The Journal of Mind and Behavior
Spring 2021, Volume 42, Number 2
Pages 109–138
ISSN 0271–0137

How Reductive Analyses Are Confused and How to Fix Them: A Critique of Varitel Semantics

Nancy A. Salay

Queen's University

The "problem of intentionality" from the vantage point of a representational understanding of mind is explaining what thoughts and beliefs are and how they guide behaviour. From an anti-representationalist perspective, on the other hand, on which cognition itself is taken to be a kind of action, intentionality is a capacity to engage in *behaviour* that is meaningfully directed toward or about some situation. That these are not in fact competing insights is obscured by the representational/anti-representational framing of the debate. This paper begins the work of shifting the conversation in two ways: (1) by arguing that it is the commitment to *internal* representations, not the acknowledgement of a role for representation per se, that is problematic; and, (2) by describing an alternative, externalist, representational approach that draws on extended, embodied, enactive insights.

Keywords: representation, content, extended, intentionality

Unhelpfully, the concept of representation continues to be divisive in the cognitive sciences. According to the mainstream view, cognition is most fundamentally a representational process: to think is to think *about* things, to use representations to plan, problem solve, and understand one's world. Anti-representational challengers, on the other hand, having their philosophical roots in phenomenology[1] and pragmatism,[2] do not separate out cognition from action in this way: to think is to engage in *behaviour* that is meaningfully directed toward or about some situation. As these foundational intuitions seem incommensurable,

I would like to thank Raymond Russ for his cheerful editorial assistance with this paper. Correspondence concerning this article should be addressed to Nancy A. Salay, Ph.D., Department of Philosophy, Queen's University, Kingston, ON, K7L 3N6, Canada. Email: salay@queensu.ca

[1] Especially Husserl (1913/1989), Heidegger (1927/1962), and Merleau–Ponty (1945/2012).

[2] Shaun Gallagher (2017, Chpt. 3) argues persuasively that pragmatist influence on phenomenological and hence enactivist ideas has been systematically overlooked. He highlights the work of Charles Sanders Peirce (1958), John Dewey (1896, 1916, 1938), and George Herbert Mead (1938) as particularly relevant sources of enactive and extended thought.

there is little reason for either side to regard the critiques or insights of the other. They are ignored, however, at the cost of general theoretical progress.

Anti-representationalists (Chemero, 2009; Dreyfus, 2002, 2007; Freeman, 2000; Gallagher, 2005; Keijzer, 1998; Varela, Thompson, Rosch, 1991) are right to reject Cartesian, internalist representationalism with its fixation on mental phenomena and indifference to critical aspects of embodiment, environment, and the interactions that take place between them. Representationalists, on the other hand, are right that a capacity to use representations is an essential feature of the high-level cognitive behaviour that modern humans exhibit. We use second-order concepts — *route A is shorter than route B*; we make inferences — *since route B is shorter than route C, route A must be the shortest of all;* we make plans — *I will take route A*. This behaviour requires a sensitivity to information that extends beyond what is available in immediate perception and thus requires representation of that information. Precisely how this capacity is implemented remains to be seen, but for a theory of cognition to be comprehensive it will need to account for this representation-use behaviour. So long as we approach the problem with a rejection of representation full stop, we will not be able to do so. Furthermore, if we do not offer viable alternatives to Cartesian-inspired ones, they will continue to be generated, as the recent flurry of analytic proposals for "naturalising content" testifies (Burge, 2005; Neander, 2017; Ryder, 2004; Shea, 2018; Skyrms, 2010) despite important critiques of them (Chemero, 2009; Clark and Toribio, 1994; Dreyfus, 2002, 2007; Freeman and Skarda, 1990; Keijzer, 1998; Van Gelder, 1995).

Theoretical progress, then, requires that we move beyond the representation/anti-representation framing of the debate to a more nuanced vantage point from which we acknowledge and attempt to account for the role of representation in cognitive behaviour while taking seriously the idea that the physical and social environments within which an individual develops play substantive, capacity-determining roles. This paper begins the work of shifting the debate in two ways: (1) by arguing that it is the commitment to *internal* representations, not the acknowledgement of a role for representation per se, that is problematic; and, (2) by describing an externalist approach to explaining representation-use.

The paper unfolds in two parts. First, I argue that the general strategy of explaining mental representation — "content" — by appeal to the conditions under which relevant internal physical mechanisms function as content vehicles is deeply misguided. Nicholas Shea's (2018) varitel semantics redresses the weaknesses of theories of content that have gone before (Dretske, 1995; Fodor, 1990; Millikan, 1984) and thus is ideal soil within which to ground my more general critique of what I shall call the internalist content reduction (ICR) strategy. Second, I outline an externalist approach that complements and expands on some of the ideas that extended mind theorists have already put forth (Clark, 2006a, 2006b; Logan, 2007; Wheeler, 2004) and draws on the insights that have been gained from embodied, enactive cognitive science. Content externalism places representations front and centre — this is the

central topic of the account after all — but it is developed with an eye to future integration with wider, enactivist theories of intentionality such as Gallagher's (2017).

ICR and Its Problems

Although Shea takes varitel semantics to be a *non*-reductive, physicalist account of content, because it purports to establish supervenience relations between content and vehicle rather than identity relations (Shea, 2018, p. 41), it treats content specification as a matter of mapping outer patterns to inner physical mechanisms, and thus, for the purpose of this critique, it is a model exemplar of the ICR approach. That it is partly externalist in the sense that it espouses a wide, relational view of content (2018, p. 39) does not alter the overall strategy of finding appropriate mappings to *internal* vehicles of content.

As a methodology for rigour in content attributions, varitel semantics is a valuable contribution to the field of cognitive science, where a set of clearly defined constraints around representation ascriptions is sorely lacking. Despite this, because it is an ICR account, varitel semantics ultimately fails.

Varitel Semantics

Shea's ultimate aim is to provide a physicalist grounding for conscious mental activity, but, recognizing that consciousness brings with it a host of confounding factors and that "the nature of representational content remains puzzling even in non-doxastic, non-conscious cases" (2018, p. 25), Shea accordingly moderates his goal to one of identifying all and only those system *behaviours* for which "representational content … offer[s] a better explanation of behaviour than … a 'factorized' explanation can provide" (2018, p. 30). Although Shea works hard to develop a rigorous methodology for content ascription that is not guided by intuition (2018, p. 28), the general approach is in fact guided by the foundational intuition of the representational theory of mind (RTM), that cognitive systems are not (only) caused to act by proximal, local, factors, but are also caused to act, in a more complex way, by distal factors — reasons, plans, and so on. Ultimately, varitel semantics is developed with an eye to being sensitive to this feature of cognitive behaviour so that it will be a useful tool for distinguishing cognitive from non-cognitive behaviour. He is also careful to navigate a middle course between, on the one hand, overly general information theoretic accounts (Dretske, 1981) on which functionally relevant structural/co-relational isomorphisms are sufficient conditions of content and, on the other, overly restrictive teleosemantic accounts (Millikan, 1984) on which only systems capable of "consuming" content-bearing states have them. For an example of the former problem, obviously non-cognitive systems such as thermostats fulfill the content condition because their bi-metallic strip/circuit mechanisms have the functional role of "carrying information" about

the ambient temperature of the surrounding air. With respect to the aim of using the content concept to distinguish between cognitive and non-cognitive systems, this is unhelpful. According to teleosemantic views, on the other hand, neurons are excluded from being potential content-bearing entities since "… it is very hard to see a principled way to identify representation consumers in the brain, if consumers are devices whose output fixes content (Cao, 2012)" [Shea, 2018, p. 19]. Shea's concern here is that if we deny a representational role to neurons, most of the representational identifications made by cognitive scientists would be excluded.[3] He is careful, then, to preserve the insights of both while correcting these key problems.

Broadly put, the general varitel semantics recipe for determining whether a physical process bears content within a system S is that it be a structural or co-relational isomorphism that plays a functional role in the context of S's teleologically robust goals. To make the notion of a teleologically robust goal more rigorous, Shea introduces two concepts: task function and robust outcome.[4]

A task function is a functional mapping, implementable in multiple ways, that is performed by a system whole. For example, the task function of mapping each element of {A, B, C} onto {1, 2, 3} can be implemented by {A-1, B-2, C-3} or {1-A, 2-B, 3-C} and a system could perform either of these implementations in several ways, with pencil and paper, with physical symbol tokens, or, as Shea argues, by way of an internal mechanism. Each of these performance methods is an "algorithm," a *concrete* representational process, "… a way of processing representations that is realized in some organism or other system" (2018, p. 34). An internal mechanism is justifiably seen as an algorithm for implementing a task function only when doing so provides a better explanatory account of system behaviour than a factorized causal one would provide. For example, while we *could* treat the task function of a thermostat to be to track the ambient temperature of the surrounding air by viewing the isomorphisms between various states of the bi-metallic strip/circuit mechanism inside it as implementing an algorithm for it, doing so would not provide any more explanatory bite than a simple, proximal causal chain account would, i.e., we wouldn't be able to predict or explain any more on the representational ascription view of the thermometer than we would without it.

