The Many Problems of Representation

I. Background.

A representation is anything that is *about* something. Philosophers call the property of being about something *intentionality* (see Intentionality, this volume). *Intentionality* and *intentional* are technical terms in philosophy; they do not mean a representation was necessarily made 'intentionally' in the everyday sense; just that it is about something. A picture of a tree in your garden is a representation because it is about the tree. When we think, we also think about things, e.g., about my birthday party yesterday, so thoughts are ways of representing. Thoughts are *personal* level representations; it is the person herself who represents the world a certain way by thinking (or seeing and so on). Psychologists also postulate sub-personal representations, where a part of the person does the representing. For example, your auditory system might represent a time difference between sounds you as a person were not even aware of. Psychologists regularly use representations in their theories: They might say prejudice occurs because of *stereotypes*, or we see the way do because of the way the visual system *encodes* information. Almost all of psychology refers to personal or sub-personal representations. But how can anything represent? Can we explain the ability of minds to represent in terms of natural science? In short, how is psychology possible?

Representations have a certain anatomy, which we will gloss as *target*, *content* and *vehicle*. When I think of my birthday party yesterday, the party plays a role in my thought in two distinct ways. On the one hand, there is the actual party I was at yesterday which my current thought is directed at or refers to. This is called the *target* (or *referent* --these terms are not quite synonymous) of the thought. On the other hand, there is how the party is presented in my thought; for example, I may think of it as, 'yesterday's party', or as 'my last birthday party.' These two thoughts have different *contents* because they present the same referent in different ways. If I have a thought with the content, 'my party was yesterday,' then the thought is true (target and content match); but if the thought had the content 'my party happened a week ago', the thought would be false. It is because target and content are distinct that thoughts can be true or false (target and content match or mismatch); and that deception, illusions and hallucinations are thus possible. Similarly, the distinction is what allows my boss and my spouses' lover to be thought of as different people (same referent person, different contents); and thus more generally allows people to be at cross-purposes and for life to be interesting. In general, all representations have content, but they need not have a target or referent. If I imagine a horse, there need be no target horse to check the accuracy of my imagination against. Finally, as well as always having content and (sometimes) a target, representations are in general made of something physical, the representational *medium* or *vehicle*. For example, if I write on a blackboard, the vehicle is the chalk and the blackboard. The distinctions between content, target and vehicle may seem obvious, but failure to keep them conceptually separate has lead to mistaken explanations in psychology (see Perner, 1991, chapter three, and Dennett, 1991, chapters five and six, for examples).

Representation is a central concept in cognitive science. The fact that some representations are not mental (e.g. photos) has motivated the idea that understanding simple physical representations might enable understanding the mind in a naturalistic way. Some feel that understanding representation might give us a handle on consciousness itself. Perhaps the goal of psychology in general is precisely to understand to what extent and in what way the mind represents. According to Perner (1991), how children come to understand the mind *as representational* is central to child development. Further, understanding representation is central to the philosophy of language, of science and of rationality. Here we will focus just on issues relevant to the nature of mental representation.

II. The Representational Theory of Mind.

The claim that people in having mental states can represent an external world is the weakest sense of a *representational theory of mind*. It is bolder to assert the best explanations of the mind will

