PAUL SCHWEIZER

School of Informatics

University of Edinburgh

paul@inf.ed.ac.uk

Conceptual Inversion and the Symbol Grounding Problem

Abstract: The paper explores the symbol grounding problem endemic to computational theories of mind. I argue that the intransigence of the problem indicates a basic incompatibility among the set of assumptions that has engendered it. Hence dissolution is the appropriate form of solution, and this requires an inversion of the conceptual framework in which the problem arises. Under this inversion, a naturalistic approach to mentality is taken as fundamental, while the traditional criterion of intentionality is abandoned.

Key words: computational theory of mind, intentionality, naturalism, problem of representation.

1. Introduction

According to the traditional conception of the mind, semantical content is an essential feature distinguishing mental from non-mental systems. In the scholastic tradition revived by Brentano, the defining aspect of mental states is their 'aboutness' or intrinsic representational aspect. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as 'mental

representations' are adopted as a primary theoretical device. For example, in classical (i.e. Fodorian) cognitive science, Brentano's legacy is preserved in the view that the properly cognitive (as opposed to implementational) level is distinguished precisely by appeal to internal representations. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars, and according to Fodor, it is only when the states of a system are treated as *representations* that we are dealing with the distinctively cognitive level.

On the classical Fodorian model, representations are thus posited as the internal structures that carry the information utilized by intelligent systems, and in addition they comprise the formal elements over which cognitive computations are performed. So these posited structures are meant to serve the dual roles of encoding semantical content, thereby supporting the idea that cognitive science is truly a science of the *mental*, and providing formal structures which can be manipulated according to well defined mathematical rules, thereby supporting the idea that it is truly a *science*.

However, as I will argue below, this crucial combination of roles incorporates a fatal inner tension. And since the tension involves the traditional conception of mind, the root of the difficulty can already be discerned in Brentano's (1874) well known position on natural versus mental phenomena. He contends (roughly) that

(1) the mind is characterized by intentional content

(2) intentional content cannot be explained naturalistically, therefore

(3) the mind cannot be explained via natural events and processes.

Tenet (3) is obviously repugnant to a large number of contemporary theorists, who strive to 'naturalize' mental phenomena and thus place the study of mind within the general context of the modern scientific worldview. Many such theorists are nonetheless concerned to insure that the phenomena thus naturalized are indeed properly *mental* in character, and as above, one way of doing this is to adopt the notion of representation as essential to the scientific understanding of the mind. Such motivations then lead to the currently mainstream position in which Brentano's conclusion (3) is denied, while some version of his central premise (1) is accepted. This in turn requires an attempted refutation of premise (2), which then fuels the quest for a computational cum naturalistic account of *intentionality*. In the case of cognitive science and artificial intelligence, this quest is often described as the attempt to solve the 'symbol grounding problem', which is the problem of how to connect the symbolic elements of internal computation to the external objects and states of affairs which they are supposed to represent.[1]

In order to be clear about the issue, it's worth mentioning that there are various possible 'weak' readings under which symbol grounding isn't a fundamental problem. For example, one could simply take an operational approach and *assign* a meaning or representational value, on the basis of what one takes to be the most convenient description of the system in question. For example, one might reasonably be inclined to say that the state of the metallic switching element in a thermostat 'represents' the room temperature. On a weak reading, this representational description picks out a straightforward correlation that is supported by the normal physical properties of the metallic element and the functioning of the temperature control device, and does not rely

on any mysterious aspects of 'directedness'. So even though there's a weak and perfectly cogent sense of representation in this case, the state of the metallic strip isn't literally 'about' anything, and the thermostat certainly isn't a mind.

On such a reading, representation talk serves a useful heuristic role, but it remains a conventional, observer-relative ascription, and accordingly there's no independent fact of the matter, and so there isn't a sense in which it's possible to go wrong or be mistaken about what an internal configuration is 'really' about. Instead, representational content is projected onto an internal structure when this plays an opportune role in characterizing the overall processing activities which govern the system's interactions with its environment, and hence in predicting its salient input/output patterns. But it is simply a matter of convention and choice.

This is in sharp contrast to the strong, literal view of intentionality, according to which there *is* an objective fact of the matter. On the strong view, an intrinsic feature of the 'directedness' of various mental states is that they are inherently about one external object rather than another, and indeed, different intentional states can be individuated in terms of their differing objects. Then the problem is to supply rigorous philosophical criteria for determining the correct or 'real' content of a given representation – ideally giving necessary and sufficient conditions for ascertaining in the general case when internal configuration *i* represents external object *o*. It also becomes pertinent to address various conceptual issues attending the full-blown notion of representation, such as a systematic account of *mis*representation, solving the disjunction problem, etc.

In this respect, I concur with the fairly widespread view that all attempts to solve the traditional problem of representation in computational and/or naturalistic terms have fallen far short of their goal, and further, that there is not much of evident promise on the theoretical horizon.[2] But rather than seeing this as an indication of the profundity and importance of the problem, and hence as a motivation for further research in the direction of solving it, instead I think it indicates that we should step back from the issue and re-examine its conceptual origins. According to the view advocated herein, the underlying import of the symbol grounding problem is that it reveals a basic flaw in the 'mainstream' position that has spawned it. Along these lines I argue that a conceptual inversion is called for, wherein *dis*solution is the appropriate form of 'solution'.