To make this distinction between explanatory/not-explanatory representational ascriptions a principled one, Shea introduces the concept of a robust

[3] As I shall argue, though they need not be excluded, the theoretical role of such representational ascriptions will need to be revisited.

[4] Shea introduces other concepts as well, to further flesh out the notion of a task function — "stabilized function" (2018, p. 64), for example. As discussion of these and similarly nuanced distinctions would take us in a direction orthogonal to the central goal of this section, namely, showing that the internalist strategy is wrong-headed, I omit such details.

outcome, a goal situation that a system pursues with persistence and in the face of obstacles. Tracking temperature cannot be construed as a robust outcome for thermostats since, not only do they not persist in it in the face of obstacles, they are not capable of such persistence: if a part breaks, the whole breaks. Thus, there is nothing the thermostat-whole does that is not exhaustively explained by a description of its parts.

In systems which do perform task functions robustly, however, representational explanations make connections between occurrent behaviour and distal effects, connections that cannot be made via factorised causal chains. To use an example of Shea's (2018, p. 124), a rat might perform the task function of reaching a certain location L — in pursuit of a food reward perhaps — in a variety of ways and in the face of obstacles: it might take the shortest route, a circuitous route, or an avoidance route, depending on the features of the situation at hand. By treating the relevant neural isomorphisms that are engaged in the course of this activity as representational components of an algorithm the rat "uses" to perform the task, we not only explain how the rat performs its actions, we also explain *why* the rat moves through the maze: to reach L. Factorised, causal explanations, in contrast, do not connect the robust outcome with the occurrent behaviour at all since, from the proximal perspective, L is just another point along the causal chain. Worse, such explanations cannot connect patterns of neural activity with patterns of system-level behaviour since patterns are invisible from the subpersonal perspective:

> An organism interacts with the environment, bringing about distal effects in its environment, doing so by reacting to distal objects and properties in the environment. There are real patterns in the environment and an agent's interactions with it that would be invisible if we looked only at intrinsic properties of the agent. (Shea, 2018, p. 32)

Shea's claim here is that the rat's personal-level, goal-seeking behaviour is inexplicable unless we see its relevant inner states as "real" physical algorithms for fulfilling task functions. We should analyse isomorphisms in representational terms, then, when the system in which they are found "exploits" them in order to fulfill a task function that it pursues robustly: "Correlations turn into content when they are exploited by a system — exploited in a very particular sense …. the content-constituting correlations are those which unmediatedly explain a system's performance of its task functions" (2018, p.110).

One conclusion we can draw from varitel semantics is this: in some cases, the ones of interest to cognitive scientists, representational descriptions of subpersonal mechanisms have an explanatory force that factorized descriptions lack. By treating the relevant subpersonal processes as representations, we explain how/why lower-level activity yields goal-oriented behaviour. Shea thinks that varitel semantics yields a further conclusion: the normative superiority of representational explanations in the cognitive context entails realism about internal content

vehicles. In what follows I will pry these two ideas apart: varitel semantics offers a rigorous methodology for making subpersonal representation ascriptions, but it fails in its ICR agenda.

Varitel Semantics from a Broader Context

Shea repeatedly insists that varitel semantics picks out cases in which representational explanations have "proprietary explanatory purchase" (2018, p. 71). But the contrast Shea has in mind here is between representational and proximal causal explanations, not between representational and non-representational explanations *tout court.*[5] Indeed, the central concepts of varitel semantics presuppose a representational framework: "content" is what a representation is about and an "algorithm" is a procedure for computing over representations. Since varitel semantics is being offered from within the representational theoretical paradigm, its success in explaining the relevant data cannot speak to the more general debate between competing theoretical frameworks, between representational and anti-representational ones. Data always underdetermine theory. Thus, while varitel semantics provides a methodology for rigorous content ascriptions from within the representational framework, Shea has not shown that representational explanations do better than those made from alternative theoretical frameworks.

And alternatives are available. The most comprehensive — dynamic systems theory — provides a framework in terms of which entirely non-representational explanations of system behaviour, including the sort of spatio–temporally extended, "goal-directed" behaviour that varitel semantics is designed to pick out, can be given (Haken, 1990; Thelen, Schöner, Scheier, and Schmidt, 2001; Townsend and Busemeyer, 1995; Vallacher, Nowak, and Kaufman, 1994). Even "representation-hungry" (Clark, 1997; Clark and Toribio, 1994) tasks such as mental imaging have been fruitfully accounted for in terms of non-representational, dynamic systems concepts (Van Rooij, Bongers, and Haselager, 2002).

Seen from this broad, multi-paradigm vantage point, the factorized, causal explanations that serve as Shea's normative comparators begin to look like straw men. If we use a low enough level of granularity to identify factorized causal chains, we guarantee that there won't be a useful mapping to the higher-level behaviour we are seeking to explain. This is to be expected when we move between low and high levels of granularity. We rarely find it useful to cite, for example, the molecular structure of a *particular* table to explain a feature of tables generally.

What, then, can the insistence on the normative superiority of representational explanation mean in this theoretically constrained context? The implicit target seems to be ascriptionism (Dennett, 1981; Egan, 2014) — the view which concedes

[5] At least this is what we must assume since, apart from one paragraph (Shea, 2018, p. 205), there is no discussion of anti-representational accounts at all in the book.

the explanatory usefulness of representations but does not take this to entail realism about the content of those representations. Unfortunately, Shea's strategy of comparing representational explanation with proximal, causal chain descriptions covers over the realist/ascriptionist divide which turns on the question of whether vehicles qua representations play a causal role in intentional action. As I show in the next section, in the cases he considers, the relevant causal relations are the factorised, proximal ones of the purported content *vehicles*. If this is right, then content realism is presupposed rather than entailed by varitel semantics. Indeed, Shea's confusing rejection of ascriptionism should make us suspicious that this is so: in the same breath he both applauds the view — "Dennett's ascriptionism is a tenable and substantial naturalistic account of representational content. In this sense we already have a good theory of content" and, without reason or argument, rejects it — "However, I will reserve the term "realism" for accounts that are committed to there being real *vehicles* of content: individuable physical particulars that bear contents and whose causal interactions explain behaviour" (Shea, 2018, pp. 14-15). No ascriptionist would deny that the physical mechanisms we describe in representational terms are *real* physical mechanisms. The point of contention here is the adverbial clause "that bear content." The ascriptionist agrees that representational descriptions of physical mechanisms can play an explanatory role, but Shea wants to conclude further that "As I have been describing the problem of mental content, realism about vehicles is a core part of what it takes to be a mental representation" (2018, p. 15). The penny drops. It is Shea's pre-theoretical commitment to realism about *mental* representations, something that Dennett and other ascriptionists do not share, that grounds his content realism. From the realist vantage point, treating content as merely an explanatory gloss "… would radically revise our conception of ourselves as reason-guided agents since reasons are mental contents" (2018, p. 6). As I have already noted, mental contents according to RTM are a part of the explanandum, the very thing being *accounted for*. Being an anti-realist about mental content at this level is just to reject RTM, as both eliminativism and ascriptionism ultimately do. But notice here that we have shifted from talk about subpersonal mechanisms to talk about the mental, something that Shea explicitly took off the table: "I won't deal with cases where a representation's being conscious is relevant to fixing its content" (2018, p. 26). Without an account of the relation between subpersonal mechanisms and mental contents, however, Shea's realist convictions are just that, convictions. Nowhere does Shea confront this issue. Indeed, since this would amount to an account of the explanatory gap, we should not hold our breaths.

This subtle movement between personal level and subpersonal level content, via content vehicles, hides the worrisome presupposition that personal level concepts can be extended to the subpersonal level at all. That mental representations and the representations we ascribe to subpersonal vehicles are both "content," even though a wide granularity-gulf extends between them, should make us

suspicious of an equivocation. Even simple reflection on the two applications of the content concept reveal that a different sense must be at play in each case. Mental representations are, by definition, essentially contentful: what it is to be a mental representation is to be *about* something. If a mental representation loses its content, the mental representation is gone. The subpersonal vehicles of content, on the other hand, are only contingently contentful (if they are at all): a purported content vehicle could lose its content-bearing relation and continue to play a functional role within its system. Neurons gain and lose synaptic connections in ways that are only sometimes in line with content relations, as the constraints of varitel semantics makes clear. This means that neurons function in ways that extend beyond their purported content-bearing properties. If they are content vehicles, then, they are only contingently so. Mapping between essentially and contingently content-bearing states should make us worry that an illegitimate equivocation has been made along the way. As I argue in the next section, this is precisely the case.

Why the Strategy of Internal Content Reduction Fails

ICR is an ill-conceived programme. To unpack why, it will be helpful to make use of Aristotle's causal analysis of explanation (*Physics* II.2.194a28–II.8.199b33). I will use the example of a sequence of neural activity, call it N, in my review of the four causes.

