found in many public forms of representation, from photographs and paintings to sculptures and maps. But what is most significant about this approach for our purposes is that when vehicles resemble their objects, the former actually *replicate* the latter in some way, either by reproducing their properties or their relational organisation (more about which in the next section). And this affords a relatively straightforward way of explaining how a physical device, in virtue of incorporating vehicles that bear resemblance relations to the world, acquires a capacity to behave selectively towards particular elements of the environment. The thermostat's bi-metallic strip reproduces—in the variations in its degree of curvature—the diachronic pattern of magnitude relations between ambient air temperature. Once this bi-metallic strip is incorporated into the thermostat, therefore, this device has a set of internal vehicles that dynamically replicates the external temperature. It is then simply a matter of rigging the innards of the thermostat so that its operation of the furnace is regulated by these internalised surrogates (Swoyer 1991).

Dretske is correct to judge this an example of content causation. It is a case in which the exploitation of a relation between environmental conditions and inner vehicles explains how the latter are capable of modifying a device's behavioural dispositions towards particular aspects of the world. But what is seldom acknowledged about this much-used example is that it demonstrates the causal efficacy of content fixed by resemblance. Despite this virtue, resemblance theories of mental content determination are unfashionable in contemporary philosophy, largely because they are widely thought to suffer from a number of fatal flaws. Before we end, therefore, it would be wise to engage in a degree of resemblance rehabilitation. This turns out to be easier than one might expect, however, once we adopt the perspective of the triadic conception of representation.

## 4 Rehabilitating resemblance

Despite the widespread assumption that they are fatally flawed, it's hard to find a sustained discussion of the problems associated with resemblance theories of content determination. Instead, one finds scattered somewhat haphazardly across the literature brief allusions to the same five objections. The canonical rendering of three of these can be found in the opening paragraphs of Nelson Goodman's Languages of Art:

> The most naive view of representation might perhaps be put somewhat like this: "A represents B if and only if A appreciably resembles B", or "A represents B to the extent that A resembles B". Vestiges of this view, with assorted refinements, per-

> sist in most writing on representation. Yet more error could hardly be compressed into so short a formula.
>
> Some of the faults are obvious enough. An object resembles itself to the maximum degree but rarely represents itself; resemblance, unlike representation, is reflexive. Again, unlike representation, resemblance is symmetric: B is as much like A as A is like B, but while a painting may represent the Duke of Wellington, the Duke doesn't represent the painting. Furthermore, in many cases neither one of a pair of very like objects represents the other; none of the automobiles off an assembly line is a picture of any of the rest; and a man is not normally a representation of another man, even his twin brother. Plainly, resemblance in any degree is no sufficient condition for representation. (1969, pp. 3–4)

In short, representation can't be based on resemblance, since the latter is *reflexive* (where the former isn't), *symmetric* (where the former isn't), and *insufficient* (all manner of objects resemble others without representing them). But however influential these three objections might be when applied to a dyadic analysis of representation, they lose all force in the context of a triadic conception. This conception agrees with Goodman that relations between vehicles and their represented objects are insufficient to confer representational status. A representing vehicle must also undergo interpretation, either by triggering thoughts in a cognitive subject or by modifying the subject's behavioural dispositions. And it is this process of interpretation, according to a resemblance theory, that also enforces the non-reflexivity and asymmetry of representation.

A fourth objection is that resemblance theories of mental content are incompatible with our commitment to physicalism:

> If mental representations are physical things, and if representation is grounded in [resemblance], then there must be physical things in the brain that are similar to (i.e., share properties with) the things they represent. This problem could be kept at bay only so long as mind-stuff was conceived as nonphysical. The idea that we could get redness and sphericity in the mind loses its plausibility if this means we have to get it in the brain. (Cummins 1989, p. 31)

But this objection is easily deflected once a proper understanding of the different forms of resemblance is in place. The most straightforward kind of resemblance—the kind that Cummins in the above quotation has in mind, for example—involves the sharing of one or more properties. This relationship can be termed *first-order resemblance*.[8] It is this kind of resemblance that grounds the content of many public forms of representation, such as paintings, sculptures, and scale models. As Cummins points out, however, first-order resemblance is clearly unsuitable as a ground of mental content, since it is incompatible with what we know about the brain.

There is, nonetheless, a more abstract species of resemblance available. The requirement that representing vehicles share properties with their represented objects can be relaxed in favour of one in which the *relations* among a system of vehicles mirror the *relations* among their objects. This relation-preserving mapping between two systems can be called *second-order resemblance*.[9] And while it is extremely unlikely

8 I am here adapting terminology used by Shepard & Chipman (1970).

9 To be more precise, suppose $S_V = (V, R_V)$ is a system comprising a set $V$ of objects, and a set $R_V$ of relations defined on the members of $V$. The objects in $V$ may be conceptual or concrete; the relations in $R_V$ may be spatial, causal, structural, or inferential, and so on. For example, $V$ might be a set of features on a map, with various geometric and part–whole relations defined on them. Or $V$ might be set of well formed formulae in first-order logic falling under relations such as identity and consistency. There is a *second-order resemblance* between two systems $S_V = (V, R_V)$ and $S_O = (O, R_O)$ if, for at least *some* objects in $V$ and *some* relations in $R_V$, there is a one-to-one mapping from $V$ to $O$ and a one-to-one mapping from $R_V$ to $R_O$ such that when a relation in $R_V$ holds of objects in $V$, the corresponding relation in $R_O$ holds of the corresponding objects in $O$. In other words, the two systems resemble each other with regard to their abstract relational organisation. As already stressed, resemblance of this kind is independent of first-order resemblance, in the sense that two systems can resemble each other at second-order without sharing properties. Second-order resemblance comes in weaker and stronger forms. As defined it is relatively weak, but if we insist on a mapping that takes

that first-order resemblance is the general ground of mental content (given what we know about the brain), the same does not apply to second-order resemblance. Two systems can share a pattern of relations *without* sharing the physical properties upon which those relations depend. Second-order resemblance is actually a very abstract relationship: essentially nothing about the physical form of a system of representing vehicles is implied by the fact that it resembles a set of represented objects at second-order. Contrary to the fourth objection, therefore, a theory of mental content determination that exploits second-order resemblance is compatible with physicalism.[10]

However, this emphasis on second-order resemblance, at least in the eyes of many theorists, takes this approach to content determination out of the frying pan and into the fire. This is because the highly abstract nature of second-order resemblance invites the charge that it entails a massive and intractable indeterminacy of mental content. And it is this fifth objection, perhaps more than any other, that accounts for the current unpopularity of resemblance theories (Sprevak 2011).

The problem here can be illustrated by returning to Dretske's thermostat. The world–mind relation that does all the heavy lifting here constitutes a second-order resemblance: the relations among the representing vehicles (the set of bi-metallic strip curvatures) systematically mirror the relations among the represented objects (the set of ambient air temperatures).[11] The worry embodied in the fifth objection, however, is that this same set of representing vehicles will second-order resemble not just the temperature surrounding the thermostat but any set of objects, regardless of their nature and location, that shares its relational organisation. This fact is entailed by the abstract nature of second-order resemblance. And this is a problem, of course, because it suggests that second-order resemblance is incapable of delivering determinate content. The most we can say about the thermostat's bi-metallic strip is that its curvature represents that potentially large and motley collection of objects with which it systematically corresponds. And this would seem to be a long way from saying it represents the temperature of the ambient air.