The symbol grounding problem arises within a theoretical project committed both to the traditional, pre-scientific characterization of the mind, and to the explanatory/ontological framework of the natural sciences, i.e. to tenets (1) and the negation of (3) above. I would suggest that the intransigence of the problem betrays a deep incompatibility between these two commitments, indicating that at least one of them must be abandoned. And from the perspective of an attempted scientific explanation of mentality, the choice of which commitment to reject is quite clear. The goal of achieving a *computational/naturalistic* theory of mind is the driving force behind cognitive science and AI, which means that a rejection of Brentano's conclusion (3) must be taken as fundamental. With this as our starting point, the quandary then arises when (1) is adopted as central to a correct analysis of the mind, which then forces the rejection of (2) and the attendant need to naturalize content.

But as maintained below, I agree with Brentano that the traditional notion of content cannot be naturalized. Hence, in sharp contrast to the views held by many writers in philosophy of mind, cognitive science and AI, a basic theme of the present paper is to endorse a generalized version of Brentano's second premise. From this standpoint, it is then a critical mistake to embrace (1) within the context of naturalism, and the salient way to promote the negation of (3) while maintaining consistency is simply to reject the traditional definition of the mind.

If the goal of our theoretical enterprise is to explain the activities of actual systems occurring in nature, then our conception of the mind requires drastic revision in order to fit within this restricted framework. And this should hardly come as a surprise, given that the customary, pre-scientific notion is an *a priori* legacy, and is not based on any empirical findings or biological constraints. In contrast, the tenor of scientific investigation is clearly more descriptive than definitional, and the mainstream predicament stems from a very odd methodology. It accepts a pre-scientific, *a priori* view of the mind, and then construes the resulting lack of fit between this very strong traditional notion and the limited resources of naturalism as some sort of deep theoretical problem. Instead, a more suitable tack would be to modify our concept of the mind in response to the empirical and metaphysical limitations entailed by normal science.

As a rejoinder, some traditionalists might contend that (1) cannot be rejected because it's a 'conceptual truth', it's an essential part of what we *mean* by the term 'mentality', and hence any theory that fails to embrace (1) will, ipso facto, fail to be a theory of the mind. But such a response is unconvincing, even from the traditional standpoint, because it embodies a confusion regarding the existential import of

conceptual analysis. Even if, for the sake of argument, it were conceded that such analysis revealed that (1) is a basic feature of what is included in the pre-scientific concept of 'mentality', still this tells us nothing about whether the concept is *correct*, i.e. whether it picks out or accurately corresponds to anything in the actual world. Within the standard framework of philosophical analysis, it's possible to elucidate a concept, an intensional specification, but in itself this cannot resolve the question of whether the concept has an extension, i.e. whether the specification is satisfied by anything in the realm of objective reality.[3]

Along similar lines, the original Greek conception of 'atom' meant an *indivisible* micro-unit of physical substance, but this original meaning has not stopped scientists from adapting the semantics of the term to correspond with the empirical facts. Similarly, the onus now lies on the traditionalist to show that there are real systems and properties that answer to the conceptual requirements of tenet (1). And I would take Brentano's initial rejection of naturalism, along with the apparent intractability of the symbol-grounding problem, as providing very firm grounds for skepticism.

Of course, a truly stalwart defender of the tradition could take a more extreme line. One could adopt Brentano's strategy and accept both (1) and (2), along with the existential affirmation that *there is something* which instantiates tenet (1). This yeilds (3), along with the ancillary claim that there is more to objective reality than the world of natural science - the 'truth maker' for (1) must then reside in some ontological realm transcending the bounds of the natural world. While this move does at least possess the virtue of maintaining a mutually compatible set of basic assertions, it takes the dialectic in a very different direction from the concerns of the present paper. In terms of the

current argument, the relevant point to note is that the mere acceptance of (2) does not force such a move, nor does it constitute a case for expanding the ontological horizons required to explain mentality, because one could instead choose the line advocated herein, which is to deny that 'original intentionality' is a genuine phenomenon.[4]

2. Computational Solipsism

In the following three sections I will present some general considerations in support of Brentano's claim that intentionality cannot be explicated in natural terms. First I'll examine the symbol grounding problem that arises within the computational paradigm, because this problem constitutes a very clear instance of the general difficulty, and it provides the most recent idiom in which the discord between intentionality and naturalism has been expressed. As mentioned in the Introduction, mental representations in cognitive science are meant to play a crucial double role - they are designed to serve as the internal repositories of meaning, and they comprise the formal elements over which cognitive computations are performed. In this manner, the semantical operations of the mind are supposed to be effected in terms of rule governed transformations on a system of internal syntax. And it is here that the fundamental discord becomes apparent: a central purpose of representations is to carry content, and yet, to the extent that they are formal elements of computation, their alleged content is completely gratuitous.

Computation is essentially a matter of manipulations performed on *uninterpreted* syntax, so that formal structure alone is sufficient for all effective procedures: the specification and operation of such procedures makes no reference whatever to the intended meaning of the symbols involved. Indeed, it is precisely this limitation to

syntactic *form* that has enabled computation to emerge as a mathematically rigorous discipline. If syntax alone is not sufficient, and additional understanding or interpretation is required, then the procedure in question is, by definition, not an effective one. But then the purported content of mental 'representations' is rendered superfluous to the algorithms that comprise the 'cognitive' processes of cognitive science. The distinguishing criterion of mentality is lost, since the intended interpretation of the syntax makes absolutely no difference to the formal mechanics of mind.[5]

And many classical *negative* results in mathematical logic stem from this separability between formal syntax and meaning. The various upward and downward Löwenheim-Skolem theorems show that formal systems cannot capture intended meaning with respect to infinite cardinalities. As another eminent example, Gödel's incompleteness results involve taking a formal system designed to be 'about' the natural numbers, and systematically reinterpreting it in terms of its own syntax and proof structure. As a consequence of this 'unintended' interpretation, Gödel is able to prove that arithmetical truth, an exemplary *semantical* notion, cannot, in principle, be captured by finitary proof-theoretic means.