be in terms of representations (a move not all are willing to make, as discussed below). Explaining the mind in terms of representations offers the appeal of having a naturalistic explanation of the mind, if only representations could be understood naturalistically. If a student thinks of a party, then presumably the student has some neural activity (the representational vehicle) that is about the party. There is reason to think we should be able to understand how physical things, like neural activity, could acquire the apparently non-physical relation of aboutness. Computer programmes for example consist of internal electrical processes that represent things outside the machine. Other familiar physical entities like linguistic inscriptions or pictures are also able to acquire aboutness. For instance, a photo of the party is a physical object (representational vehicle: piece of paper with patterns of colour) that shows a scene of the party (target) in a certain way (content). Moreover, it seems clear what each of these elements (vehicle, content and target) is. However, when one asks why the photo shows the scene of the party the answer defers to the beholders of the photo: they see that scene in it. Similarly for language: the readers understand the linguistic inscriptions in the book, etc. If it is us human users of these representations that determine content, then this makes them useless to explain the origin of intentionality of the mind, because it is only derived from their users' intentionality (*derived intentionality*, see Searle 2005). There is no user endowed with intentionality reading our nervous activity. Philosophers have since made various proposals of how intentionality can be naturalized by explaining how mental representations can have content without the help of an agent imbued with intentionality. A failure to answer the question pushes one to either adopt dualism (intentionality is an inexplicable primitive property of minds) or to deny a representational explanation of the mind is possible.

According to the *representational theory of mind* proposed by Fodor (amongst others; see Sterelny, 1990) a mental state like thinking of a party consists of a mental representation of the party as well as the *attitude* of thinking, captured by the *functional role* the representation of the party plays in relation to other mental states. Different functional roles define different attitudes, including wishing, wondering, dreaming and so on (in fact, all the different types of mental states recognised in ordinary language, i.e. as recognised in 'folk psychology'). Some people who endorse explaining the mind in terms of representations nonetheless reject the claim that the folk categories of believing, wanting etc are real natural kinds (for example, Churchland). Fodor's suggestion of a representational theory of mind included the claim that the mental representations were of a special sort, namely symbolic. Others have denied mental representations are symbolic, an issue we also comment on below.

III. Foundational Problems of Naturalizing Representation:

III.1 Representational content.

The content of a representation is how it portrays the world as being. We discuss four main approaches to determining how the content of a representation is determined. There is not wide agreement yet that any of them has solved the problem.

III.1.a. According to Dennett, there is no fact of the matter as to what the content of a representation is, or even whether it is a representation (a position called *irrealism*). It is sometimes useful for an observer to view a system as an intentional one; in that way the observer takes an *intentional stance*. For example, I might say 'my computer is thinking about the problem', 'my car wants to stop now', if I find this useful in predicting behaviour. In taking this stance one works out what beliefs and desires the system ought to have if it were a rational agent with a given purpose. Hence one can work out what behaviour the system ought to engage in. On other occasions it might be more useful to take a *physical stance*, analysing the system in terms of physical laws and principles, a process that can take

considerable effort. What type of stance we take is largely a matter of convenience. Similarly, even the thoughts we seem introspectively to be having are actually interpretations we make: stories we spin about ourselves (for example, that we are having a certain thought) for which there is no fact of the matter. The account raises a number of issues: Why is the intentional stance so good at making predictions (at least for people) if it fails to describe and explain anything real? How can we as observers make interpretations of ourselves without already having intentionality? And given people are not always optimally rational how do we constrain the stories that are spun?

The remaining theories are *realist* about representations: They presume there is some fact of the matter as to whether something represents and what its content is.

III.1.b Causal/informational theories, for example Dretske's early work, assume that a neural activity 'cow' becomes a representation meaning COW if it is reliably caused by the presence of a cow, hence reliably carries the information that (indicates) a cow is present. A recognized problem with this approach is that on this theory representations cannot misrepresent. If 'cow' is sometimes caused by the presence of a horse on a dark night then it simply accurately represents the disjunction of COW-OR-HORSE-ON-A-DARK-NIGHT (disjunction problem) instead of misrepresenting the horse as a cow as one would intuitively expect.

III.1.c Functional role semantics, for example Harman's , assume the meaning of a representation is determined by the permissible inferences that can be drawn from it or, in more liberal versions, how the representation is used more generally. 'Cow' means COW because it licenses inferences that are appropriate about cows. The idea can be motivated by considering a calculator. If one inspected the wiring diagram, one could work out which button or state corresponded to the number '3', which to the process of addition, and so on, just by determining how the states were related to each other. To be able to represent the number '3', surely one needs to have appropriate relations between that representation and the one for '2', for the concept of 'successor', etc. And maybe it is just the same for concepts of cows, bachelors, assassins, and so on.