Fortunately, the triadic analysis again offers a way to surmount this difficulty. On this account, representations aren't manufactured solely from relations between vehicles and the objects they represent. Rather, the process of interpretation must also be thrown into the mix. We've seen that interpretation is discharged ultimately in terms of modifications to a system's behavioural dispositions. But not any old modifications will do—representing vehicles must modify the system's dispositions towards their represented objects. Consequently, interpretation plays an important content-limiting role. Specifically, a system's behavioural dispositions will anchor its representing vehicles to particular represented domains. Once a domain is secured in this way, second-order resemblance relations determine the content of the individual vehicles. In the case of the thermostat, for example, the behavioural dispositions of the system restrict the represented domain to ambient air temperature, and the second-order resemblance relations determine what temperature each vehicle represents.

*every* element of $V$ onto some element of $O$, and, in addition, preserves *all* the relations defined on $V$, then we get a strong form of resemblance known as a *homomorphism*. Stronger still is an *isomorphism*, which is a one-to-one relation-preserving mapping such that every element of $V$ corresponds to some element of $O$, and every element of $O$ corresponds to some element of $V$. When two systems are isomorphic their relational organisation is identical. In the literature on second-order resemblance the focus is often placed on isomorphism (see e.g., Cummins 1996, pp. 85–111), but where representation is concerned, the kind of correspondence between systems that is likely to be relevant will generally be weaker than isomorphism. For a much fuller discussion of second-order resemblance, see O' Brien & Opie (2004).

10 Two early theorists who sought to apply second-order resemblance to mental representation are Palmer (1978) and Shepard (Shepard & Chipman 1970; Shepard & Metzler 1971). More recently, Blachowicz (1997), Cummins (1996), Gardenfors (1996), O' Brien & O' Brien (1999), O' Brien & Opie (2004), and Swoyer (1991), have all defended second-order resemblance theories.

11 Notice that in this case, the second-order resemblance is sustained *structural* relations among the set of representing vehicles (i.e., the set of bi-metallic strip curvatures). This is an example of what Palmer (1978) calls *natural isomorphism*, since the second-order resemblance relations are sustained by constraints *inherent* in the vehicles, rather than being imposed *extrinsically*. Elsewhere I have used the term *structural resemblance* to describe this kind of second-order relationship and to distinguish it from *functional resemblance*, where the second-order resemblance relations are sustained by *causal* relations among the vehicles—see O' Brien & Opie (2004).

## 5 Conclusion: How mind matters

It is time to take stock. We began with three commonplace theses about mental phenomena and their physical realization in the brain that together generate a profound puzzle about mental causation. This is the content causation problem: that of explaining how the specifically representational properties of mental phenomena can be causally efficacious of behaviour. This problem has an air of insolubility about it because it appears to require something impossible: an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain. It has been the foundational conjecture of this discussion, however, that this despair issues from the impoverished understanding of content that attends the dyadic analysis of mental representation, and that once we adopt the perspective of the triadic conception our view of the content causation problem is transformed.

The insight offered by triadicity is that the relational character of mental content is to be discharged ultimately in terms of our behavioural dispositions towards features of the world. This offers a way forward with the content causation problem because it suggests that, rather than seeking to explain some kind of mysterious action-at-a-distance, the task for a theory of content determination is to explain how the obtaining of world-mind relations can dispose cognitive systems to respond selectively to certain elements of their embedding environments.

According to most naturalistically-inclined philosophers, there are just two candidate mind–world relations available: causal relations and resemblance relations. Causal theories of content determination dominate the contemporary landscape, but our analysis confirms what many have suspected—namely, that causal theories offer no prospect of a solution to the content causation problem. The reason for this, however, is not because such theories appeal to relations that incorporate factors beyond the brain. All theories of mental representation, in their efforts to explain the relational character of mental content, are forced to invoke world–mind relations of some kind. The problem with causal theories, at least from the triadic perspective, is the disconnect between world–mind causal relations and a system's behavioural dispositions. The obtaining of causal relations between external conditions and inner vehicles cannot explain how the latter endow systems with the capacity to respond in a discriminating fashion towards the former.

This leaves us with resemblance relations. The problem here is that resemblance theories of content determination have for many years been deeply unpopular in philosophy. But this is another point where the triadic conception of representation pays rich dividends. Most of the problems associated with resemblance theories don't look so severe when viewed from the triadic perspective. This is encouraging, because resemblance does offer some prospect of a solution to the content causation problem. The key here is that the mere obtaining of the resemblance relation entails that representing vehicles replicate their represented objects. This ensures that the former have properties that can be exploited to shape the behavioural dispositions of cognitive systems towards the latter.

Consequently, if the line of argument presented here is on the right track, then resemblance theories of mental content determination must be rehabilitated and subjected to scrutiny and development. It goes without saying that there are a great number of significant hurdles yet to be overcome. I have focused, for instance, on just one very simple example of a representation-using system. There remains, accordingly, a large question mark over whether the resemblance solution to the problem of content causation scales up to more sophisticated cognitive creatures, let alone to the immense complexities of our own mental phenomena. But we have to start somewhere. And as things currently stand, resemblance theories appear to be obligatory, since they alone offer some prospect for explaining how mind matters.

## References


Adams, F. & Aizawa, K. (2010). Causal theories of mental content. http://plato.stanford.edu/entries/content-causal/

Armstrong, D. (1973). *Belief, truth and knowledge.* Cambridge, UK: Cambridge University Press.

Baker, L. R. (1993). Metaphysics and mental causation. In J. Heil & A. Mele (Eds.) *Mental Causation* (pp. 75-96). Oxford, UK: Oxford University Press.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22* (3), 295-318. 10.1207/s15516709cog2203_2

Blachowicz, J. (1997). Analog representation beyond mental imagery. *Journal of Philosophy*, *94* (2), 55-84.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, *10* (1), 615-678. 10.1111/j.1475-4975.1987.tb00558.x

——— (1989). Can the mind change the world? In G. Boolos (Ed.) *Meaning and method: Essays in honor of Hilary Putnam* (pp. 137-170). Cambridge, UK: Cambridge University Press.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, *47* (1-3), 139-159. 10.1016/0004-3702(91)90053-M

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4* (1), 73-122. 10.1111/j.1475-4975.1979.tb00374.x

——— (1986). Individualism and psychology. *Philosophical Review*, *95* (1), 3-45. 10.1007/978-94-009-2649-3_3

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, *78* (2), 67-90. 10.2307/2025900

Crane, T. (1995). The mental causation debate. *Proceedings of the Aristotelian Society*, *69*, 211-236.

Cummins, R. C. (1989). *Meaning and mental representation.* Cambridge, MA: MIT Press.

——— (1996). *Representations, targets, and attitudes.* Cambridge, MA: MIT Press.

Cummins, R. C. & Roth, M. (2012). Meaning and content in cognitive science. In R. Schantz (Ed.) *Prospects for meaning* (pp. 365-382). Berlin, GER: de Gruyter.