These (and a host of other) powerful results on the inherent limitations of syntactical methods would seem to cast a rather deflationary light on the project of explicating *mental content* within a computational framework.[6,7] Indeed, they would seem to render hopeless such goals as providing a computational account of natural language semantics or propositional attitude states. Non-standard models exit even for such rigorously defined domains as first-order arithmetic and fully axiomatized geometry. And if the precise, artificial system of first-order arithmetic cannot even impose isomorphism

on its various models, how then could a *program*, designed to process a specific natural language, say Chinese, supply a basis for the claim that the units of Chinese syntax posses a *unique* meaning?

Computational formalisms are syntactically closed systems, and in this regard it is fitting to view them in narrow or solipsistic terms. They are, by their very nature, independent of the 'external world' of their intended meaning and, as mentioned above, they are incapable of capturing a unique interpretation, since they cannot distinguish between any number of alternative models. This can be encapsulated in the observation that the relation between syntax and semantics is fundamentally *one-to-many*; any given formal system will have arbitrarily many different interpretations. This one-to-many character obviates the possibility of deriving or even attributing semantical content merely on the basis of computational structure, and hence the symbol-grounding problem is insoluble within the solipsistic framework of pure computation.

John Searle's celebrated Chinese Room Argument (CRA) directly exploits this solipsistic aspect of rule governed symbol manipulation, and his critique of 'strong' AI and the allied computational paradigm in cognitive science first brought symbol grounding to the fore as a serious theoretical problem for these approaches. So from the perspective of the present discussion, it is instructive to recast Searle's influential argument in terms of the separability of syntactical structure from its intended meaning. Thus formulated, it can clearly be recognized as a computational variation on Brentano's general theme. In what follows I will abstract away from the pictorial details of Searle's original version and express the logical core of the CRA via two premises and a conclusion:

(1) the mind possesses original semantical content

(2') syntactical manipulations cannot capture this content, therefore

(3') the mind cannot be reduced to a system of syntactical manipulations.

According to the foregoing metamathematical considerations, premise (2') is true because of the one-to-many relation. There is no determinate model which can be posited or recovered merely on the basis of syntax, so computational theories of mind are unable to yield mental semantics. Furthermore, on Searle's view it is the mind's original intentionality which supports various derivative forms of intentional phenomena such as linguistic semantics, and hence the reference relation for natural languages will also fail on the computational account. The CRA entails that mere syntactic manipulation of the string 'h-a-m-b-u-r-g-e-r' will never be sufficient to determine the culinary transgression to which the term refers. Hence if Searle is granted Brentano's central thesis, it then follows that *we* have something essential that eludes the resources of the computational approach.

But, to clarify my position with respect to the longstanding conflict between Searle and the computationalists, it's important to note that my critique applies *both* to Searle and to the cognitive science doctrine that he attacks. This is because both parties to the dispute accept some form of premise (1), while denying some form of (2) and hence (3). Classical cognitive science views the mind according to the model of rule governed symbol manipulation, and premise (1) is embraced insofar as the manipulated symbols are supposed to possess representational content. Searle's dispute with cognitive science centers on his rejection of the idea that internal computation can shed any real light on mental content, which leads to his specialized conclusion (3') and his concomitant

dismissal of the research paradigm central to cognitive science and AI. In turn, the standard line for defenders of this paradigm is to try and defuse the CRA by arguing against (2'), which of course engenders the symbol grounding problem.

However, I would urge that this is a serious mistake for those who wish to defend the computational approach, since, for the reasons already given, I think that Searle's (2') is clearly true. So I agree with the negative point that computation is too weak to underwrite any interesting version of (1), and I concur with Searle's reasoning to the extent of accepting the salient *conditional* claim that *if* (1) is true *then* (3') is true as well. So the crux of the issue lies in the truth-value of (1), without which the consequent of the conditional cannot be detached as a free-standing conclusion. Only by accepting the traditional, *a priori* notion of mentality does (3') follow from the truth of (2'). And it's here that I diverge from the views of both Searle and orthodox cognitive science.

In contrast to typical defenders of cognitive science and AI, I'd retain computationalism by rejecting (1) rather than (2'). And even though I accept Searle's pivotal premise (2'), his views still fit the 'mainstream' template to which my overall discussion stands in opposition. Searle famously eschews computationalism, but he nonetheless accepts Brentano's central thesis, while simultaneously embracing his own particular version of naturalism. So while Searle's views and my own are in accord with respect to the truth of the very specialized (2'), Searle would nonetheless reject a wider reading of Brentano's second premise.

According to Searle (1990, 1992), it is the fact that the human mind sustains conscious presentations of content which serves as the basis for his claim that the mind possesses original intentionality. It is consciousness rather than computation which Searle

takes to be the basis for the truth of (1), and he further believes that consciousness arises from physical brain activities rather than from symbol manipulation. Hence intentionality is tethered to brain processes via consciousness, and Searle thereby attempts to naturalize the traditional notion of mentality, while at the same time discrediting the computational paradigm.[8]

And while I agree with Searle's view that consciousness arises from physical brain activities rather than from abstract computational structure, I would nevertheless argue that conscious experience, just like symbol manipulation, is too weak to underwrite any interesting version of tenet (1). In accord with the view that conscious experience is the cornerstone of intentionality, the CRA presupposes that the homunculus Searle, replete with conscious presentations, *really does* understand English in some special way. Searle appeals to himself as the locus of genuine intentionality in the Chinese Room, and he would support this by citing the fact that he is consciously aware of the meanings of English expressions. Ostensibly, this special understanding of English enables him to follow the program and manipulate the 'meaningless' Chinese symbols. Hence lack of conscious presentation with respect to the semantics of Chinese constitutes the real asymmetry between the two languages, and this underlies Searle's claim that genuine understanding occurs in the case of one language and not the other.