Suppose that N has been found to be co-related with the presence of a pheromone, for example androstenol, in an organism's environment. We might frame our analysis of N for organism O with the following questions:

1. Material Cause: Out of what material is N made? — N is made up of neurons.
2. Formal Cause: What is the structure of N? — N is a pattern of impulses across a cluster of neurons.
3. Efficient Cause: What is the proximal cause of N, that is, what event or situation causes N to come about? — Olfactory neural activity that is triggered by the presence of androstenol is the proximal cause of N.
4. Final Cause: Why does N occur, that is, what function does the neural activity play in the organism? — N tracks/represents the presence of potential mates for the organism.[6]

In terms of Shea's discussion, factorized explanations are efficient causes, the proximal triggers of the behaviour to be explained. In scientific practice, efficient

[6] Note that for Aristotle, a final cause or "that for the sake of which" is a purposively teleological end — "It is plain then that nature is a cause, a cause that operates for a purpose" (Physics II.8.199b32–33); however, final cause non-purposively understood, in its functionally teleological sense, is an enormously fruitful aspect of biological explanation and this is the sense in which I use it here.

cause is what is typically meant by "cause." The representational explanations that Shea favours, on the other hand, are functional descriptions of subpersonal activity; final causes on the Aristotelian analysis. As different aspects of such an analysis, representational explanations are neither better nor worse than factorized ones; they are complementary, together contributing to a richer account of the behaviour than either one alone would provide. To compare them, as Shea does, is to confuse these two aspects of explanation.

To someone committed to realism about mental content, of course, whether content is a final cause *only* is precisely the point of contention. Realism about content is motivated by the observation that, for intentional agents, final causes *are* efficient causes of system behaviour, not just explanations of them. From this vantage point, varitel semantics individuates intentional and non-intentional systems by way of the content vehicles that serve as efficient causal mechanisms of behaviour. The robustness constraint on task functions does the heavy lifting of distinguishing systems whose proximal behaviour is *directed* toward some distal goal from those whose proximal behaviour is merely consistent with it. Only when a task function is a robust outcome are the proximal actions that lead to it performances of an algorithm that the system has developed to achieve it (Shea, 2018, p. 55). If task functions were merely final causes, there would be no justification for realism about the algorithms — the content vehicles — since the proximal path to the final cause would be causally determined by other proximal causes, not by the final cause itself. In other words, the behaviour would not be directed toward the distal goal.

To see what is wrong with identifying final and efficient causes in this way, consider once again the four causes. As Aristotle himself observed,[7] final cause explanations are not required for some natural processes. To give an analysis of a bead of water as it trickles down a pane of glass, we need only cite material, formal, and efficient causes: nothing beyond the interactions between the molecules of water, their structural properties, and the proximal physical conditions is needed to explain how and why the water moves in the way and direction that it does. But some systems do admit of (some might say "require") further explanation. Organized collectives such as organisms,[8] for example, act as unities that exhibit spatio-temporally extended patterns of behaviour — locomotion, hibernation, predation, and so on — that call for explanation. Because such unities interact with their environments in non-random ways, that is, they learn or evolve to respond in specific ways to specific situation-types, isomorphisms develop between their behaviour and features of their environment. When environmental conditions offer few opportunities for survival, for example, many species develop

[7] Some natural phenomena, such as rain, do not have a final cause; they occur out of necessity (*Physics* II.8.198b18–24).

[8] Non-living examples work here too, e.g., crystalline structures.

responses that exploit the opportunities that do exist — hibernation and migration are two common adaptations. A bear hibernates through the winter months when food is scarce, a strategy for conserving the energy it needs to maintain homeostasis, something it must do to survive. Without the context of this broader explanatory framework, local behaviour such as excessive eating and nest building is mysterious. When we treat hibernation as the final cause of excessive eating during the autumn, however, we connect the proximal and distal behaviours in an explanatorily satisfying way.

Notice that this is not an appeal to purposes or goals: final cause explanations are teleological[9] in that they appeal to ends, but they are not thereby purposive. The bear does not increase its food intake because it is moved by the goal of hibernation, nor does it hibernate because it intends to conserve its energy, nor does a plant grow leaves because it intends to shade its fruit. The bear's and the plant's "actions" are system-level manifestations of a multitude of lower-level proximal causes and effects — sensitivity to subtle changes in temperature, humidity, and light — that together constitute the behaviour. These behaviour patterns have evolutionary staying power not because the individuals exhibiting them are moved by "a drive to survive," again a purposive notion, but because those who do happen to survive pass on their genes and thereby their evolved or learned behaviours along with them.

The same lesson holds for the relationship between neural and personal level isomorphisms. One could, from the scientist's vantage point, notice that a personal level pattern of behaviour, the frog's FLY–FLICK–TONGUE[10] behaviour, for example, is co-related, roughly, with a pattern of neural activity. Functional hypotheses that connect the organism-level behaviour with the neural activity provide satisfying final cause explanations of why the neural activity occurs. Such hypotheses need not appeal to purposively teleological functions, that the neural activity is *directed* by the goal of tracking flies, to gain traction: the neural activity becomes entrenched because those organisms in which it occurs survive to pass on their genes, thereby ensuring the proliferation of the neural–organismal behaviour bundle across the species. As in the bear and plant cases, the efficient cause of the neural activity is the chain of proximal neural events that result from the presence of flies.

This lesson applies also to analyses of artificial neural networks (ANNs) or machine learning systems more generally. A classifier that has "learned" a

[9] We might also call them design stance explanations, following Dennett's scheme of distinguishing explanations according to their level of abstraction: "… the design stance, where one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that it will behave as it is designed to behave under various circumstances" (Dennett, 1981, p. 60).

[10] I will use a convention of all-caps to indicate a generalisation, which might be a concept, a word, or a response pattern.

response, an output usually represented by a real number, to some input, a vector representing some object in the world, such as the pixels in a photograph, will co-classify inputs that share input features. For example, since dogs typically share many features — four legs, fur-covered, two eyes, two floppy or pointy ears, etc. — a network that has learned to classify dog photographs will have become sensitive to variations in these and more fine-grained features, depending on the resolution of the input. Pictures of cats, since they share many dog features, will also elicit some activation of the "dog" output node if the bias threshold is set low enough. This is what is known as "free generalisation" in machine learning: inputs that share (detectable-to-the-system) features will elicit similar responses. Such behaviour, although powerful and useful in myriad application contexts, is not an indication that such systems are now moved by the concepts themselves. Interpretations that describe machine learning behaviour in terms of human concepts — the system has "learned" what a dog is — offer explanatory models of an otherwise confusing morass of arithmetical operations. These interpretations offer final causes that are not purposively teleological: a network that can respond consistently to instances of the class DOG has not, thereby, transcended its architecture to apprehend the class behind the instance. Its output behaviour is the result of myriad training sessions each of which is designed to adapt the system's responses toward consistency with the human concept it is being trained to "learn." As in the bear, plant, and frog cases, the efficient cause of the network activity is the input and the chain of proximal events that is triggered thereby. The final cause plays only an explanatory role.

More generally, final cause explanations draw on spatio–temporally extended patterns of behaviour in the context of which the local behaviour can be understood. Such explanations are rewarding precisely because they connect the dots, so to speak, between proximal behaviour and distal behaviour. But explanatory usefulness cannot turn final causes into efficient ones. The cases that Shea is trying to distinguish, of course, are the ones in which final and efficient cause seem to overlap, in which systems that perform task functions with robust outcomes do so intentionally.

Can "robust outcome" carry this heavy burden? One way to find out is to ask whether a system's robust performance of a task function *could* be a merely functionally teleological final cause of its actions, and therefore not one that should be identified as an efficient cause. According to Shea's definition, "An output F from a system S is a *robust outcome function* of S iff (i) S produces F in response to a range of different inputs; and (ii) S produces F in a range of different relevant external conditions" (2018, p. 55). Consider machine learning classifiers again. Depending on how broadly we specify "range of different inputs" and "range of different relevant external conditions," some classifiers will meet the robustness requirement and others will not. Many can reliably detect the features they have been trained for in a wide range of noisy input situations: when the input vectors

are only partially complete, when the input vectors are complete, but partially incorrect, when the input vector is embedded in a wider context vector. But it is a well-known problem in the field of AI that such systems might display brittleness in the face of certain sorts of noise conditions (Eykholt, Evtimov, Fernandes, Li, Rahmati, Xiao, Atul, Tadayoshi, and Song, 2018). What sorts of noise conditions? Precisely those for which the system has not been trained. Our justification for treating such a system's internal mechanisms as content vehicles turns on what we take the relevant robustness factors to be. But when we stop to consider that, in terms of architecture and behaviour, there is not much difference between a classifier that correctly picks out a feature under one set of robustness conditions and one that does so under another — because each has been trained using different examples — it seems implausible to suppose that the one that meets our conditions pursues its goals while the other that does not merely acts in accordance with them. Robustness of outcome, in other words, could just as well be a function of varied learning conditions, and not, therefore, a sufficient condition of purposiveness.