Dennett, D. (1978). *Brainstorms.* Cambridge, MA: MIT Press.

——— (1982). Styles of mental representation. *Proceedings of the Aristotelian Society*, *83*, 213-226.

——— (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Devitt, M. & Sterelny, K. (1987). *Language and reality.* Cambridge, MA: MIT Press.

Dretske, F. (1981). *Knowledge and the flow of information.* Oxford, UK: Clarendon.

——— (1986). Misrepresentation. In R. Bodgan (Ed.) *Belief: Form, content, and function* (pp. 17-36). Oxford, UK: Oxford University Press.

——— (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge, MA: MIT Press.

——— (1995). *Naturalizing the mind.* Cambridge, MA: MIT Press.

Fodor, J. (1984). Semantics, Wisconsin style. *Synthese*, *59* (3), 231-250. 10.1007/BF00869335

——— (1986). Banish discontent. In J. Butterfield (Ed.) *Language, mind, and logic* (pp. 1-24). Cambridge, UK: Cambridge University Press.

——— (1987). *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge, MA: MIT Press.

——— (1989). Making mind matter more. *Philosophical Topics*, *17* (1), 59-79. 10.5840/philtopics198917112

——— (1990). *A theory of content and other essays.* Cambridge, MA: MIT Press.

——— (2007). The revenge of the given. In B. McLaughin & J. Cohen (Eds.) *Contemporary debates in the philosophy of mind* (pp. 105-116). Malden, MA: Blackwell.

Gardenfors, P. (1996). Mental representation, conceptual spaces and metaphors. *Synthese*, *106* (1), 21-47. 10.1007/BF00413612

Goodman, N. (1969). *Languages of art.* London, UK: Oxford University Press.

Hardwick, C. (Ed.) (1977). *Semiotics and significs: The correspondence between Charles S. Peirce and Victoria Lady Welby.* Bloomington, IN: Indiana University Press.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich & D. Rumelhart (Eds.) *Philosophy and connectionist theory* (pp. 171-206). Hillsdale, NJ: Lawrance Erlbaum.

Heil, J. & Mele, A. (1991). Mental causes. *American Philosophical Quarterly*, *28* (1), 61-71.

Hohwy, J. (Ed.) (2008). *Being reduced: New essays on reduction, explanation and causation.* New York, NY: Oxford University Press.

Jackson, F. & Pettit, P. (1990a). Causation and the philosophy of mind. *Philosophy and Phenomenological Research Supplement*, *50*, 195-214.

——— (1990b). Program explanation: A general perspective. *Analysis*, *50* (2), 107-117.

Kim, J. (1992). Multiple realization and the metaphysics

of reduction. *Philosophy and Phenomenological Research*, *52* (1), 1-26. 10.1017/CBO9780511625220.017

——— (2000). *Mind in a physical world.* Cambridge, MA: MIT Press.

——— (2005). *Physicalism, or something near enough.* New York, NY: Princeton University Press.

——— (2006). *Philosophy of mind.* Boulder, CO: Westview Press.

Kriegel, U. (2008). Real narrow content. *Mind and Language*, *23* (3), 304-328. 10.1111/j.1468-0017.2008.00345.x

LePore, E. & Loewer, B. (1989). More on making mind matter. *Philosophical Topics*, *17* (1), 175-191. 10.5840/philtopics198917117

Loar, B. (1981). *Mind and meaning.* Cambridge, UK: Cambridge University Press.

——— (1982). Conceptual role and truth conditions. *Notre Dame Journal of Formal Logic*, *23* (3), 272-283. 10.1305/ndjfl/1093870086

Menary, R. (Ed.) (2010). *The extended mind.* Cambridge, MA: MIT Press.

Millikan, R. G. (1984). *Language, thought and other biological categories.* Cambridge, MA: MIT Press.

——— (2004). *Varieties of meaning.* Cambridge, MA: MIT Press.

O'Brien, G. & O'Brien, J. (1999). Putting content into a vehicle theory of consciousness. *Behavioral and Brain Sciences*, *22* (1), 175-196.

O'Brien, G. & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In P. S. Clapin & P. Slezak (Eds.) *Representation in mind: New approaches to mental representation.* Amsterdam, NL: Elsevier.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. Lloyd (Eds.) *Cognition and categorization* (pp. 259-303). Hillsdale, NJ: Lawrence Erlbaum.

Putnam, H. (1975). The meaning of 'meaning'. In H. Putnam (Ed.) *Mind, language, and reality* (pp. 131-193). Cambridge, UK: Cambridge University Press.

Shepard, R. & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1* (1), 1-17. 10.1016/0010-0285(70)90002-2

Shepard, R. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171* (3972), 701-703.

Sprevak, M. (2011). Review of Willian H. Ramsey's 'Representation Reconsidered'. *British Journal for the Philosophy of Science*, *62* (3), 669-675.

Stampe, D. (1977). Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy*, *2* (1), 42-63. 10.1111/j.1475-4975.1977.tb00027.x

——— (1986). Verificationism and a causal account of meaning. *Synthese*, *69* (1), 107-137. 10.1007/BF01988289

Stich, P. (1983). *From folk psychology to cognitive science.* Cambridge, MA: MIT Press.

——— (1992). What is a theory of mental representation? *Mind*, *101* (402), 243-261. 10.1093/mind/101.402.243

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87* (3), 449-508. 10.1007/BF00499820

von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA: MIT Press.

Wilson, R. (1994). Wide computationalism. *Mind*, *103* (411), 351-372. 10.1093/mind/103.411.351

# Does Resemblance Really Matter?

## A Commentary on Gerard O'Brien

Anne-Kathrin Koch

In this commentary on Gerard O'Brien's "How does mind matter?—Solving the content causation problem", I will investigate the notion of *representational content* presented in the latter. With this notion, O'Brien aims at giving an explanation of how mind matters in physicalist terms. His argumentation is motivated by, and supposedly directed towards, a problem he calls *the content causation problem*. Regarding this, I am most interested in reconstructing how his account relates to the presuppositions that make this problem so pressing in philosophical enquiry. O'Brien provides a very interesting answer to the question of "why mental content matters", as motivated by the content causation problem. In particular, I will try to show that by making use of the notion of dispositions, it provides an interesting way of avoiding the presupposition that understanding content causation always requires the reduction of individual relational properties to individual intrinsic properties—probably because it is presupposed that such a reduction is impossible.

**Keywords**
Dispositions | Mental causation | Mental representation | Reduction

Commentator

Anne-Kathrin Koch
anne-kathrin.koch@gmx.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Gerard O'Brien
gerard.obrien@adelaide.edu.au
University of Adelaide
Adelaide, Australia



## 1 Introduction

Gerard O'Brien's paper "How does mind matter?—Solving the content causation problem" (this collection) is situated at the border of philosophy and cognitive science. The subject matter, as announced by the title, is the causal efficacy of mental content, especially of representational content. O'Brien approaches this subject in three argumentative steps: first he introduces us to a problem called the *content causation problem*, then he proposes a conception of representation that he calls the *triadic conception of representation*, and, third, he enriches this concept by proposing a second-order similarity theory of content determination.