But it is clearly possible to concede that Searle has episodes of conscious awareness which attend his processing of English, while still denying that these episodes are sufficient to establish intrinsic content, or to ground the semantics of natural language expressions. Instead, what consciousness does provide is the foundation for the subjective *impression*, had by Searle and others, that the human mind enjoys some mysterious and

transcendent form of intentionality. Thus when Searle contends that our mental states are 'really about' various external objects and states of affairs, this is merely an expression of the fact that, introspectively, it *seems to us* as if our mental states had some such special property. Conscious experience is clearly sufficient to provide the source for this belief, since conscious experience determines how (some of) our mental states appear to us. But it cannot provide a basis for concluding that the belief is *true*, unless consciousness is something much more mysterious and powerful than naturalists can consistently allow. Brentano dismissed naturalism, and he thereby gave himself some room for the claim that consciousness underwrites the mind's essential intentionality. However, if one accepts naturalism and views consciousness as a phenomenon supported by, say, the causal properties of electrochemical reactions taking place inside the skull, then one should just bite the bullet and admit that it is too weak to support Brentano's central thesis.

3. Syntax Contextualized

A standard move in response to the symbol grounding problem as expressed in the previous section is to append a naturalistic exegesis to the computational analysis, whereby cognitive syntax is assigned a 'meaning' via the consideration of assorted causal chains and physical interactions between the system and its environment. These considerations appeal to a myriad of factors, including direct relations of cause and effect, evolutionary trajectories, functional teleologies, initial baptisms, covariation, asymmetric dependence, communal linguistic behavior, natural signs, information flow, innate structures triggered by external stimuli, embedded computation, etc.

Abstracting over variations in the many different strategies put forward, the overall goal is to imbue internal processes with content by grounding them, somehow or other, in an encompassing causal or biological nexus.[9] But, whatever other merits this type of project may possess, it seems clear that it doesn't answer the question at issue, because, as I argue below, the strategy of formalism plus environment is still afflicted by the same one-to-many syndrome as before. The original symbol grounding problem arose because the relationship between syntax and semantics is intrinsically underdetermined; for any given formal system there will be arbitrarily many different models that satisfy it, and this obviates the notion that a unique interpretation can be grounded on formal structure alone. And this is what drives the need to add new factors to the story. The hope is that formal structure *plus* some type of salient causal ties will suffice to fix the intended referents.[10]

However, this sanguine hope is critically undermined by the fact that the computational paradigm makes cognitive science and AI deeply committed to a narrow reading of mental states. If the mind is to be identified with a physically realized computational procedure, then the boundaries of the mind are set by its formal input/output specifications. The mind cannot directly interact with objects in the environment; rather, these objects must impinge upon its 'surface' and be translated or transduced into the appropriate form of input *signal*, and the mind operates by processing these structured inputs rather than their remote causes. In turn, the input signals cannot be exploited in the reverse direction to uniquely specify the objects that produced them, because these surface effects cannot distinguish between any number of *different* remote sources capable of producing the very same signals. So, for example, they are in principle

incapable of distinguishing between sufficiently refined virtual environments and real ones.

Similarly, the philosophical literature is replete with imaginative scenarios wherein envatted brains are fed computer generated stimuli resulting in subjective experiences that are identical to those accompanying veridical perception in normal agents. But we don't even need to invoke fanciful sci-fi scenarios to generate the problem: the same one-to-many syndrome also occurs in more mundane, low-tech cases, such as the failure to discriminate between two qualitatively 'identical' but numerically distinct objects that have been swapped without the subject's knowledge. The sensory inputs are the same while the respective causes are two physically distinct entities, with the result that the subject's perceptual experience is insufficient to determine its distal object. This shows that the 'aboutness' of sensory representation cannot be grounded by appeal to causal traces to be found *within* an environmentally embedded cognitive structure. Even in the simplest case of direct sensation, the one-to-many aspect of the symbol grounding problem cannot be solved along causal/environmental lines, because these factors simply give rise to the structurally parallel and equally intransigent 'sensory input grounding problem.'

So, if the mind is construed narrowly, then we're back in the same boat as before. Even when cognitive states are viewed not just as computational processes, but as complex internal effects of outside circumstances, still the internal state itself is incapable of determining a unique cause as its source. From the perspective of the effect, the relation with possible *natural cause* is still one-to-many, and therefore, within a narrow framework, causal stories suffer from exactly the same general underdetermination that

afflicted the pure computational approach. Given any internal configuration of a cognitive system, be it computational, neurophysiological or conscious/ phenomenal, no unique physical event can be recovered as its cause.

And theoretical commitment to *narrow* mental states runs much deeper than the computational paradigm alone. As Fodor, Block, Stich and many others aptly maintain, the causal/explanatory role of mental states (independently of specialized computational assumptions) requires them to be defined narrowly, as *internal* configurations of a cognitive system. This is because the ultimate goal is to explain a system's intelligent *behavior*, and the causal locus of such behavior lies within the organism. This idea is succinctly expressed in Stich's 'principle of psychological autonomy' whereby "the properties and relations to be invoked in an explanatory psychological theory must be supervenient upon the *current, internal physical* properties and relations of organisms…" (p. 260, his italics). The principle of psychological autonomy underlies the cognitive science research paradigm in particular, as well as most orthodox frameworks of psychological explanation. And if a computational or (narrowly) naturalistic model is adopted to explain intelligent human behavior, then the symbol grounding problem cannot be solved. Rather, a conceptual inversion is called for, under which the problem is dissolved by denying that *human* mental states possess intrinsic content or aboutness.