Nor is robustness a necessary condition of purposiveness. Any number of factors can side-line a seemingly goal-driven activity. Fred wants a piece of cake and gets up to go to the kitchen to get one, but the phone rings and he stops to answer it. By the time the call is over, he has lost his desire for cake or he has forgotten it. Fred moves to the kitchen to get the cake but stubs his toe on the coffee table en route. The pain in his foot wipes away all desire for cake. And so on. Do we say that because getting the cake was not a robust outcome for Fred, it was never pursued with intent? Counterfactuals will not help here. What if Fred is the sort of person who is easily distracted, does not stick to goals, and is generally what we might call flighty? Is he therefore not an intentional being? Robustness of outcome, it seems, cannot shoulder the burden that Shea sets.

But even were there some strategy for rigorously separating out functional from purposively teleological systems, supposing that the distinguishing characteristics will be found in their internal mechanisms — as ICR does — is a mistake: (1) final causes are not candidates for physical instantiation; and, (2) low-level mechanisms are not causally related to system level action. I will examine each of these ideas in turn.

A strategy, such as the one used in varitel semantics, which seeks to identify the conditions under which some subset of a system's physical processes has transformed into an *efficient-cum-final* cause plays, to use Dennett's (2003, p.18) apt description, a "conceptual sleight-of-hand." When we describe neural patterns of behaviour in final cause terms, e.g., FLY tracker, we use broad explanatory abstractions — functional–teleological hypotheses — that come into view only from the scientist's spatio–temporally extended vantage point. The presence of a fly causes a frog to flick its tongue because the fly is a bundle of features to which frogs have an evolved behavioural response. But

though individual frogs behave in ways that march in step with this pattern of behaviour, possibly even in ways that persist in the face of diverse obstacles, the efficient cause of an individual frog's behaviour is always the current situation via its ongoing perceptual interaction with its environment. To call the neural clusters that underwrite this behaviour *vehicles* of content is deeply misleading since content — the concept FLY — is an explanatory heuristic for us qua scientific observers and plays no role at all in the system's occurrent behaviour. The neural mechanisms that we label "FLY tracker" are low-level, cause–effect chains triggered by occurrent features in the environment, not responses to spatio–temporally extended abstractions such as FLY.

Worse, the supposition that efficient causes might be found among internal mechanisms at all rests on a confused picture of the origination of agency. If agent S does something for reason R, then R is (one of) the efficient cause(s) of S's subsequent actions. If mental contents are to function as efficient causes, ICR assumes that some internal content vehicle "carries" this content, that this content vehicle plays the efficient causal role in the picture. For example, if some neural activity N functions as the content vehicle, then N is the efficient cause of S's subsequent actions. But this is a sloppy analysis of how parts contribute to system behaviour. Parts of systems do not *cause* system wholes to do anything; they are constitutive of system behaviour (Craver and Bechtel, 2007). My feet moving forward are not the efficient cause of my forward movement; they are part of what it is for me to move. My lungs contracting and expanding do not cause me to breathe; they are constitutive of my breathing. Likewise, the purported content vehicles are constituents of system behaviour, not causes of it.

Varitel semantics provides a methodology for rigour in our (final cause) content ascriptions, but it does not give us a tool for picking out "real," content-bearing, internal mechanisms. It cannot: the idea that internal mechanisms can "bear" content and thereby function as efficient causes of agent action confuses descriptions with causes on the one hand and constituency relations with causal ones on the other. Contrary to the standard critique, however, this analysis shows that the root of the problem is the insistence on internalist reductions, not the concept of representation.

If the central insight of RTM is correct, that intentional agents act for reasons, an account of how representations qua reasons function as efficient causes of action must be possible. Since an internalist account is not forthcoming, we should consider externalist possibilities. In the next section I provide a sketch of one. Elsewhere (Salay, 2019) I have begun to work through the details of what I offer only an overview of here. Much work remains to be done.

Content Externalism: A Sketch[11]

Content

Shea takes "content" to be "shorthand for the objects, properties and conditions that a representation refers to or is about" (2018, p. 6). As a broad characterisation, this is surely right, but it draws focus away from an important feature of content *use*, namely that it requires interaction with high-level, "intellectual patterns" (Dennett, 1991, p. 41). To understand Shea's example — "… the content of one of my thoughts about dinner is: *each person needs 150g of pasta*" (2018, p. 6) — an individual must, in a way that will need explanation, abstract over a range of possible pasta dinners in which the salient feature is how much pasta is being eaten. We do not need to become bogged down with ontological questions about the nature of these abstractions to appreciate that abstracting is involved when we speak and think, when contents are being *used*. Shea is correct, then, that explaining intentionality requires explaining how contents are determined, but incorrect that determining the content of a mental representation requires explaining how internal thoughts hook on to the things they are about, how the "thought in my head connect[s] up with quantities of pasta" (Shea, 2018, p. 6). The entailment follows only if the assumption that mental representations are fully "in the head" holds. But it does not — to suppose otherwise is the mistake of the ICR approach. The contents central to any inquiry into intentionality are those spatio–temporally extended abstractions that function as efficient causes of actions. To explain how contents are determined, then, we first need an account of how content users come to recognize and develop responses to these high-level patterns. Content fixation and individuation will follow from this practice.

Perception

On all internalist representational theories of mind, perception is the most basic intentional relation: to perceive an object, even non-conceptually, is to have a mental representation with conceptual and/or non-conceptual content. For an externalist account of content to be viable, then, it must begin by offering an alternative to this intentional view of perception, what is called in the literature "representationalism." Fortunately, non-representationalist theories have gained traction over the past decade and, taken as a family of "relational" views (Brewer, 2017; Hurley, 1998; Martin, 2002; Noë, 2010; O'Regan and Noë, 2001; Travis, 2013), offer fertile theoretical ground for content externalism. Common to all is a shift in the kind of process perception is understood to be: rather than see it as

[11] Not to be confused with externalism about the content of experience, which is version of ICR since experiences, e.g., perceptions, are taken to be contentful in themselves.

reconstructive, which invites questions about the nature of the resultant representations, perception is treated as an interactive process, as something organisms *do*. Locomotion and reproduction are similarly active, self-originating processes: organisms locomote, reproduce, and perceive; they do not *have* locomotions, reproductions, or perceptions. Before we can look at how relational perception side-steps the problems of content internalism, we need to get clearer about where those problems gain a foothold on representationalist views of perception. As one might expect from an area that has been under analytic scrutiny for some time, there are many nuanced distinctions made within representationalism and sorting out which are relevant to the view being sketched here will help to illuminate the space I am carving out.

Representationalism divides the concept of perceptual content into representational content on the one hand and phenomenal aspect on the other. Representational content is, very roughly, what a perception is about — my perception of the mug of coffee is *of* the coffee mug — while phenomenal aspect, a.k.a. qualia, is the subjective nature of a perceptual experience — when I look at the mug, my experience is constituted by certain colours, sounds, smells; it *feels* a certain way. Both veridical and non-veridical perceptions can share "common" representational content and sometimes phenomenal aspect as well: both my accurate perception of the real mug on the desk and my mug hallucination are *of* the coffee mug and might *feel* the same way to me. Representationalism thus offers a parsimonious account of perception and misperception since the same representational mechanism is taken to underlie both.

According to the externalist view being developed here, however, that efficiency is purchased at the cost of accuracy: by grounding content in perception, a biologically fundamental capacity, a commitment to ICR is made. To be clear, the rejection of contentful perception is not a repudiation of the appeal to representations that cognitive scientists make in the context of theorizing about perception. As Tyler Burge rightly says, "Most of the explanation … at both the explicitly representational levels and the more 'engineering oriented' algorithmic levels — invokes intentional or representational content" (2005, p. 20).[12] In cognitive science, perception is physically cashed out as a complex, re-constructive, binding process that begins with the stimulation of an organism's sensory receptors. Neurons represent, either singly or in groups, the various features of an organism's environment: low-level edges, colours, and textures as well as high-level inferences about object distance, blind-spot gap-filling and smoothing of saccade smears. So long as we do not confuse these final cause representational ascriptions with the contents that function as efficient causes of behaviour, we need have no quarrel with this use of the representation concept in the cognitive sciences.

[12] Although in treating "representational" and "intentional" synonymously, Burge makes it clear that he is advancing a representationalist approach.

Matters become problematic when representational appeals are called upon to do double duty, however, when the abstract patterns — contents — we ascribe to the lower-level mechanisms in our explanations are, in addition, taken to play an efficient causal role in the actions of the systems themselves, by way of these lower-level mechanisms. This is the ICR strategy, one that I have been arguing is deeply mistaken. Representation might play a fruitful explanatory role in cognitive science accounts of perception and nevertheless not be a source of *intentional* content. Roberta Locatelli and Keith Wilson put it well: "if we take representationalism to offer a theory of the content of conscious perceptual experiences, and not merely the subpersonal mechanisms involved in perceptual processing, … it is necessary not only to show that experiences possess contents, or have accuracy conditions, but that these contents are in some way manifest, or cognitively available, to the perceiving subject at a first-personal level" (2017, p. 209).