In this commentary I will try to reconstruct how these three points relate to each other, focusing in particular on the role that the content causation problem plays in the other two argumentative steps. O'Brien's theory of representational content and its causal efficacy is doubtlessly interesting even when considered in isolation, as I will briefly outline in section 2. In section 3, I will try to show that the content causation problem demands us to be more specific than when just investigating the question of how mind matters. In section 4, I will try to show how O'Brien's account of the relational character of mental content, which is at the core

Koch, A. K. (2015). Does Resemblance Really Matter? - A Commentary on Gerard O'Brien.
In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 28(C). Frankfurt am Main: MIND Group. doi: 10.15502/9783958571051

of his argumentation, gains its philosophical relevance from implicit assumptions that form the conceptual foundations of the content causation problem as here formulated. In an attempt to assess whether his account must really be regarded as *solving* the content causation problem, I will highlight in section 5 how important it is to be specific about what we really mean if we suppose that representation is somehow *relational* in character.

## 2 How mind might matter

In his paper in this collection, Gerard O'Brien confronts the task of "explaining how mind matters" (p. 12). He does so, because he—rightly, I believe—identifies the fact "that our minds matter—that our beliefs and desires, and our perceptions and thoughts ultimately have a causal impact on our behaviour" as a ubiquitous and well-accepted, but unexplained phenomenon (p. 1).

O'Brien's investigation is motivated by the following question: "[h]ow can mental phenomena be causally efficacious of behaviour in virtue of their representational contents if those contents are not determined by intrinsic properties of the brain?" (O'Brien this collection, p. 2) He calls this question the "content causation problem" (ibid., p. 2). This specific way of approaching the matter of mental causation is set in the context of "three widely accepted theses about mental phenomena and their physical realization in the brain" (ibid., p. 2): (i) the supposed causal efficacy of mental phenomena is grounded in their representational contents, which, (ii) are taken to be *relational* properties of those phenomena (ibid., p. 2–3); and (iii) the results of neuroscience, which already provide us with an explanation of how behaviour is caused, only make use of the brain and its *intrinsic* properties in their explanation (ibid., p. 2). Hence, there is a need for an explanation of behaviour being caused by mental phenomena in virtue of their relational properties, and this explanation cannot easily make use of the explanation of the causation of behaviour that has already been provided by contemporary neuroscience. At first, it looks as if this shortfall is exactly what O'Brien is addressing.

Philosophical mainstream accounts of representation, O'Brien reminds us, are built on an understanding of representation as a two-place relation. Representational content is thus described in terms of *aboutness* and/or *reference.* O'Brien, however, advises us to abandon the traditional understanding of the notion of representation, i.e., the idea that representation is a two-place relation and adequately phrased in terms of one thing being *about* another (O'Brien this collection, p. 3–4). Instead, he proposes a triadic conception of representation, making representation a three-place relation between a represented object, a representing vehicle, and an interpretation (ibid., p. 5).

In the triadic picture, interpretation is "a cognitive effect [of the object] in the subject", thereby establishing a relationship between this subject and the represented object (O'Brien this collection., p. 5). The ingenious move here, of course, is that interpretation is explained in *causal* vocabulary. We should think of interpretation as "presumably implicating the production of mental representing vehicles" possessing new properties, which should in turn be thought of as "bring[ing] the subject into some appropriate relationship to the original vehicle's represented object" (ibid., p. 5), i.e., "modifying [the subject's] behavioural dispositions" (ibid., p. 6). At first, when O'Brien further describes those vehicles as "hav[ing] [...] cognitive and ultimately behavioural effects" (ibid., p. 6), it isn't clear exactly which category we are dealing with. I take the relata of the causation relation to be events, but understand talk of vehicles to be talk about objects.[1] I suggest that we understand the vehicles as modifying the system's behavioural dispositions insofar as, once produced, they have certain properties that are directly and specifically relevant for a causal process to take place (given that some sort of stimulus initiates the causal process). If we adopted a view of dispositions as second-order properties, i.e., the property of having certain properties that

1 If we want to avoid reification of the "vehicles", we might look at it them as time-slices in a complex, internal chain of events, i.e., of dynamic inner processes modifying a subject's global dispositions.

can be causally relevant (cf. Choi & Fara 2014), this would allow us to think of the vehicles as modifying global dispositions of a system as a whole, in the sense of *providing new ones*—that is, by themselves having novel dispositional properties. This way, we can analyze the obtaining of the representation relation as a specific causal process having taken place: the first step of that process is the triggering of the cognitive effect by the representandum (the first relatum); the second relatum is the event of interpretation itself; and the third relatum is the new vehicle produced during the event of interpretation that provides the subject with a new behavioural disposition towards the representandum. The representational character of mental content, in this picture, just rests on what we call "content" resulting from the multi-layer causal process described above.[2]

O'Brien's triadic account of representation describes the obtaining of the representation relation not in terms of our everyday intuitions about representation, but in terms of the job it is supposedly doing for us: bridging the gap between whatever is going on in the sphere of "the mental" and the external world by alluding to the causal chain that unites the two. It is thus understandable why the dyadic conception might be accused of hiding behind terms like "aboutness" or "reference": saying that something mental is about something external is just saying that there is a gap being bridged. Saying that something external sets a three-step causal chain in motion with the result that a subject has undergone a specific change in her behavioural dispositions seems much closer to saying what the bridge is made of.

Yet we should still dispose of the vague language of "specific change in her behavioural dispositions". What exactly makes this change specific? It was called "specific" because it selectively relates back to the object that set the causal chain in motion in the first place and which we would like to keep calling "the represented object."[3] But how can the change in a subject's behavioural dispositions make them pick out the exact same object from which this change originates?[4] The answer to this lies in the theory of content determination.

When holding that the representing vehicle brings about a change in the subject's behavioural dispositions, causal theories of content determination are supposedly to be abandoned because of a "disconnect between world–mind causal relations and a system's behavioural dispositions" (O'Brien this collection, p. 12). An appropriate theory of content determination—so says a desideratum that we gain from the results of the triadic analysis of representation—must "explain how [inner vehicles] endow systems with the capacity to respond in a discriminating fashion towards [external conditions]" (ibid., p. 12). For fulfillment of this criterion, O'Brien turns to *resemblance theories* of content determination, which "hold that representing vehicles are contentful in virtue of resembling their represented objects" (this collection, p. 9).

Within the triadic conception of representation, O'Brien identifies two hurdles for a resemblance theory that still need to be overcome: it must be shown how the theory can be compatible with physicalism, and it must be secured that the theory does not leave content indeterminate (ibid., pp. 9–11).

In order to secure the compatibility of a resemblance theory of content determination with physicalism, O'Brien turns away from the notion of *first-order resemblance* and instead makes use of *structural* or *second-order resemblance* (ibid., pp. 10–11). He thus avoids the seemingly naive and implausible thesis that

2 It might be said that this understanding of representation is plausible for paradigmatic cases of representation, like the representation of material objects, but that it is less clear whether it is also fit to capture cases that differ strongly from those paradigms, like the representation of abstract "objects".
A similar worry, which has been pointed out to me by an anonymous reviewer, concerns cases of *fictional* representations, e.g., future events. For this specific example, she/he suggests that we allow for the causal chain to be reversed. This would make the representandum part of the final event. However, this solution applies only to those cases of fictional representation where the representandum does *not yet* exist, but leaves the majority of cases of fictional representation inexplicable.