Lest this inversion appear too extreme, it is worth recalling the weak sense of 'representation' mentioned in the Introduction, under which symbol grounding is not a fundamental quandary. To take an oversimplified example merely for the sake of illustration, suppose that researchers observed a regular correlation in which the presence of a giraffe in some system's perceptual field always caused a specific region of it's brain

to light up. Then for various reasons it might be expedient to say that the region in question 'represents' the presence of a giraffe. A dependable correlation obtains, which in turn may be useful for explaining the internal processing and subsequent behavior of the system. But in the end, this correspondence is simply a matter of cause and effect, and such causal relations are notoriously too weak to sustain the traditional notion of 'real' aboutness. They may supply a useful heuristic for predicting the behavior of a system in its customary surroundings, but this falls fall short of grounding original intentionality.[11]

Of course, it's possible to invoke more subtle and sophisticated considerations, such as evolutionary history and biological teleology, to try and refine the correspondence relation and support a richer version of representational content. It's well beyond the scope of the present discussion to attempt a detailed critique of all such manoeuvres, or to give decisive arguments against each of the many alternative strategies put forward by those who wish to naturalize intentionality. However, my general position is that, to the extent that such strategies are truly naturalistic and do not presuppose or smuggle intentionalistic primitives in through the back door, then what has been naturalized will turn out to be some version of the weak, operational notion. Hence these strategies do not overturn Brentano's second premise, because what has been scientifically rehabilitated is in fact not the genuine article.

Brentano utilized metaphors such as the mind's 'aim' to explicate the traditional notion of directedness, as if an intentional arrow were to emanate from the head and connect to a chosen external referent. And indeed, something like this is exactly what would be required to underwrite the intuitive notion: internal state *i* represents external object *o* iff state *i* emits an intentional arrow connecting the mind to *o*.[12] And clearly this

is a relation that cannot, even in principle, be rendered in naturalistic terms - there are no tangible correlates for this intentional arrow or 'noetic ray'. Naturalistic accounts are not able to yield directedness *towards* something external, because the natural 'lines of force' are all in the direction from object to mind. There are no physical forces that go outward, from mind to object, to establish that the thing correlated to the organism by the relevant causal history is also the *object* of an intentional state. In terms of energy and 'information' flow, the mind is simply a recipient and processor of outside influences. This processing may well lead to outputs such as bodily motions, but the combination of solipsistic processing and *physical* action is not enough to ground a concomitant relation of 'aboutness'. Causal and biological histories may, in some cases, supply a necessary condition, but without a non-natural link in the direction from internal state to object, they will never suffice to capture the traditional notion.

In order to locate my view on the salient philosophical landscape, it is worth noting that it is distinct from the positions held by various proponents of dynamical systems theory (e.g. Van Gelder, 1996), behavior based robotics (e.g. Brooks, 1991), etc., who advocate 'intelligence without representation'. The latter views concern the basic form of cognitive architecture required to sustain intelligence, and this is largely an empirical question regarding what types of structure are present and/or required in actual cognitive systems. In contrast, the inversion called for in the present discussion is not concerned with architectural details *per se*, but rather with how such details should be philosophically construed.

On my view, there could well be internal structures that fulfil many of the requirements that people would ordinarily expect of representations, and this is especially

true at the level of sensorimotor control, perception and navigation – things like spatial encodings, somatic emulators, internal mirrorings of salient aspects of the external environment. So, unlike the anti-representationalists, I do not deny that there may be internal structures and stand-ins that various people would be tempted to *call* 'representations'. But I would argue that this label should always be construed in its weak, operational sense, and not be conflated with the traditional conception. There is nothing about these internal structures that could support the notion of original intentionality, and there is no independent fact of the matter regarding their content or status. If this point were to be condensed into a competing slogan, then instead of 'intelligence without representation' I would propound 'representation without intentionality'.

So what I deny is not that there may be internal mechanisms that reflect external properties in various systematic and biologically useful ways. Instead I would deny that there is anything more to this phenomenon than highly calibrated relations of cause and effect within some specialized environmental context. If one is truly committed to naturalism, then there is only a difference of degree and not kind between, say, the reflection of moonlight in a pond and the retinal image of the moon in some organism's visual system. Proponents of the mainstream view are inclined to think that a sufficient difference in degree and complexity somehow yields an esoteric difference in *kind*, a difference that allows us to cross the conceptual boundary from mere causal correlations to 'genuine aboutness'. But I would contend that naturalism itself supplies an asymptotic limit for this curve, and that the boundary can be crossed only by invoking non-natural forces.

According to the position advocated herein, Fodor's characteristic insistence on representational *content* embodies a serious confusion within the context of naturalistic explanation. The crucial point to notice is that these internal 'representations' do all their scientifically tangible *cognitive* work solely in virtue of their physical/formal/mathematical structure. There is nothing about them, qua efficacious elements of internal processing, that is 'about' anything else. Content is not an explicit component of the input, nor is it acted upon or transformed via cognitive computations. All that is explicitly present and causally relevant are computational structure plus supporting physical mechanisms, which is exactly what one would expect from a naturalistic account.

In order for cognitive structures to do their job, there is no need to posit some additional 'content', 'semantical value', or 'external referent'. From the point of view of the system, these internal structures are manipulated directly, and the notion that they are 'directed towards' something else plays no role in the pathways leading from cognitive inputs to intelligent outputs. Hence the symbol grounding problem is a red herring – it isn't necessary to quest after some elusive and mysterious layer of content, for which these internal structures serve as the syntactic 'vehicle'. Syntactical and physical processes are all we have, and their efficacy is not affected by the presence or absence of meaning. Indeed, the postulation of content as the essential feature distinguishing mental from non-mental systems should be seen as the last remaining vestige of Cartesian dualism, and, contra Fodor, naturalized cognition has no place for a semantical 'ghost in the machine'.