One problem with functional role semantics is how learning anything new, by changing the inferences one is disposed to draw, may change the meaning of all one's representations (a consequence called holism). As I am always learning, I can never think the same thought twice. Nor can any two people think the same thought, given their belief structures are different, if only slightly. Does this make it impossible to state useful generalisations concerning mental states and behaviour? How can we respect the difference between changing one's belief about a concept and changing one's concept? It is also not clear how conceptual role semantics can allow misrepresentation: because the meaning of a representation is determined by its uses, how can a representation meaning one thing be used as if it meant something else? The solution to these problems is to restrict which uses determine meaning, but there is no agreement how in general to do this.

III.1.d Dretske and Millikan proposed teleosemantics as a theory of naturalising content. Whereas causal/informational theories focus on the input relationship of how external conditions cause representations, and functional role semantics focuses on the output side of what inferences a representation enables, teleosemantics sees representations as midway between input and output and focuses on function as determined by history. Representations do not represent simply because they reliably indicate, or simply because they are used, but because of functions brought about by an evolutionary or learning history, namely a function to indicate. Because of an evolutionary history, a heart is an object with the function of pumping blood. So certain patterns of neural activity may come to have the function of indicating, for example, cows. The basic assumption is that there must be enough covariation between the presence of a cow causing a 'cow' token, so that the token can guide cow-appropriate behaviour often enough that there is a selective advantage without which the cow-

'cow' causal link would not have evolved. The 'cow' token *thereby* acquires the function of indicating cows. Just as historically defined function allows a heart to both function and malfunction, a representation can malfunction making misrepresentation possible. If a horse happens to cause a token naturally selected for indicating cows, then the token misrepresents the horse as a cow. The token still means COW (disjunction problem solved) because the causing of 'cow' by horses was not what established this causal link. Although this theory seems to be the current leading contender it is far from uncontested (see Neander, 2004). For example, it seems conceptually clear that a representation could be adaptive but not accurate, or true but maladaptive to believe. The distinction between accurate and adaptive makes no sense to teleosemantics.

III.2. Broad versus narrow content:

To the extent the content can be fixed by what happens in our head, by having the same vehicle state, the content is called narrow. To the extent that the content can be fixed only by reference to the world as well as the vehicle state, it is called broad or, equivalently, wide. People who believe content is in general narrow are called internalists; those who believe content is in general broad are called externalists. The two camps have been vigorously debating for some decades now (see Content Externalism, this volume).

Externalism or internalism are motivated by different theories of content. On teleological theories, for example, content is necessarily broad because content is determined by a selectional history. Exactly the same neural wiring may constitute a mouse detector or a poteroo detector in different possible worlds in which it was selected for mapping its on/off state onto mice or poteroos, respectively. No amount of examination of the vehicle alone need answer the question of which it represented. Similarly, two different neural wirings may both be poteroo detectors given relevant selectional pressures. Conversely, on a strict functional role semantics, content depends on the inferences that could be drawn, so content depends only on what goes on inside the head.

III.3. Thoughts about the non-existent.

The attempts to naturalize representational content has almost exclusively focused on the use of representations in detecting the state of the environment and acting on the environment. In other words the analysis—in particular of causal and informational approaches—is restricted to cases of real, existing targets. The teleosemantic approach can be extended to the content of desires since these representations can also contribute to the organism's fitness in a similar way to beliefs. Papineau made desires the primary source of representation. However, little has been said how natural processes can build representations of hypothetical, counterfactual assumptions about the world or even outright fiction. One intuitively appealing step is to copy informative representations (caused by external states) into fictional contexts, thereby *decoupling* or *quarantining* them from their normal indicative functions. Although intuitively appealing, the idea presupposes the standard assumption of symbolic computer languages that symbols exist (but based on derived intentionality) that can be copied and keep their meaning. It remains to be shown how such copying can be sustained on a naturalist theory.