When we let go of the idea that perceptions are internal objects to which an organism stands in some relation, the conviction that perception entails a system-level pseudo-understanding[13] of objects perceived also falls away. Letting go of this idea, however, is difficult to do. Even Burge (2005), whose anti-individualism is sensitive to the critical role that external factors play in content grounding, assumes that an ability to perceive instances qua types requires a (presumably biological) capacity to generalise: "To count as a perceptual system, the system must have objectifying capacities.… The proximal light arrays cannot alone (even taken sequentially) suffice to distinguish among different types of possible distal causes. They cannot alone determine a single distal, objective property under different conditions" (p. 10).

On the contrary, in the absence of any system-level, objectifying capacities whatsoever, machine learning classifiers distinguish between different distal causes by processing fine-grained, low-level, inputs. Organism-level actions might co-relate with instances of classes — a cat hisses when a dog is present — but it does not follow from this that the system thereby understands the general. Bill Brewer's "thin" perception nicely captures this distinction: "an object of acquaintance, *o, thinly looks F* iff *o* has, from the point of view and in the circumstances of perception in question, appropriate *visually relevant similarities* with paradigm exemplars of *F*" (Brewer, 2017, p. 216). We might find it useful in our final cause explanations to call "thin perceivers" generalisers, to talk about them as "having concepts," but it is only their behaviour over time that fits such descriptions. In developing responses to bundles of features, organisms act in accordance with the general and this has important effects, e.g., success in obtaining resources

[13] Non-conceptualist representationalism (Block, 2003; Crane, 1992; Peacocke, 2001) in the context of a final cause analysis is to be applauded for extending explanations to language-naive systems such as many non-human animals, but views that take non-conceptual awareness as playing a causal role in behaviour are again making the ICR mistake.

for survival; but at any moment of activity, and without further resources, such systems are merely responding to what is spatio–temporally present.

Relational accounts of perception shift the focus of analysis away from systems narrowly taken and onto system interactions in a way that complements content externalism.[14] On this view, perception is an active and dynamic process according to which an organism interacts with its environment. Perceptual capacity is determined by a host of factors, including embodiment details — e.g., placement, size, manoeuverability of the eyes, ears, nose, mouth — as well as modality-specific, sensorimotor contingencies that guide and constrain an organism's perceptual activity. Perception is the ongoing exercising of these sensory modes of environmental engagement. There are no resultant, static, percepts that a perceiver "has" on this view since "the outside world acts as an external memory that can be probed at will by the sensory apparatus" (O'Regan and Noë, 2001, p. 946). Thus there are no objects of perception — contents — only objects of engagement.

With respect to phenomenal aspect, what a sensory modality "is like" for an organism is largely a matter of the sensorimotor contingencies of the modality in question. For example, during vision the sensory stimulation of retinal photoreceptors will shift and change in lawful ways as eyes rotate. The "flow of information" in consequence will expand and contract as an organism moves forward and backward. Animals with eyes on the backs of their heads or on their knees will have different kinds of visual experiences than do those with eyes on the front of their heads. Phenomenal aspect will vary across modalities as well. Unlike the tactile modality, visual stimuli are received only from the parts of objects that are facing an organism's visual sensory receptors. An organism skilled in vision and touch will develop expectations that accord with the idiosyncrasies of each of these modalities, e.g., partial objects when visually searching but whole objects when touching. For different organisms, different expectations will develop.

Notice that on both representationalism and relationalism there can be an acknowledgement of the hard problem: sentient systems perceive their worlds rather than merely act in a way consistent with their sensory receptor activity. Perception implies an organism-level, pre-reflective awareness, an *experiencing*, which is lacking in mere sensory systems. Automobiles with sophisticated sensory systems, for example, can track the presence or absence of objects in the immediate surround of the vehicle, but they do not have unified experiences of these objects. In other words, cars, thermostats, and motion-detectors *sense* the features in their environments, but they do not perceive them.

We can acknowledge the subjective vantage point, as well as the need to provide a satisfying scientific account of it, however, while resisting the move to reify

[14] As with representationalist views, there are nuanced distinctions between the different brands of relationalism, the details of which need not detain us here. Here I describe just those features of the relational account that will be relevant for the subsequent discussion of content.

it, to treat perception *itself* as something over and above the activity of becoming aware. We might use words to describe perceptual activity — "Fred's perception of the dog caused him to move away"— in the same way that we might use words to describe activity of any sort — "Fred's drinking that glass of water caused him to urinate" — but the applicability of such descriptions does not entail that there is anything over and above the spatio–temporally extended bundle of activity which the words label.[15] Importantly, we need not commit to the existence of some content over and above the perceptual activity. Just as Fred's drinking the glass of water is nothing over and above the sum of the actions and processes, across multiple levels of granularity, that constitutes the drinking, Fred's perception of the dog is nothing over and above the sum of the actions and processes, across multiple levels of granularity, that constitutes the perceiving.

If perception does not yield representations of the sensory data our receptors collect, we might wonder why we do not see gaps corresponding to our retinal blind spots, saccadic smears as our eyes literally jump around, and colour fade-outs at the edges of our vision. On representationalism these are all neatly explained by appeal to the reconstructive process that removes these by-products of our sensory apparatus. If we did not have such internal representations, would visual experience be of a jagged, discontinuous, Picassoesque world?

No. To be a perceiver is to be an organism *skilled* in interacting with its environment. When we visually perceive an object, we see it without gaps because "the slightest flick of the eye, or of attention, renders it visible" (O'Regan and Noë, 2001, p. 947). In perception we interact with actual objects in the world, objects that are whole and continuously present. We move and adjust our bodies, our heads, our eyes so that we attend to different aspects of our locality to sense what is there. On this view, we never "have" a complete, uninterrupted, visual experience; rather, we continually experience the world — interact with it — not at the subpersonal level, at which there are discrete, gappy, moments of sensation, but at the organism level, a vantage point from which the chaos of micro-level activity appears calm, smooth, and slow to change. As a result, the traditional problems of binding, grounding, and intentionality disappear: "what explains the conceptual unity of experience is the fact that experience is a thing we are doing, and we are doing it with respect to a conceptually unified external object" (p. 967).

What then of non-veridical perception? As we have seen, one of the virtues of representationalism is its simple explanation of such phenomena. Relationalists have responded to this challenge in a variety of ways, but the general strategy is disjunctivist: since hallucination does not involve interaction, while misperception does, different mechanisms must underly them (Brewer, 2008; Johnston, 2004; Martin, 2004). More recent predictive models have made interesting

[15] I will not enter into the complex epistemological/metaphysical issue here of where the boundaries of such bundles of activity are to be drawn.

suggestions about what some of these different mechanisms might be;[16] but, since this is an appeal to a relational view of perception rather than a defense of it, I will leave the topic of hallucinations there and describe briefly how misperception might be understood on the relational view.

The general strategy is to take accuracy conditions to apply to descriptions of perception rather than to perceptions themselves. Since perception is an interactive process, something organisms do, it is not the right kind of subject for truth condition predicates: Does one *act* truly or falsely? We might act more or less skillfully, and this in turn might have consequences that could be cashed out in terms of truth and falsity, but then, as Brewer notes, it is the interpretation, a subject's response/judgement *to* its interactions, that is accurate or not: "… the erroneous nature of perceptual illusions is explained in terms of 'misinterpretation'" (Brewer, 2008, p. 169). In other words, a misperception is a description of occurrent perceptual activity as false in some way.

For example, suppose that a cat, on perceiving a distant stump in the forest, begins to hiss and exhibit a series of behaviours that we, third-person observers who study animal behaviour, classify as DEFENSIVE BEHAVIOUR. Why, we might wonder, is the cat responding in this way when no obvious threat is present? If we have observed repeated instances of cats exhibiting this behaviour in the presence of dogs, we might offer up the final cause hypothesis that the cat has misperceived the stump as a dog. This would link up the cat's current behaviour with past behaviour in a way that makes the range of behaviour internally consistent. On a relational view of perception, however, this final cause explanation does not entail that the cat has understood DOG, has had some pseudo-DOG thought, in the context of its perceiving activity. For the cat there are merely the details of the occurrent situation. Although we might profitably describe the behaviour as a mistaken response to the current situation, as a misperception of the stump, we go awry when we look for some lower-level constituent of the behaviour to which we can attach the label "DOG representation." The misrepresentation is an abstraction that comes into view only in the spatio–temporally extended context of the cat's behaviour over time and, having this context available to us, we use it to explain the behaviour. But there is no *thing* that is the misrepresentation, that causes the response. The entire response itself is the misrepresentation, so to speak.