4. Behavior versus Meaning

As observed in previous sections, the symbol grounding problem arises because of the persistent one-to-many relation between internal cognitive factors and external objects of reference and representation. The 'objective' content of a mental state cannot be recovered from any feature of the state itself, so the unique two-place relation that purportedly obtains between mind and world, between internal structure and represented object, cannot be reduced to any one-place property of the mind.

Mental representations and natural language semantics clearly have many intimate philosophical connections, and the foregoing one-to-many relation has acute consequences for the linguistic theory of meaning. If one accepts the principle of psychological autonomy, then the mind is too weak to determine what its internal components are

'really about', and this extends to the case of expressions in natural language as well. The famed conclusion of Putnam's Twin Earth argument is that 'meanings ain't in the head', and this is because narrow psychological states are incapable of determining the reference relation for terms in our public languages. But rather than abandon natural language semantics in light of the problem, the externalist quite rightly abandons the traditional idea that the intentionality of mental states provides the foundation for linguistic reference.

Putnam's well known strategy is to directly invoke external circumstances in the characterization of meaning for natural languages. The externalist strategy exploits direct, ostensive access to the world, thus circumventing the difficulty by relieving mental states of their referential burden. On such an approach, the object of reference can only be

specified by indexical appeal to the object itself, and in principle it *cannot* be determined merely from the psychological states of the language user. Direct appeal to the actual environment and linguistic community in which the cognitive agent is situated then plays the principal role in determining the match-up between language and world.

Putnam's strategy offers a viable account of linguistic reference *precisely because* it transgresses the boundaries of the mind intrinsic to the explanatory project of cognitive science. The externalist must invoke broad environmental factors, since nothing internal to a cognitive system is capable of uniquely capturing the purported 'content' of its representations and thereby semantically grounding its internal states. And from this it follows that original content is not a property of the representation *qua* cognitive structure, and hence it is not the cognitive structure itself that provides the theoretical basis for meaning. Indeed, outside factors then do the real work, and the *semantical* role of internal configurations is trivialized.[13]

However, in normal, everyday practice, we continually use sentences of public language to ascribe various content bearing mental states, both to ourselves and others. A defender of the tradition might argue that the *truth* of such ascriptions shows that there is still a genuine fact of the matter regarding mental content, and hence that there is an objective match-up problem remaining to be solved. When an agent is correctly attributed a given propositional attitude, such as the belief that φ, this describes an authentic feature of their doxastic configuration, and must be supported by some corresponding aspect of their internal make up.

But such a line of argument would gravely misconstrue our common sense practices, because the age-old customs of 'folk psychology' are independent of any

assumptions about internal symbols or representational structures. Observable behavior and context are the relevant criteria, and the truth-conditions for such ascriptions are founded on macroscopic, operational considerations. As in everyday life, one can use behavioral and environmental factors to adduce that, say, Jones believes that lager quenches thirst, but this practice makes no assumptions about the nature or even existence of internal representations. The attribution concerns Jones as an unanalyzed unit, a black box whose actions take place within a particular environmental and linguistic setting. It gives no handle whatever on postulating hidden internal cogs and levers that generate Jones' actions, and it's perfectly compatible with an agnostic disregard of such inner workings.

At this stage, an entrenched representationalist such as Fodor is likely to invoke the belief-desire framework of psychological explanation to defend a realist account of meaning. Not only do we ascribe various content bearing states to ourselves and others, but furthermore we habitually *use* such ascriptions to explain and successfully predict behavior. According to this widely accepted framework, psychological states individuated in terms of their *content*, such as beliefs and desires, are *causally* responsible for a host of rational actions. Thus the belief-desire framework can successfully predict behavior from the outside, because it mirrors the internal processing structure that causes it.

This is a key motivation behind Fodor's illustrious 'Language of Thought' hypothesis (LoT), which serves as a very clear and straightforward exemplar of many of the ideas behind a realist approach to mental content. So in the ensuing discussion I will scrutinize the LoT in particular, but the same points apply to a wide class of positions within the representational theory of mind. The LoT proposes a system of internal syntax

as the computational medium of psychological processes, where sentences of this internal language bear the content of the propositional attitudes. Thus when, from the outside, we justifiably ascribe to Jones the belief that lager quenches thirst, Fodor would have it that a token of some mentalese sentence, say 'n%^7 £#~ %&!+', which encodes the same semantical content as the English ascription, has been duly etched into her 'belief box'. This physical implementation of mentalese syntax is then poised to interact with other physically implemented tokens in her desire box to produce assorted forms of rational action, such as standing up and reaching for a pint. In this manner, the truth of propositional attitude ascriptions is directly correlated with salient internal configurations of the agent.

But this purported correlation breaks down at its most vital point – the level of semantical content. For the story to work, the sentences 'lager quenches thirst' and 'n%^7 £#~ %&!+' must both have the same meaning. Yet as a medium of classical computation, the Language of Thought is just a scheme for rule governed symbol manipulation. Syntax churning within a formal system is fundamentally different from the operation of a public language, and it is a significant mistake to impute to the former the same semantical properties conventionally attributed to the latter. English is acquired and exercised in an intersubjectively accessible context to which the entire sociolinguistic community has indexical appeal. There are shared criteria for the correct use of natural language sentences and the rules under which various expressions are deployed, and there are direct, ostensive ties between publicly produced syntactic tokens and their referents. In vivid contrast, there are no such criteria nor ostensive ties for the hidden, internal sentences of mentalese. The LoT serves as an extreme example of a *private* language, and

as such it has no communal truth conditions nor standard semantic properties. Indeed, the LoT is so private it's even hidden from the introspective awareness of the individual agent, and it thereby also eludes Searle's traditional association of linguistic meaning with agent-based intentionality.