III.4 Implicit vs explicit representation.

A representation need not make all of its content explicit. David Kirsch proposed that the degree to which information is explicitly encoded in a representation is related to the computational cost of recovering or using the information. According to Dienes and Perner (see our 2002) something is implicitly represented if it is part of the representational content but the representational medium does not articulate that aspect. This can be illustrated with bee dances to indicate direction and amount of nectar. From the placement in space and the shape of the dance bees convey to their fellow bees the direction and distance of nectar. The parameters of the dance, however, only change with the nectar's

direction and distance. Nothing in the dance indicates specifically that it is about nectar (an object which nonetheless plays a causal role in the bees behaviour). Yet nectar is part of the meaning. It is represented implicitly within the explicit representation of its direction and distance

Bee dances only make the properties of distance and direction explicit. The English sentence 'the nectar is 100 yards south' makes explicit the full proposition that the location is predicated of a certain individual, the nectar. Similarly, if a word 'butter' is flashed to a person very quickly, they may form an active BUTTER representation, but not represent explicitly that the word in front of them has the meaning butter (subliminal perception). The sentence 'I see the nectar is 100 yards south' or 'I see the word in front of me is butter' makes even more explicit; namely, the fact that one knows the proposition. Explicit representation is fundamental to consciousness (for the relation between representation and consciousness, see Representational Theories of Consciousness, and Contents of Consciousness, this volume).

IV. Application in Sciences:

Despite foundational problems, research in psychology, functional neuroscience and computer science has proceeded using the notion of representation. The tacit working definition is in most cases covariation between stimuli and mental activity inferred from behaviour or neural processes.

IV.1. Cognitive Neuroscience:

Cognitive neuroscience investigates the representational vehicle (neural structures) directly. The intuitive notion of representation relies heavily on covariation. If a stimulus (including task instructions for humans) reliably evokes activity in a brain region it is inferred that this region contains the mental representations required to carry out the given task. Particularly pertinent was the success of finding single cortical cells responsive to particular stimuli (single cell recording), which led to the probably erroneous assumption that particular features are represented by single cells, parodied as "grandmother cells". In the research programme investigating the Neural Correlates of Consciousness (often abbreviated NCC), conscious contents of neural vehicles are determined by correlating verbal reports with neural activities. Where the NCC are, and if any such thing exists, is an open and active current area of research.

IV.2. Cognitive Psychology:

Cognitive psychology assumes that experimental stimuli and task instructions are mentally represented. Theories consist of postulating how the representations are transformed. These representations are modelled and task performance (errors or reaction times) predicted. The quality of predictions provides the evidential basis of mental representations. For the investigation of longer reasoning processes sometimes the introspective verbal report is taken as evidence for the underlying mental representations. The enterprise is predicated on the notion that not only do people represent by having mental states, but thinking, perception and memory in turn can be explained in terms of sub-personal representations. The implicit assumption is that sub-personal intentionality will be naturalised, uniting cognitive psychology with the rest of the sciences.

One extensively argued issue is whether or in what domains thinking is symbolic or non-symbolic (normally, connectionist). A representation is symbolic if a token of it can be copied in different contexts and it retains the same meaning (normally explained because it has the 'same shape'). The notion originates from the use of a central processor in normal serial computers. Because different tokens of a symbol in different lines of program will only be processed when they pass through the central processor, of course different tokens can all be treated the same way. It is not clear that the same logic applies to the brain, for which there is no CPU. Nonetheless, Fodor has argued that

the symbolic style of representation is the only theory we have for why people think in systematic and indefinitely productive ways.