3 I will, inspired by O'Brien's terminology, keep refering to the representandum as the represented *object*. The term "object" is thereby used in a very wide sense and not intended to be restricted to single material objects. What exactly can take the place of an object in O'Brien's story of representation is yet to be determined.

4 This question presupposes that the established representation is actually correct and not a case of misrepresentation.

mental representations must actually share properties with what they represent (ibid., p. 10). Resemblance is taken to a more abstract level where, for example, something red can be mentally represented *with the representation resembling the representandum*, but *without* them both sharing the property of being red (ibid., pp. 10–11).

The second hurdle might seem redundant at first glance. Explaining how content is determined is basically the job description of a theory of content determination. It is still worth mentioning this as an obstacle, however, because the reliance on second-order resemblance makes this job look particularly difficult: second-order resemblance is too easily established. If a set of mental representations second-order resembles a pattern of colour shades, it might in virtue of the same relational organization also second-order resemble a pattern of locations in a two-dimensional space. Nevertheless, O'Brien trusts that within the triadic conception of representation, second-order resemblance will do the job. The idea is that some of the possibilities for the content of a vehicle that are left open by second-order resemblance are ruled out in the process of interpretation—interpretation is "content-limiting" by "anchoring vehicles" in "domains" (O'Brien this collection, p. 11). The preexisting behavioural dispositions influence the newly developing ones, so that they are not directed towards all domains with a specific relational organisation, but towards a selection of these.

So far, O'Brien has provided us with an interesting account of how mental phenomena are causally efficacious in virtue of their representational contents: the property

> $x$ has representational content

is analyzed as the property

> $x$ results from a causal process that brings about behavioural dispositions towards the object that triggered the causal process.

These behavioural dispositions, given their respective stimuli, can now yield causal effects.

But if we took this as O'Brien's only accomplishment, we would miss the most interesting part of his argument. Furthermore, we would take the second step before the first.

## 3 Content and causation: from a question to a problem

So far, I have interpreted O'Brien's formulation of the content causation problem ("How can mental phenomena be causally efficacious of behaviour in virtue of their representational contents if these contents are not determined by intrinsic properties of the brain?" O'Brien this collection, p. 2) as something along the lines of "How does mental causation in virtue of representational content work, if not in the way we already know it sometimes works?" In so doing, one already engages in an interesting discussion about content causation. But closer examination reveals that this understanding of the problem is an oversimplification. The content causation problem is supposed to be much more severe. It is not a problem about finding alternative explanations to the ones we already have, but about the *consistency* of all available explanations. A better understanding of the problem will help us to evaluate whether O'Brien's suggestions, which are doubtlessly interesting, are really motivated by the problem at hand.

The three theses, that (i) the causal efficacy of mental phenomena is grounded in their representational contents, that (ii) these "are not determined by the intrinsic properties of the brain" (O'Brien this collection, p. 2), and that (iii) the brain's causal efficacy for behaviour is grounded only in its *intrinsic* properties, supposedly "form an inconsistent triad" (ibid., p. 2). Yet, strictly speaking, they are not inconsistent: why not say that what we do in theses (i) to (iii) is gathering information about (human) behaviour—our object of enquiry—but on two levels of description? On both levels, we attempt, metaphorically speaking, to travel back along the causal chain that brings behaviour about. On the one level, we then discover that intrinsic properties of the brain are responsible for it to cause behaviour (ibid., p. 2, thesis 3). On another level, we trace behaviour back to

mental phenomena, which owe their capacity to cause it to their representational contents (ibid., p. 2, thesis 1). Why not assume that these two levels both provide us with (true) formulations of what is happening, but which—since they depend on two conceptual frameworks that are not intertranslatable—must be regarded as *nomologically incommensurable*? If so, they could never both be part of a unified causal theory (see Davidson 1970). This picture seems perfectly intelligible at first. But on both levels, we talk about causes of (presumably the same) behaviour.

Within a physicalist framework, we take behaviour to fall, in the end, under the description of a physical event. As such, it is subject to *the principle of causal closure*: if it is caused, it has a sufficient physical cause (Kim 1989, p. 43). Hence, we should pay close attention to the fact that "our best neuroscience informs us that the changes to musculature that constitute our behavioural responses are wholly determined by the intrinsic properties of the brain to which they are causally connected" (O'Brien this collection, p. 2). Thus it is not only assumed that the brain *can be* causally efficacious of behaviour in virtue of its intrinsic properties, but also that w*henever* behaviour is caused, it is *always* caused in the brain and in virtue of the brain's intrinsic properties. Yet mental phenomena, which are of a non-physical kind, are also mentioned in (i) as a cause of behaviour. But with brain states already providing a sufficient cause for behaviour, what role in causation can they possibly play (Kim 1989, pp. 43–44)? As long as we cannot answer question, we are forced to reject the possibility that behaviour is *over-determined,* or, in other words, we are forced to accept both mental phenomena and brain states as two *separate* causes of behaviour. So causally efficacious mental phenomena should be reducible to the physical cause already provided by the states of the brain, or we must conclude that they are not a cause at all. In the latter case, this would mean that we would have to deny "that our beliefs and desires, and our perceptions and thoughts ultimately have a causal impact on our behaviour" (O'Brien this collection, p. 1) and we would have to accuse every discipline accepting this tenet—O'Brien names folk psychology and the computational theory of mind (ibid., p. 2)—of operating with a faulty ontology, pointing out causes that do not really exist.[5]

Now we have made explicit a metaphysical constraint that was left implicit in the formulation of the content causation problem: mental properties and all their capacities, e.g., their capacity to cause behaviour, must be reducible to properties of the physical brain. Knowing this, we see where the supposed inconsistency comes from: we traditionally think of representational content not as determined by the brain's intrinsic properties, but rather as determined by what it is about (O'Brien this collection, p. 2–3). But if the content causation problem dares us to integrate these two things, namely the description of mental phenomena as causing behaviour in virtue of their representational contents and our theory of the same behaviour being caused by processes in the brain in virtue of the brain's *intrinsic* properties, then we might conclude with O'Brien:

> A solution to the content causation problem requires something that prima facie appears impossible: an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain. (this collection, p. 3)

This is what turns the content causation problem as formulated by O'Brien from an interesting question into an urgent philosophical problem. The apparent impossibility of a solution relies on the idea of a sharp distinction between relational and intrinsic properties. If our best shot at understanding whatever we describe as the "relational character" of representational content is to understand it as a relational property,[6] then its irreducibility to intrinsic properties of the brain is built into it—and so is the insolvability of the content causation problem, given the metaphysical constraint just men-

5 The threat lurking in the background is, of course, eliminative materialism (cf. Churchland 1981).

6 Remember that this kind of property is even referred to as "not determined by the intrinsic properties of the brain" (O'Brien this collection, p. 2).

tioned. Nevertheless, O'Brien aims to provide a solution.