As elements in a formal system, there is no fact of the matter concerning what the internal sentences of mentalese 'really mean': here the one-to-many syndrome explored in the pervious two sections applies in full force. At best, these conjectured tokens of computational syntax would successfully govern our behavior in familiar surroundings, but they would not govern our behavior successfully if we were placed in radically different circumstances. So they are merely calibrated with the environment in which they happened to develop, and this fact doesn't imbue them with 'genuine content'. To the extent that these hypothetical symbols successfully govern behavior, they do so purely in terms of their formal, syntactical properties, and as noted before, there is no work left to be done by their intended interpretation. On a computational approach to the mind, it is processing *structure* and not semantics that is the cause of human action.

So as a piece of formal syntax, the mentalese string 'n%^7 £#~ %&!+' may be efficacious in terms of cognitive processing and behavior, but it doesn't *mean* the same as 'lager quenches thirst', any more than the physical state of the metallic strip in a thermostat has the same semantical properties as the English string 'the room is too cold and the heating should go on'. There are no adequate grounds upon which to assign a canonical content to mentalese syntax, and there is certainly no basis upon which to specify a mapping between mentalese and a public language such as English that preserves 'sameness of meaning'. Indeed, the LoT could have a semantics comparable to

English only if it served as part of a shared public language and were then *used* to make propositional attitude ascriptions from the outside.

The LoT hypothesis takes representational content and a belief-desire style of internal processing as its starting point, and it sees macroscopic behavior as a straightforward manifestation of these inner workings. Other proponents of the 'mainstream' view would favor a more operational, behavior-based approach to the issue of intentionality and computation. As above, we constantly ascribe various intentional states to other human beings, and this practice is based on observable activity in a particular physical environment and sociolinguistic setting. It has been argued that comparable standards should be applied to computational artifacts, and that attributions of intelligence and/or intentionality are warranted when the behavior displayed by the artifact would earn such attributions if it were human. This is the rationale behind the original Turing Test (TT), where conversational behavior that fools an interlocuter is deemed an adequate standard for intelligence.

An obvious fault with the original TT is that only *verbal* inputs and outputs are involved, and even on purely behavioral grounds this falls far short of the observable evidence we have with regard to fellow human beings, who *use* language in a complex world and perform many types of intelligent non-verbal behavior. In order to warrant ascriptions of intentionality on grounds approaching those used in daily life, at the very least we would need to include language-entry and language-exit abilities in the test. It's one thing to produce appropriate verbal input/output patterns regarding hamburgers – this is still operating within a closed syntactic bubble, as in the Chinese Room. It's quite another thing to break out of the bubble and *identify* an actual hamburger in one's spatial

vicinity, to reach for *it* rather than for a pizza or a kebab. This type of publicly observable language/object association would constitute primary evidence for the 'directedness' of the system.

Such considerations motivate an obvious scaling-up of the original test, to what Steven Harnad (1991) has dubbed the Total Turing Test (TTT). Under the TTT, the scrutinized artifact is a *robot*, and the relevant behaviors coincide with the full range of activities of which normal human beings are capable. This combined linguistic and robotic test constitutes a vast improvement over Turing's original format, since the scope for empirical data gathering now includes all those forms of complex and varied interaction typically exhibited by human beings. Harnad expresses a widely held view when he says that the TTT is "… is no less (nor more) exacting a test of having a mind than the means we already use with one another…" (p. 49).

Indeed, precisely this type of claim is made in the 'robot reply' to Searle's original Chinese Room Argument. Perhaps a program governing mere conversational behavior isn't enough, goes the reply. But suppose we could construct a robot equipped with television camera's enabling it to 'see', and arms and legs enabling it to 'act', and the robot were controlled by a computer and could function indistinguishably from a human. Then 'Such a robot would … have genuine understanding and mental states' (p. 192). The implicit reasoning behind this type of scenario is apparently something like this:

(a) human minds have original intentionality,

(b) the robot can do everything a human can do, so

(c) the robot must have this special property as well.

However, if one finds such an argument from analogy at all persuasive, then the conceptual inversion argued for in the present paper would run it in the reverse direction, *viz.*:

(b) the robot can do everything a human can do,

¬(c) the robot has no special property, so

¬(a) human minds don't either.

The only evidence we have for human intentionality is behavior, plus our own subjective experience of some of the conscious psychological states that attend this behavior. And the issue of robot consciousness can be factored out of the equation, since I have already argued that on a naturalistic approach, qualitative experience in the *human* case is too weak to ground the traditional notion, and hence it won't tip the scales for a robot either.

So we're left with behavioral evidence, and by hypothesis, the robot's behavior is indistinguishable from our own. The ascription of 'real' mental states would be based purely on the robot's performance, on its operational success. But the designers would achieve this success by utilizing (rather ingenious) computational and engineering solutions to a host of problems faced by physical organisms such as ourselves. So if the robot is deemed a genuine engineering possibility, then we should take this as a compelling reason to deny that the strong, traditional notion of intentionality truly applies to ourselves. And if the robot is not deemed to be a possibility, at least in principle, then I think one's credentials as a naturalist are brought into serious doubt.