Connectionist networks potentially have two types of qualitatively different representation types: patterns of activation of units and patterns of weight strength between units. Given a teleological theory of content, for example, a learning algorithm may bring it about that a pattern of activation maps onto some external state of affairs so that the rest of the network can do its job. Then the pattern of activation represents that state of affairs. Similarly, the learning algorithm may bring it about that a pattern of weight strengths maps onto enduring statistical structures in the world so that the rest of the network can do its job. Then the pattern of weight strengths represents the external statistical structures. Psychologists naturally think in terms of activation patterns being representational because they naturally think in terms of representations being active or not. In the same way, sometimes they regard weights as non-representational. Weights on a teleological story are representational. They can misrepresent statistical structure in the environment so they can represent it. The process by which knowledge becomes embedded in weights in people is an example of implicit learning, the acquisition of unconscious knowledge.

Cognitive approaches to perception can be contrasted with ecological and sensorimotor approaches; these latter attempt to do without a notion of representation at all. The anti-representationalist style of argument often goes like this: We need less rich representations than we thought to solve a task; therefore representations are not necessary at all. The premise is a valuable scientific insight; but the conclusion a non-sequitur. To take a case in ecological psychology, while people in running to catch a ball do not need to represent the trajectory of the ball, as might first be supposed, they do represent the angle of elevation of their gaze to the ball. The representation is brutally minimal, but the postulation of a content-bearing state about *something* (the angle) is necessary to explain how people solve the task. There is a similar conflict between cognitive (representational) and dynamical systems (anti-representational) approaches to the mind, e.g. development. For example, how children learn to walk may be best understood in terms of the dynamic properties of legs getting bigger, the muscles getting elastic in certain ways; nothing need change in how the child represents how to walk. The debate has focussed people's minds on a very useful research heuristic: Try to work out the minimal representations needed to get a task done. Only when such a simple story has been discredited, then increase the complexity of the content of the postulated representations. Finally note that all accounts must eventually explain how personal level representation is possible, as most people accept that people do represent the world.

IV.3. Artificial Intelligence:

The traditional approach (symbolic AI, or Good Old Fashioned AI, GOFAI) assumed that the mind can be captured by a program, which is designed to have building blocks with representational content (derived intentionality). The New Robotics is a reaction to trying to program explicitly all required information into an AI, a strategy that largely failed. The new strategy is to start from the bottom up, trying to construct the simplest sort of device that can actually interact with an environment. Simplicity is achieved by having the device presuppose as much as possible about its environment by being embedded in it; the aim is to minimise explicit representation where a crucial feature need only be implicitly represented. In fact, the proclaimed aim of New Robotics is sometimes to do without representation altogether (see Clark, 1997).

Words: 3900

Zoltán Dienes
University of Sussex

Josef Perner
University of Salzburg

RECOMMENDED READING

Clark, A. (1997). *Being there: Putting brain, body and world together again*. MIT Press.
Crane, T. (2005). *The mechanical mind*, 2nd Edition. Routledge.
Cummins, R. (1996). *Representations, targets, and attitudes*. MIT Press
Dennett, D. (1991) *Consciousness explained*. London: Penguin.
Dienes, Z., & Perner, J. (2002). A theory of the implicit nature of implicit learning. In French, R M & Cleeremans, A. (Eds), *Implicit Learning and Consciousness: An Empirical, Philosophical, and Computational Consensus in the Making?* Psychology Press (68-92).
Greenberg, M. & Harman, G. (2006). Conceptual Role Semantics. In E. Lepore & B. Smith (eds) *The Oxford Handbook of Philosophy of Language*. Oxford University Press.
Kim, J. (2006). *The philosophy of mind*. 2nd edition. Westview Press.
Neander, K. (2004). Teleological theories of mental content. In the *Stanford Encyclopedia of Philosophy*.
Perner, J. (1991). *Understanding the representational mind*. MIT Press.
Searle, J. (2004). *Mind: A brief introduction*. Oxford University press.
Sterelny, K. (1990). *The representational theory of mind: An introduction*. Blackwell.