## 4 The relational character of representational content

We see now that an attempt to solve the content causation problem must address the question of how the specific character of representational content can be analyzed in a way that invokes only the intrinsic properties of the brain (instead of being understood as "being a relational property and not an intrinsic property of the brain"). However, O'Brien advertises a theory of content determination that draws on second-order similarity between mental vehicles and the outside world as necessary for a solution to the content causation problem. In fact, he admits that "all theories of mental representation, in their efforts to explain the relational character of mental content, are forced to invoke world-mind relations of some kind", where the latter term seemingly refers back to "relations that incorporate factors beyond the brain" (O'Brien this collection, p. 12). But how does this relate to the explicit goal of providing "an explanation of the *relational* character of mental content that invokes only the *intrinsic* properties of the brain", which would only "prima facie [appear]" to be, but not—as the content causation problem is supposed to have a solution—actually *be* "impossible" (O'Brien this collection, p. 3)?

The answer to this might lie in a view which can be found in O'Brien & Opie:

> Von Eckardt observes that the triadicity of representation in general, and mental representation in particular, can be analysed into two dyadic component relations: one between representing vehicle and represented object (which she calls the *content grounding* relation); the other between vehicle and interpretation [...] This suggests that any theory of mental representation must be made up of (at least) two parts[7]: one that explains how the content of mental vehicles is grounded, and a second that explains how they are interpreted. (2004, p. 5)

When O'Brien writes that every theory of representational content, including his own, must make use of factors extrinsic to the brain, he most likely refers to the content grounding relation—in his case, second-order resemblance. Yet when he promises us "an explanation of the relational character of mental content that invokes only the intrinsic properties of the brain" (O'Brien this collection, p. 3), the scientific explanation mentioned most likely refers to the other part of the theory of mental representation: the internalist theory of interpretation. If O'Brien takes it that only this theory, which provides us with a reconstruction of the causal processes involved in mental representation, needs to be presented in terms of intrinsic properties of the brain, then he provides an account within the "narrow content program" (ibid, p. 3): this research program accepts the thesis that "[t]he representational contents of mental phenomena are not determined by the intrinsic properties of the brain" (ibid., p. 2) but— quite plausibly, I think—relaxes the metaphysical constraint made explicit in section 3 insofar as it only demands "an account of mental phenomena according to which *(at least the causally relevant component of) their representational properties* are determined by intrinsic properties of the brain" (ibid., p. 3, my emphasis).

If this is a correct reconstruction of O'Brien's steps towards a solution to the content causation problem, then he has reached his goal if he:

a) has provided an account of the causally-relevant components of representation that makes use of only the intrinsic properties of the brain, and
b) can make sure that this account still deserves to be called an account of representation, i.e., captures the specific characteristics of representational content that we have so far called "relational".

O'Brien claims that "[t]he insight offered by triadicity is that the relational character of mental

7 A third relation that one might want to look at when "taking apart" the triadic relation of representation is the relation between interpretation and represented object.

content is to be discharged ultimately in terms of our behavioural dispositions towards features of the world" (this collection, p. 12). While it might not be clear at first sight why this provides a solution to the problem at hand, I am convinced that a view of dispositions as second-order properties—such as the property of having a property that becomes relevant for the causing of a certain manifestation once a certain stimulus is provided—helps us to see how O'Brien's account provides a solution. I hope to have shown in section 2 how the adoption of the view that dispositions are second-order properties fits into his picture of representation. Thus I believe that it allows us to regard the first of the two requirements mentioned above as fulfilled: I see no reason why such a second-order property should not be understood as an intrinsic property of the brain. Furthermore, I hold that this view allows us to regard the second requirement as fulfilled, too: the dispositions in question seem to deserve the label "relational" insofar as, when combined with a certain stimulus, they are manifested in terms of overt, observable behaviour of a biological organism. When so manifested, they turn into concrete chains of events linked by causal *relations.* One can now argue that these potential relations are what let us intuitively characterize representation as relational. So understood, O'Brien's explanation does justice to the project of providing an account that captures the specific character that makes representational content deserve the label "representational", but without characterizing its causally efficacious components as being relational properties.

## 5 The content causation problem and second-order resemblance relations

If this is to be seen as a successful analysis of representational content in terms of intrinsic properties of the brain, we can conclude that the triadic picture alone—with its analysis of the relational character of representational content in terms of dispositions—already solves the content causation problem. Hence, this problem, by itself, offers criteria that could be turned into an argument for or against any specific theory of content determination.

Such a theory, as I understand it, serves two purposes: it explains how the contentful vehicles of which the theory of interpretation makes use are individuated, and it explains "how relations between mental vehicles and their represented objects can endow subjects with the capacity to respond selectively to those very features of the world" (O'Brien this collection, p. 6). It thus provides the background information necessary for understanding why the theory of interpretation is able to do what is required of it. The second desideratum is only made clear if the triadic account of content causation is adopted. According to O'Brien, it can only be fulfilled if we adopt the resemblance theory of content determination, because "[t]he obtaining of causal relations between external conditions and inner vehicles cannot explain how the latter endow systems with the capacity to respond in a discriminating fashion towards the former" (ibid., p. 12), whereas "resemblance does offer some prospect of a solution to the content causation problem. The key here is that the mere obtaining of the resemblance ensures that the former have properties that can be exploited to shape the behavioural dispositions of cognitive systems towards the latter" (ibid., p. 12).

Let us recapitulate the steps that took us from the content causation problem to the second-order resemblance theory of content determination. The content causation problem motivates the triadic account of representation if we assume that it is a problem about reduction and if we assume that there is a sharp distinction between relational and intrinsic properties that forbids an analysis of the former in terms of the latter, thus preventing the reduction of a theory of causally efficacious mental phenomena to a theory of brain-based causation of behaviour. The triadic account of representation solves this problem by showing us that we need not assume that representational content owes its specific character to relational properties. Dispositions, understood as non-relational second-order properties, do justice to our concept of "relational character". The triadic ac-

count then needs to be enriched by a theory of content determination, and its use of the concept of a "disposition" leads to a new requirement: to explain how inner vehicles can enable a subject to respond selectively towards external objects. Supposedly, only second-order resemblance can fulfill this requirement. Thus understood, a second-order resemblance theory of content-determination is only indirectly relevant to a solution of the content causation problem.

Yet I would like to point out a way in which the second-order resemblance theory itself relates to the content causation problem. Second-order resemblance between vehicles and objects tells us that there is a mapping from objects to vehicles that is pattern-preserving or, in other words, some objects and some vehicles are alike in some of their relational properties. Nevertheless, the kind of pattern involved is to be "sustained by constraints inherent in the vehicles, rather than being imposed extrinsically" (O'Brien this collection, p. 11, footnote 11). The relations constituting a structure or pattern collectively supervene on the distribution of intrinsic properties of objects and vehicles, although individual extrinsic (and specifically relational) properties of objects and vehicles do not. This strategy of explanation is in principle also available to our understanding of content: contents, fixed by a structural organisation of vehicles, are relational in the same, completely unproblematic sense. Representational contents may not *individually* be determined by intrinsic properties of the brain, but there is a sense in which they are so collectively. But this might count as evidence against the second thesis of the content causation problem: that "[t]he representational contents of mental phenomena are not determined by the intrinsic properties of the brain" (ibid., p. 2). One might then even say that content is *not relational at all*, and that the puzzle that actually troubles us is the question of how representations can have something to which they are applied, namely a "*target*" (Cummins 1996, p. 8).[8] This could still be accounted for by the triadic picture of representation, but it would then not amount to *solving* the content causation problem, but to rejecting it.