This is a break from tradition and common sense to be sure. According to representationalism, perception is the mechanism by which organisms *receive* information from their environments. The resulting percepts, on that view, are *had* by organisms. In looking at an apple, my perception is "as of" something

[16] In chapter seven of *Surfing Uncertainty*, Clark explains that on the predictive processing model, delusions and hallucinations "flow from a single underlying cause: falsely generated and highly weighted (high-precision) waves of prediction error" (Clark, 2015, p. 206).

small, round, and red — these are impressions *in* my mind. If Fred, walking through the forest at night, recoils upon seeing a distant stump and begins to change direction, it is Fred's perception of the stump "as of" a dog that causes Fred to change direction, that is, it is Fred's *misrepresentation* of the current situation, that, at least partly, causes his behaviour.

From the vantage point of the critique I have been developing here, however, such analyses confuse final and efficient causes. When we talk about Fred's perception of the stump "as of" a dog, we are generalizing over the entire bundle of spatio–temporally extended behaviour beginning with the learning of the MOVE AWAY response to instances of DOG and ending with the current instance of the MOVE AWAY response to a stump. At any given moment of perceiving activity, however, this generalization does not exist: there is only the complex bundle of behaviour that constitutes Fred's perceiving as he moves through the forest. As Fred moves, the contours of the situation he is in extend spatio–temporally in different directions. Fred may have entered the forest in an anxious state, perhaps because he is uncomfortable walking through the forest at night by himself. Fred might have had a recent experience with a hostile dog, priming him for a DOG-related response. To use a dynamic systems term, responses such as MOVE AWAY, connected as they are to past dog experiences, are attractors in the current situation, which is to say, they are response patterns that Fred, as a dynamic system, will tend to enact. Thus primed, a distant, shadowy stump might indeed efficiently cause Fred to react with (what we describe as) a DOG response. From the perspective of our final cause analysis — our interpretation of the entire situation — we might say that something has gone wrong, that Fred has misperceived the stump as a dog, but from the perspective of occurrent Fred, it is the physical stump in Fred's environment that is the (partial) efficient cause of his MOVE AWAY response. In other words, it is perfectly consistent for Fred's perceiving the stump as a DOG to explain (final cause) his response without the perception of the stump as a DOG thereby being the (efficient) cause of Fred's behaviour.

This, at least, is the beginning of the relational response to the "problem" of misperception. A representationalist might object that such accounts fail to explain conscious perceptual activity and the actions that seem to be guided directly by the awareness of that activity: Fred, in misperceiving the stump "as of" a dog, becomes *aware* of the stump/dog and its potential threat and this awareness seems to play a role in *keeping* Fred moving away from the stump. In Shea's terminology, moving away from the stump/dog is a robust outcome for Fred and his occurrent behaviour of continuing to move in a trajectory continually away from the stump is most comprehensively viewed as implementing an algorithm for MOVE AWAY, in other words, as an efficient cause of his actions.

Ultimately, explaining how distal causes become efficient causes is the goal of both internalist and externalist theories of content. Because distal causes are, well, distal, they must yield effects by way of other, proximal, factors. On the internalist

approach, these are the physical mechanisms that function as internal representations of distal conditions. On the externalist approach, the representations of distal conditions are themselves external. Thus, the externalist owes an account of how external resources take on this role. I turn to this next.

Vehicles of Content

A person develops, over time and within a linguistic community within which there is a *practice*[17] of using words in particular ways, a response association between words and the objects to which they refer.[18] One way in which responses, either instinctive or learned, become associated with different feature bundle types is through stimulus–response conditioning. After repeated experience with the constant conjunction of instances of different feature bundle types, the response to one becomes a (conditioned) response to the other. Pavlov's classic example of operant conditioning involves a dog's unconditioned response (R), salivation, to instances of MEAT (M). M is the feature bundle type that triggers R. The dog has no initial response to instances of BUZZER (B). When the dog is repeatedly exposed to a conjunction of instances of the feature bundle pair, M and B, the dog learns an association between the two and salivates when *either* M or B is present (Pavlov, 1927). In predictive processing terms, hearing B elicits for the dog an expectation for M: at different levels of granularity and across various networks of associations — neural, muscular, respiratory — the dog responds to B as though M were already present. A word, as another feature bundle type in the landscape, can also become a trigger of associated response patterns. A dog that learns to associate instances of the word "ball" with instances of BALL, will exhibit the same behaviour pattern — e.g., become excited and restless — when hearing the word as it would on seeing a ball. In other words, hearing the word will prime the dog's executive BALL responses; the dog's attention will be easily drawn by balls.

As we have seen, responses to instances of feature bundle types, e.g., balls, can develop in the absence of an understanding of the type, BALL, to which the instances conform. A classifier can consistently and reliably sort BALL pictures from BOX pictures without thereby using any knowledge of BALL and BOX in so doing, that is, it processes and outputs in accordance with these types, but it is not driven by them. In the same way, response patterns to words can develop in the absence of conceptual understanding of word *meaning*. Words qua utterances

[17] Elsewhere (Salay, 2019), I elaborate upon the role that such a practice plays in establishing the functional role of words as content vehicles.

[18] I must use words here to describe this account and, given the standard view, most of the words I must use are heavy with internalist, representational baggage. As this account unfolds, it should become clear to what "object" and "reference" amount.

are simply feature bundles in an organism's environment that can be interacted with and, thereby, become bundles of significance with which associations can be formed. Rover, hearing the word "ball," shifts his attentional focus to the possible presence of balls: he wags his tail, looks around expectantly, and prepares for movement.

From the perspective of individual word use, then, words function as efficient causes of behaviour: perception of them quite literally triggers responses. But since word associations are made possible by the extended linguistic context within which there is a *practice* of using words in this way, words function as vehicles of content. "Vehicle" is a metaphor here, of course, since words do not *carry* anything, and "content" is not a thing to be carried. As the use of a word proliferates through a linguistic community, individual usage of the word will become increasingly constrained to meet the standard that emerges. "Look, there's a dog," says Sally. "That's a cat, not a dog," corrects her mother. Content is the abstract, spatio–temporally extended pattern of use–response that emerges out of this linguistic practice. As a result of this system, skilled word users can respond to things that are neither spatially nor temporally present: Fred, walking rapidly away from the stump qua dog, continues to tell himself, "there's a dog,"[19] and in this way the stump qua dog exerts a causal influence on Fred's behaviour, even after it has faded from immediate view.

In the contemporary human context, words are by far the most ubiquitous item in an individual's environment. For someone who is linguistically sophisticated, words are ready-to-hand tools that, because of the potent combination of experience, learning, and habit, draw the attention. Fred hears the front door open and close and yells, "I'm in the study." Does he desire to make his presence known to whomever has entered? Is this desire an internal representation of some ideal state of affair that he now, in speaking, has moved to bring about? On the externalist view such explanations are, as always, final cause accounts of behaviour: there is no inner *intention* that Fred is bringing about by these verbal actions; rather, his desire to make his presence known is constituted by his actions.

Over time, as an individual develops into a sophisticated word-tool-user, a new capacity begins to emerge, one that is possible only with the aid of these content vehicles. Words qua objects of perception function to draw our attention and trigger responses, but as qua content vehicles they function to draw our attention to what the words represent. When we are perceptually engaged with words

[19] That self-talk is used in this way has been extensively studied and, in a review of the developmental literature, Zelazo (2015, p. 61) concludes that "Contemporary research generally supports the seminal ideas of Vygotsky (1934/1962) and Luria (1959, 1961) concerning the importance of verbal processes in the exercise and development of self-regulation, finding, for example, that with age children increasingly use verbalization strategically to maintain task information in mind (Karbach and Kray, 2007), and that blocking the use of inner speech disrupts cognitive control in children and adults (Emerson and Miyake, 2003; Kray, Eber, and Karbach, 2008)."

— thinking — we are continually being drawn to the other ready-to-hand words in our environment. As we reach for them, we perpetuate our ongoing perceptual interaction with them: our attention is continually being drawn back to the words — they are tools that are *there* — and as we reach for more, we sustain the interaction.

Words, then, in the context of a deeply entrenched, linguistic practice, yield a deeper capacity for inhibitory control of attention. Each time we use a word, we draw attention *away* from whatever else is also pulling our focus, the stuff in our immediate surround, and *toward* what the word represents.[20] With each utterance, new, deeply interconnected, possibilities draw our attention. As our linguistic proficiency grows, words become increasingly ready-to-hand until, eventually, the interaction between self and words is a coupling that, for the tool-user, feels natural. We feel confident in our ability. We describe this feeling as *understanding.* On the relational account of perception, we experience our environment in a smooth, consistent, uninterrupted way partly because we are confident in our sensory skills (O'Regan and Noë, 2001, p. 946). Likewise, word use is experienced as understanding, or not, because we are skilled word users, or not. The more connections we have learned between words and possible responses, the more profound is this sense of understanding.