## 6 Conclusion

In his target article, Gerard O'Brien addresses the question of "how the specifically representational properties of mental phenomena can be causally efficacious of behavior" (this collection, p. 12). When he does so, there are two parts of the problem to be considered: the first is explaining how mind matters, and the second is showing how an answer can prevail in the light of the content causation problem. Considering the first part in isolation, O'Brien provides an interesting answer. He translates our talk of representation into causal vocabulary, thereby making it possible to reach a concept of causally efficacious representational content. In order to understand how O'Brien's account needs to be assessed with regard to the second part, one first needs to reconstruct which background assumptions make the content causation problem so pressing.

I am convinced that the issues of *reduction* and the *relational/intrinsic property distinction* need to be addressed in order to understand whether and how the content causation problem can motivate an account like O'Brien's. His account takes as a starting point that representational content has a relational character, but should not be understood in terms of relational properties. Rather, as we have seen, it should be understood in terms of dispositions—which can, if manifested, establish causal relations, but are not relational by themselves. I hope to have provided a reconstruction of how this starting point is used to reach the conclusion that, as O'Brien formulates it, "resemblance theories appear obligatory, since they alone offer some prospect for explaining how mind matters" (this collection, p. 12).

If this is correct, there remains one question: whether resemblance theories of the proposed kind might themselves indicate that the content causation problem rests on a mistake. The problem presupposes further problems about the role of relational and intrinsic proper-

8 I owe this point to an anonymous reviewer.

ties that need not be addressed in order to account for the causal efficacy of representational content. The content causation problem's not arising in the first place would, of course, not undermine O'Brien's highly interesting account. It is only that this problem could no longer be used to motivate the argumentative steps he takes. Still, his account is illuminating for many other reasons, such as translating mysterious talk about "being about" into naturalistic terminology. However, whether we can regard the content causation problem as solved or rather as successfully rejected is not clear; but instead of worrying about this problem, we might now turn towards the details of O'Brien's account. An interesting starting point for such further inquiry might be to try to reach a better understanding of the role and the kind of *dispositions* and *vehicles* involved in causal processes, for they form two of the key concepts in O'Brien's theory.

## References


Choi, S. & Fara, M. (2014). Dispositions. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2014 Edition)*. http://plato.stanford.edu/archives.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78* (2), 67-90.

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.

Davidson, D. (1970). Mental events. In L. Foster & J. W. Swanson (Eds.) *Experience and theory* (pp. 79-101). Atlantic Highlands, NJ: Humanities Press.

Kim, J. (1989). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, *63* (3), 31-47.

O'Brien, G. (2015). How does mind matter? - Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

O'Brien, G. & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In H. Clapin, P. Staines & P. Slezak (Eds.) *Representation in mind: New approaches to mental representation* (pp. 1-20). Amsterdam, NL: Elsevier.

# Rehabilitating Resemblance Redux

## A Reply to Anne-Kathrin Koch

Gerard O'Brien

Anne-Kathrin Koch's insightful commentary places a great deal of pressure on the connection between my deployment of the triadic analysis of representation to solve the content causation problem and my contention that it makes mandatory the rehabilitation of the resemblance theory of mental content determination. She argues that if the relational character of mental content can be captured in terms of brain-based behavioural dispositions, as I claim, then this manoeuvre in its own right solves the content causation problem and hence offers no support for resemblance or any other theory of content determination. In this reply, I argue that the relation between the proposed solution to the content causation problem and the resemblance theory of content determination is stronger than Koch allows.

**Keywords**
Content determination | Mental causation | Mental content | Mental representation | Resemblance

Author

Gerard O'Brien
gerard.obrien@adelaide.edu.au
University of Adelaide
Adelaide, Australia

Commentator

Anne-Kathrin Koch
anne-kathrin.koch@gmx.de
Johannes Gutenberg-Universität
Mainz, Germany



## 1 Introduction

There is a paradoxical air surrounding mental content. On the one hand we take it to be a localized property of our minds—of our mental states—distinct from the world in which we are embedded. Yet on the other hand, it is the means by which our minds reach out and make "cognitive contact" (Kriegel 2003) with this surrounding environment. How is such action-at-a-distance possible? The standard solution to this conundrum is to assume that the **relational character** of mental content can be explained by the fact that mental content is a **relational property** of our mental states. This line of thought leads to **content externalism**, according to which mental content is determined in part by factors beyond our heads. But once content externalism is combined with a couple of unexceptional theses about (i) the role of content in mental causation and (ii) the brain-basis of the causal determinants of behaviour, we encounter the **content causation problem**—the problem of explaining how the content of mental states can be causally efficacious of behaviour when it doesn't supervene on what's in our heads.

The solution I offered in my target paper was to sever the connection between the relational character of mental content and the as-

sumption that the latter is a relational property of our mental states (O'Brien this collection). My suggestion was that unlike a **dyadic** story that seeks to explain representation solely in terms of relations between vehicles and their represented objects, a **triadic** account of representation opens up space to explain the relational character of mental content in terms of brain-based behavioural dispositions—specifically, dispositions to respond selectively to specific features of the external environment. According to this triadic account, the **aboutness** of mental content is not some mysterious relational property that brings our minds into contact with various aspects of the surrounding environment; it is the relatively straightforward cognitive capacity, bestowed by the intrinsic properties of our brains, to regulate our behaviour in response to specific environmental conditions.

In her insightful commentary, Anne-Kathrin Koch, after carefully rendering explicit some of the background assumptions on which I rely, focuses on the connection between the proposed solution to the content causation problem and my further contention that it makes mandatory the rehabilitation of the resemblance theory of mental content determination (Koch this collection). Her counter claim is that if the relational character of mental content can be successfully captured in terms of the brain's behavioural dispositions, then this manoeuvre in its own right solves the content causation problem and hence offers no support for resemblance or any other theory of content determination. In this reply, I will show that the relation between the proposed solution to the content causation problem and the resemblance theory of content determination is stronger than Koch allows.

## 2 Rejecting resemblance (and content causation)

The great insight of Charles Sanders Peirce's analysis of representation is his claim that aboutness can't be explained solely in terms of relations between representing vehicles and represented objects (Hardwick 1977). Instead, vehicles are about their objects in virtue of having a certain kind of effect on a cognitive subject—specifically, vehicles either trigger thoughts about objects (in cases of public representation) or they engender behavioural dispositions towards them (in cases of mental representation). According to Peirce, it is this additional relatum—known as **interpretation**—that renders representation triadic.

But once interpretation is added into the representational mix, it has the potential to overwhelm any content-grounding relations that may obtain between vehicles and objects. This is the thread that Koch astutely pulls on in her commentary. To the extent that one can appeal to the manner in which a representing vehicle modifies a subject's behavioural dispositions in order to capture the relational character of content, it seems as though one can also appeal to these dispositions to fix the content of this vehicle. In short, the triadic account would appear to make those theories of content determination that appeal to vehicle-object relations—such as resemblance—redundant (or, at least, "only indirectly relevant", as Koch charitably puts it; this collection, p. 8).

Koch is not alone in drawing out this consequence from the triadic nature of representation. It is precisely this idea about the role of behavioural dispositions in content fixation that forms the foundation of the instrumentalist approach to mental representation that Daniel Dennett has defended over many years (1978; 1987). Dennett was one of the early proponents of triadicity, insofar as he argued that it was only in virtue of their roles in cognitive systems that representing vehicles can be interpreted as bearers of information:

> There is a strong by tacit undercurrent of conviction [...] to the effect that only by being rendered explicit [...] can an item of information play a role. The idea, apparently, is that in order to have an effect, in order to throw its weight around, as it were, an item of information must weigh something, must have a physical embodiment [...]. I suspect, on the contrary, that this is almost backwards. [Representing vehicles]... are by themselves quite inert as

> information bearers [...]. They become information-bearers only when given roles in larger systems. (Dennett 1982, p. 217)

Dennett has also famously argued that the consequence of taking the triadic account seriously is the rejection of any story that takes mental content to be determined independently of a cognitive creature's patterns of behaviour.

Dennett's instrumentalist approach to mental representation, however, has another famous consequence. If the full burden of content determination falls on the shoulders of interpretation—if, that is, it is a cognitive system's behavioural dispositions ultimately fix the content of its representing vehicles—then content is a product of cognition, not an ingredient, and hence cannot be casually implicated in the production of behaviour.

This last point, of course, represents Dennett's own solution to the content causation problem: he abandons the thesis that mental phenomena are causally efficacious of behaviour in virtue of their representational contents (see O'Brien this collection, fn. 1). This is also what Koch is hinting at when she indicates that my proposal to invoke the triadic account of representation might be better interpreted as **rejecting** the content causation problem rather than **solving** it (Koch this collection, p. 8). That is, far from showing that rehabilitation of the resemblance theory of content determination is mandatory, her (implicit) objection is that my proposed solution to the content causation really shows that there is no such thing as content causation in the first place.

## 3 Rehabilitating content causation (and resemblance)

To reiterate, the problem associated with the triadic analysis of representation is that once behavioural dispositions are invoked in order to explain the relational character of mental content, they threaten to overwhelm any other story about mental content determination. But if mental content is determined by such behavioural dispositions, it can't play a robust causal role in their production. In short, the triadic account seems to suggest that there is no content causation **problem** because there is no **content causation**.

In this context, however, it is pertinent to note that, despite his insistence that representation is triadic, Peirce expends a good of effort investigating the relations between representing vehicles and represented objects. His analysis of public forms of representation famously yields three different kinds of vehicle–object relations —convention, causation, and resemblance—associated with symbols, indexes, and icons, respectively (see Hardwick 1977 and Von Eckardt 1993, Ch. 4). If content determination is ultimately just a matter of interpretation, why would Peirce have been so bothered about these vehicle–object relations?

The answer, of course, is that Peirce was concerned not just with the fact that public representing vehicles effect interpretations in cognitive subjects, but with **how** they do so. The point here is that interpretation isn't magic —it requires explanation. Consider, for example, Leonardo da Vinci's **Mona Lisa**. According to the triadic story, the painting that hangs in the Louvre is not about anything on its own. Its standing as a representing vehicle hinges on its capacity to trigger interpretations in cognitive subjects. When we look at this painting it causes us to think about a dark-haired woman with a famously enigmatic smile. But what is it about this painting that endows it with this capacity? Part of the explanation here invokes our recognition of the resemblances between the painting and a woman with a certain kind of physical appearance. The painting wouldn't have the same impact on us if these resemblances didn't obtain. So a complete account of the painting's aboutness must go beyond the fact that it triggers certain thoughts in us to include an explanation of how it does so. And it is here that vehicle–object relations such as resemblance are compulsory.

The general lesson to take away from this (far too brief) analysis is that the interpretation of public forms of representation cannot be disconnected from the cognitive subject's (conscious or unconscious) recognition of what are generally known as **content grounding** rela-

tions between vehicles and represented objects. And what goes for the interpretation of public representing vehicles also goes for the interpretation of mental vehicles. On the triadic story being entertained here, mental vehicles, just like the **Mona Lisa**, aren't about anything considered in isolation. Their aboutness is a consequence of the multifarious behavioural dispositions they create in us towards selective features of the world—dispositions to physically interact with these features, for example, or to make utterances about them. Since this form of interpretation likewise isn't magic, a complete account of mental representation must explain how mental vehicles establish these behavioural capacities. And just as with the case of public representing vehicles, it is impossible to do this without recourse to content-grounding relations (something that is demonstrated by even exceedingly simple representation-using devices such as the humble thermostat—see O'Brien this collection, pp. 7–9).

Precisely because content-grounding relations must be invoked to explain how the former endow cognitive systems with behavioural dispositions towards the latter, content causation is back in business. But what kind of vehicle–object relations can turn this trick? This, of course, was one of the central questions that animated much of my discussion in the target paper (O'Brien this collection). Of the three grounding relations that Peirce found to be implicated in public forms of representation—convention, causation, and resemblance—the first is widely assumed to be unavailable for mental representation since it violates the naturalism constraint.[1] Despite is popularity in contemporary philosophy, the second, I argued at some length, is actually powerless to explain how mental vehicles create the requisite behavioural dispositions (this collection, pp. 6–7). This just leaves us with resemblance. Fortunately, this third vehicle–object relation is up to the task, or at least so my argument went, since the structural properties of mental vehicles that ground second-order resemblance relations can be exploited to shape the behavioural dispositions of a cognitive system towards worldly objects (this collection, p. 8). This is where the rubber of resemblance meets the road of content causation. And it is why a resemblance theory of content determination is mandatory if we are to explain why mind matters.

1 This is the requirement that mental representation be explained without appeal to further forms of representation. If a vehicle is related to its object by convention, the cognitive subject must deploy a *rule* that specifies how the vehicle is to be interpreted. In the case of non-mental representation, where for example the vehicle is a word in a natural language, the application of such a rule is a cognitive achievement that must be explained in terms of processes defined over mental representing vehicles. When this same account is applied to mental vehicles, therefore, it would seem to generate an infinite regress of further representing vehicles, and hence interpretation is never achieved (see Von Eckardt 1993, p. 206).

## References


Dennett, D. (1978). *Brainstorms.* Cambridge, MA: MIT Press.

——— (1982). Styles of mental representation. *Proceedings of the Aristotelian Society, New Series*, *83*, 213-226.

——— (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Hardwick, C. (Ed.) (1977). *Semiotics and significs: The correspondence between Charles S. Peirce and Victoria Lady Welby.* Bloomington, IN: Indiana University Press.

Koch, A.-K. (2015). Does resemblance really matter? – A commentary on Gerard O'Brien. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Kriegel, U. (2003). Real narrow content. *Mind and Language*, *23* (3), 304-328. 10.1111/j.1468-0017.2008.00345.x

O'Brien, G. (2015). How does mind matter? - Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA: MIT Press.