Contents are determined, then, by the way in which a word can be used in a community that has a linguistic practice of associating words with the high-level patterns to which they have developed responses. Content users — intentional agents — are the members of those communities; they are individuals who use language as a representational tool (Clark, 2005; Clark and Chalmers, 1998; Wheeler, 2004). We are intentional beings when we are augmented by the language tool. Reasons are "out there," but because the word tool is so ubiquitous, so ready-to-hand, it is tempting to treat reasons and thoughts as internal to us. This is an illusion. "If the phenomenological picture of ourselves as beings-in-the-world is correct, then, when we augment it with a world that is replete with sign vehicles, the traditional Cartesian mind/body dualism comes into soft relief: we are beings who, most fundamentally, act; however, as it happens, most of our tools for action are tools of reason, namely, words" (Salay, 2019, p. 141). As with any paradigm shift, questions and objections from the perspective of the reigning framework loom large. I conclude with a brief response to two questions that will likely be looming large for those attracted by ICR.

Doesn't Intentionality Presuppose a Capacity for Recursive Thought?

Giving an account of intentionality is, many think (Kriegel, 2009; Lycan, 1996; Rosenthal, 1986; Van Gulick, 2004), just to give an account of a *capacity*

[20] Of course, there are uses of language besides reference!

for second-order thought.[21] According to content externalism, this is right in one way and deeply mistaken in another. It is right to observe that a representational tool such as language gives one a capacity to talk and think about things. If "ball" stands for BALL, then Fred talks *about* balls when he utters the word. But it does not follow from this that Fred now has a second-order capacity; rather, it is Fred + "ball" qua representational tool in Fred's linguistic community that yields the second-order capacity. As Clark (2006b, p. 294) observes, even simple symbol tokens can convert difficult second-order tasks into simple first-order ones. Language-naïve chimpanzees who are taught to associate first-order relations such as sameness and difference with symbol tokens are subsequently capable of solving the hitherto opaque second-order task of identifying object groupings that instantiate sameness and difference properties (Thompson, Oden, and Boysen, 1997). These chimps have not thereby gained a new capacity, an inner understanding of some abstract relation; rather, with the aid of the symbol tokens, the second-order problem is reduced to the first-order one of grouping pictures associated with one type of token separately from pictures associated with another type of token. Language does not endow us with a second-order capacity either; it reduces second-order tasks into first-order ones. Words are stand-ins for complex, abstract, spatio–temporally extended patterns. We are not the sorts of beings that can interact with such entities, but by developing responses to the words that stand for them, we are able to solve problems that require them.

If Fred Intends to Eat Dessert Tonight, Isn't There Something About Fred Now That Represents This Intention?

What does this externalist account say of Fred, who plans to eat dessert after dinner tonight? For materialist RTMers who are committed to realism about mental representations, contentful mental states must be physically manifested in some way, that is, if Fred plans to order dessert at the restaurant, there must be something about Fred now that *represents this future intention*. In other words, a successful reduction of mental level content to lower-level physical processes, i.e., vehicles of content, is a necessary condition of realism about the mental realm. But, I've argued, this ICR programme cannot succeed, resting as it does on an equivocation between content qua efficient cause and content qua final cause. Does this mean that mental representations are not real?

If what we mean by "real" is a version of "Fodor's industrial-strength realism" (Dennett, 1991, p. 42) then, no, since on that view "beliefs and their kin would not be real unless the pattern dimly discernible from the perspective of

[21] Note that I am not speaking about self-consciousness here, which we might think is a specific kind of second-order thought. There is not room to begin a discussion of this complex issue here.

folk psychology could also be discerned (more clearly, with less noise) as a pattern of structures within the brain" (p. 42). On the other hand, if what we mean by "real" mental representation includes something extended, that is the result of a tool-dependent cognitive process, then, yes, mental representations are real: "The process that produces the data of folk psychology … is one in which the multidimensional complexities of the underlying processes are projected *through linguistic behaviour*, which creates an appearance of definiteness and precision, thanks to the discreteness of words" (p. 45).

Is there something now about Fred that represents his future intention to eat dessert? Yes and no. In having the intention, Fred uses words to express to himself that he will eat dessert tonight. This act of expression is something that is constituted by a multitude of activity at different levels of granularity: the low-level neural activity that is active when Fred is using the relevant words; the mid-level circulatory and/or digestive activity that is triggered by the word/act associations; the high-level perceptual activity that Fred experiences; the higher-level reflective activity that Fred might engage in if he keeps his attention focused on the "intention" by repeatedly using words — "Yes, I'll have dessert tonight" — which in turn trigger responses — salivation, hunger, episodic flashes of past dessert experiences — that heighten the focus on dessert. By the time Fred is at dinner, unless something more powerful has drawn his attention, the loop of activity will reoccur, perhaps less consciously if it has been deeply engrained, and Fred will fulfill this intention to eat dessert. At no point is there an inner intention that causes Fred to eat dessert, but the ongoing, extended, multi-layered activity together constitutes his intention to eat dessert. The view is thus realist about content, but eliminativist about "Real" content.

Conclusion

Contra the current zeitgeist within which representation is the polarizing concept, it is the assumption that representations are *internal* resources of intentional agents that is mistaken. The strategy of explaining the content of mental representation by appeal to the conditions under which relevant internal physical mechanisms function as content vehicles, I have argued, confusedly identifies spatio–temporally extended patterns of behaviour with occurrent causes of action. The former do cause actions, but only by way of the *external* content vehicles to which we develop responses.

On content externalism, representational contents are abstractions, high-level patterns that emerge out of the ongoing, shared interactions of individual members of linguistic communities. The physical, personal-level media of these transactions, the content vehicles, are words. As triggers of responses, words function as efficient causes of behaviour. Intentional agents act because of reasons, but

they accomplish this feat with the aid of representational tools that expand their cognitive capacities in myriad ways.[22]

An individual wholly without sign tools, then, is not an intentional agent; a skilled user of language in the context of a sophisticated linguistic community is. In between these two extremes is a continuum of intentionality: the more representation tools there are at hand and the more skilled one is with them, the greater is the possibility for talking and thinking *about* things; the fewer tools there are at hand and the less skilled one is with them, the lesser the possibility for talking and thinking *about* things.[23] The landscape of the modern human is replete with words that are delivered in multiple modalities — visual, aural, tactile. Young children who have not yet learned language have nevertheless been exposed to an astonishingly large number of words, through interaction with people, books, billboards and street signs, television, and the internet.[24] Thus it is not an exaggeration to say that there are more words in a modern human's everyday environment than there is anything else. As words become increasingly ready-to-hand, they begin to feel like internal resources, but this is an illusion. They are a part of the rich, representational scaffolding of language that grounds our intentionality.

References

Anderson, J. R. (2005). *Cognitive psychology and its implications*. New York: Worth Publishers.

Aristotle. (1955). *Physics.* In W. D. Ross (Ed.), *Aristotle*: *Selections.* New York: Scribner.

Austin, J. (1962). *How to do things with words: The William James Lectures delivered at Harvard University in 1955*. (J. O. Urmson and M. Sbisà, Eds.). Oxford: Clarendon Press.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy X*(1), 615–678.

Block, N. (2003). Mental paint. In M. Hahn and B. Ramberg (Eds.), *Reflections and replies: Essays on the philosophy of Tyler Burge*. Cambridge, Massachussetts: MIT Press.

Bower, M., and Gallagher, S. (2013). Bodily affectivity: Prenoetic elements in enactive perception. *Phenomenology and Mind, 2*, 108–131.

Brewer, B. (2008). How to account for illusions. In A. Haddock and F. Macpherson (Eds.), *Disjunctivism: Perception, action, knowledge* (pp. 168–180). Oxford: Oxford University Press.

Brewer, B. (2017). The object view of perception. *Topoi, 36*, 215–227.

Broadbent, D. (1982). Task-combination and selective intake of information. *Acta Psychologica, 50*, 253–290.

[22] Much has been written about this in the extended mind literature. See Clark, 2006b; Salay, 2019; Wheeler, 2004 for discussion of how a toolbox of symbols makes reasoning amodally and abstractly, planning, and regretting possible.

[23] I have not spoken about non-human animals at all here, but I take it that many animals use a variety of representation tools, e.g., calls and gestures, and therefore certainly do exhibit intentional behaviour.

[24] Literacy projects such as the Thirty Million Word Initiative (https://tmwcenter.uchicago.edu/tmwcenter/research/) are founded on empirical data showing a strong co-relation between early word exposure and later-life cognitive aptitude (Sénéchal and LeFevre, 2002; Suskind, 2015). Given content externalism, this is to be expected, since its central claim is that language is the external scaffolding on which our key cognitive capacities depend.

Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics, 33*(1), 1–78.
Cao, R. (2012). Teleosemantic approaches to information in the brain. *Biology & Philosophy, 27*, 49–71.
Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, Massachussetts: MIT Press.
Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge, Massachussetts: MIT Press.
Clark, A. (2005). Intrinsic content, active memory and the extended mind. *Analysis, 65,* 1–11.
Clark, A. (2006a). Language, embodiment, and the cognitive niche. *Trends in Cognitive Science*, *10*(8), 370–374.
Clark, A. (2006b). Material symbols. *Philosophical Psychology, 19*(3), 291–307.
Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.
Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.
Clark, A., and Toribio, J. (1994). Doing without representing? *Synthese, 101*, 401–431.
Crane, T. (1992). The non-conceptual content of experience. In T. Crane (Ed.), *The contents of experience* (pp. 136–157). Cambridge: Cambridge University Press.
Craver, C., and Bechtel, W. (2007). Top–down causation without top–down causes. *Biology and Philosophy*, *22*, 547–563.
Davidson, D. (1975). Thought and talk. In S. Guttenplan (Ed.), *Mind and language* (pp. 7–23). Oxford: Oxford University Press.
Dennett, D. (1981). True believers: The intentional strategy and why it works. In J. Haugeland (Ed.), *Mind design II: Philosophy, psychology, artificial intelligence* (pp. 57–79). Cambridge, Massachussetts: MIT Press.
Dennett, D. (1991). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.
Dennett, D. (2003). Explaining the "magic" of consciousness. *Journal of Cultural and Evolutionary Psychology, 1*(1), 7–19.
Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review, 3*(4), 357–370.
Dewey, J. (1916). *Essays in experimental logic*. Chicago: University of Chicago Press.
Dewey, J. (1938). *Experience and education*. New York: Macmillan.
Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, Massachussetts: MIT Press.
Dretske, F. (1988). *Explaining behaviour: Reasons in a world of causes*. Cambridge, Massachussetts: MIT Press.
Dretske, F. (1995). *Naturalizing the mind*. Cambridge, Massachussetts: MIT Press.
Dreyfus, H. (2002). Intelligence without representation: Merleau–Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences, 1*(4), 367–383.
Dreyfus, H. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence, 171,* 1137–1160.
Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170*, 115–135.
Emerson, M., and Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, *48*, 148–168.
Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Atul, P., Tadayoshi, K., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1625–1634.
Fodor, J. (1987). *Psychosemantics*. Cambridge, Massachussetts: MIT Press.
Fodor, J. (1990). *A theory of content and other essays*. Cambridge, Massachussetts: MIT Press.
Fodor, J. (1994). *The elm and the expert*. Cambridge, Massachussetts: MIT Press.
Freeman, W. (2000). *How brains make up their minds*. New York: Columbia University Press.
Freeman, W., and Skarda, C. (1990). Representations: Who needs them? In J. McGaugh, N. Weinberger, and G. Lynch (Eds.), *Brain organization and memory* (pp. 375–380). Oxford: Oxford University Press.
Friston K., Kilner J., and Harrison L. (2006). A free energy principle for the brain. *Journal of Physiology*, *100*(1–3), 70–87.
Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford University Press.
Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford: Oxford University Press.
Haken, H. (1990). Synergetics as a tool for the conceptualization and mathematization of cognition and behaviour — how far can we go? In H. Haken and M. Stadler (Eds.), *Synergetics of cognition* (pp. 2–31). Berlin: Springer.

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.
Heidegger, M. (1962). *Being and time* [J. Macquarrie and E. Robinson, Trans.]. New York: Harper & Row. (originally published 1927)
Herbers, J., Cutuli, J., Supkoff, L., Heistad, D., Chan, C., Hinz, E., and Masten, A. (2012). Early reading skills and academic achievement trajectories of students facing poverty, homelessness, and high residential mobility. *Educational Researcher, 41*(9), 366–374.
Hurley, S. (1998). *Consciousness in action*. Cambridge, Massachussetts: Harvard University Press.
Husserl, E. (1989). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: Second book studies in the phenomenology of constitution* [R. Rojcewicz and A. Schuwer, Trans.]. Dordrecht: Kluwer Academic. (originally published 1913)
Johnston, M. (2004). The obscure object of hallucination. *Philosophical Studies, 120*(1–3), 113–183.
Karbach, J., and Kray, J. (2007). Developmental changes in switching between mental task sets: The influence of verbal labeling in childhood. *Journal of Cognition and Development*, *8*, 205–236.
Keijzer, F. (1998). Doing without representations which specify what to do. *Philosophical Psychology, II*(3), 269–302.
Kray, J., Eber, J., and Karbach, J. (2008). Verbal self-instructions in task switching: A compensatory tool for action-control deficits in childhood and old age? *Developmental Science*, *11*, 223–236.
Kriegel, U. (2009). *Subjective consciousness*. Oxford: Oxford University Press.
Locatelli, R., and Wilson, K. (2017). Introduction: Perception without representation. *Topoi, 36*(2), 197–212.
Logan, R. (2007). *The extended mind: The emergence of language, the human mind, and culture*. Toronto: University of Toronto Press.
Luria, A. (1959). The directive function of speech in development and dissolution. *Word*, *15*, 341–352.
Luria, A. (1961). In J. Tizard (Ed.), *The role of speech in the regulation of normal and abnormal behavior*. New York: Liveright Publishing Corporation.
Lycan, W. (1996). *Consciousness and experience*. Cambridge, Massachussetts: MIT Press.
Martin, M. (2002). The transparency of experience. *Mind & Language, 17*(4), 376–425.
Martin, M. (2004). The limits of self-awareness. *Philosophical Studies, 120*(1–3), 37–89.
Mead, G. H. (1938). *The philosophy of the act*. Chicago: University of Chicago Press.
Merleau–Ponty, M. (2012). *Phenomenology of perception* [D. A. Landes, Trans.]. London: Routledge. (originally published 1945)
Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, Massachussetts: MIT Press.
Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge, Massachussetts: MIT Press.
Noë, A. (2010). Vision without representation. In N. Gangopadhyay, M. Madary, and F. Spicer (Eds.), *Perception, action, and consciousness: Sensorimotor dynamics and two visual systems* (pp. 245–256). New York: Oxford University Press.
O'Regan, K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*, 939–1031.
Papineau, D. (1987). *Reality and representation*. Oxford: Blackwell Publishers.
Pavlov, I. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* [G. V. Anrep, Trans.]. London: Oxford University Press.
Peacocke, C. (2001). Does perception have a nonconceptual content? *Journal of Philosophy, 98*, 239–264.
Peirce, C. S. (1958). *Collected papers of C. S. Peirce* (C. Hartshorne, P. Weiss, and A. Burks, Eds.). Cambridge: Harvard University Press.
Rasmussen, C., Baydala, L., and Sherman, J. (2004). Learning patterns and education of Aboriginal children: A review of the literature. *Canadian Journal of Native Studies, 24*(2), 317–342.
Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329–359.
Ryder, D. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind & Language, 19*, 211–240.
Salay, N. (2019). Learning how to represent: An associationist account. *Journal of Mind and Behavior, 40*(2),121–145.
Sénéchal, M., and LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73*(2), 445–460.
Shea, N. (2018). *Representation in cognitive science*. Oxford: Oxford University Press.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.
Stalnaker, R. (1984). *Inquiry*. Cambridge: MIT Press.
Suskind, D. (2015). *Thirty million words: Building a child's brain*. New York: Dutton Books.
Thelen, E., Schöner, G., Scheier, C., and Schmidt, L. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences, 24,* 1–86.
Thompson, R., Oden, D., and Boysen, S. (1997). Language-naive chimpanzees (*Pan troglodytes*) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behaviour Processes, 23*, 31–43.
Townsend, T., and Busemeyer, J. (1995). Dynamic representation of decision making. In R. F. Port and T. van Gelder (Eds.), *Mind as motion* (pp. 101–120). Cambridge, Massachussetts: MIT Press.
Travis, C. (2013). The silence of the senses (revised). In C. Travis, *Perception: Essays after Frege* (pp. 23–58). Oxford: Oxford University Press.
Vallacher, R., Nowak, A., and Kaufman, J. (1994). Intrinsic dynamics of social judgement. *Journal of Personality and Social Psychology, 67*, 20–34.
Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, *92*(7), 345–381.
Van Gulick, R. (2004). Higher-order global states (HOGS): An alternative higher-order model of consciousness. In R. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 67–93). Philadelphia: John Benjamins.
Van Rooij, I., Bongers, R., and Haselager, F. (2002). A non-representational approach to imagined action. *Cognitive Science*, *26*(3), 345–375.
Varela, F., Thompson, E., and Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, Massachussetts: MIT Press.
Vygotsky, L. (1962). *Thought and language* [E. Hanfmann and G. Vakar, Trans.]. Cambridge, Massachussetts: MIT Press. (Originally published 1934)
Wheeler, M. (2004). Is language the ultimate artefact? *Language Sciences, 26*, 693–715.
Wittgenstein, L. (2009). *Philosophical investigations* (fourth edition) [P.M.S. Hacker and J. Schulte, Eds.]. Oxford: Wiley–Blackwell. (originally published 1953)
Zelazo, P. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review, 38*, 55–68.