In closing, it is worth distinguishing my view from some forms of 'eliminativism' with respect to the propositional attitudes and the belief-desire framework of psychological explanation. In contrast to the Churchlands and others, I do not deny that

there *may* be processing structures that play the role of Fodor's belief and desire boxes, internal sentences, etc. So I would not deny (in advance of weighty empirical evidence) that there may be some type of *operational* reduction of folk concepts to functional or neurophysiological states that were useful in predicting behavior. As in the case of the anti-representationalists, my position is not based on speculations about the non-existence of various elements as revealed by future scientific findings. Instead, my point is that even if there were such neural structures implementing an internal LoT, this still wouldn't ground traditional semantics and genuine aboutness – these structures would have the relevant causal/syntatctic properties but not the semantical ones.

Presumably there must be *something* going on inside of Jones that is causally responsible for her behavior, and which allows us to use propositional attitude talk to explain and predict relevant portions of it. But it simply doesn't follow that these internal states and mechanisms have the same *semantical* properties as the expressions in natural language that we use to provide our common sense gloss. So, while I wouldn't deny that there may be internal structures that do much of the work normally assigned to 'representations', there would still be no fact of the matter regarding, say, whether the frog's fly representation 'really means' fly, fly or BB pellet, or fly or BB pellet or anything else that makes it activate. The disjunction problem is ill-posed: it's motivation is from the side of those who want to challenge the idea that the traditional notion of representation has been satisfactorily captured, and it is a false step on the part of naturalists to take the challenge on board and try to show that it has. The salient response is to admit that it hasn't, and to transfer the onus back to the traditionalist by challenging Brentano's central thesis.

5. Conclusion

A primary theme of the paper has been to argue against the idea that original intentionality is a trait implemented by any system occurring in the natural world. In a closely related vein, I've argued against the classical Fodorian view that internal representations constitute the hallmark of the cognitive. For a number of reasons, the traditional notion of 'content' is unsuitable as a primary theoretical device in the scientific approach to mentality. One of the most serious problems with content is its total inefficacy. As elaborated in section 2 above, computation is essentially a series of manipulations performed on *uninterpreted* syntax, and formal structure alone is sufficient for all effective procedures. The purported content of mental representations is superfluous to mental processes viewed as computations, and the interpretation of internal syntax makes absolutely no difference to the formal mechanics of mind.

Indeed, if content weren't gratuitous, then computational versions of cognitive processing would be lamentably deficient in terms of their specification of the inputs, since content *per se* is entirely absent from the stimuli recognized by the operations that can be expressed within the computational paradigm. If representational content enjoyed any genuine role in the model, then cognitive systems would have to process content *itself*. But on the contrary, content is not specified with the inputs, it has no tangible influence on syntactic structure, nor does it play any role in internal processing. It is syntax or structure rather than meaning which is physically realized, and it is structure which governs cognitive computations. Semantical content must simply float above its formal and neurophysiological 'vehicles' like some sort of inert Cartesian substance,

doing no work but nonetheless comprising the alleged essence of the mind. Surely this is an exemplary occasion for invoking Ockham's razor. The science of mind doesn't need a semantical ghost in the machine - when it comes to computation and content, only the vehicle is required, not the excess baggage.[14]

This is not to deny that the cognitive level of description must abstract and generalize over the details of mere physical mechanisms in order to capture the patterns and regularities relevant to intelligent systems. But instead of positing intentional content as the criterion that distinguishes the cognitive level, I would retain the computational paradigm and jettison the Cartesian millstone. This leaves internal processing structure and macroscopic behavior as the appropriate level of abstraction.[15] All we need to be concerned with are the computational procedures that yield the right types of input/output patterns relative to the context in which the system is located, and with how these abstract procedures can be physically embodied. These are the factors that actually do the work, and they accomplish this using the normal theoretical resources of the natural sciences.[16]

Notes:

1. In a more general setting, Putnam (1988) calls this the 'hook-up' problem.

2. In the present context I construe a naturalistic account of intentionality and mental content to be a theory of these phenomena that does not itself depend on any primitive or ineliminable notions which are themselves intentional or mentalistic. Hence all such theoretical primitives are part of the basic framework of the natural sciences and are compatible with a materialist ontology, and with physical conservation laws applied

to the human body. Under a less stringent reading, which I do not adopt, anything which is 'entailed' or 'required' by the natural sciences is itself automatically naturalized.

In turn, there are a number of different ways to construe what it is to be a *theory* of intentionality and mental content. I shall interpret it broadly to mean a systematic method for specifying natural circumstances and configurations which at least reflect the same truth conditions as statements of purported fact involving mentalistic terminology.

3. In making this critique I am not endorsing the metaphysical or semantical presuppositions that often go along with a standard view of conceptual analysis and the intension/extension distinction. Instead I am simply using the traditionalist's own conceptual ammunition against their anticipated objection.

4. Along comparable polemic lines, the traditional religious conception of 'human being' holds that possession of an immaterial soul is an essential trait of the thing in question. But surely the scientific understanding of *Homo sapiens* need not be constrained by this traditional religious conception. Furthermore, the denial of this trait does not mean that the subject under investigation has thereby shifted and we are no longer talking about 'real' human beings. Instead, the preferred conclusion is that the traditional specification has no extension, and hence we must modify our *concept* to fall within the purview of the most productive and successful theories currently available.

5. According to the Church-Turing thesis, every computable function is computed by some Turing machine. And every Turing machine is expressible as a finite table of instructions for manipulating the symbols '0' and '1', where the 'meaning' of the manipulated symbols is entirely ignored. There is nothing internal to the machine that would indicate whether the syntactic transformations were computations of arithmetical

functions, tests for the grammaticality of linguistic expressions, proofs of theorems in first-order logic, answers to questions posed during a session of the Turing test, etc. The interpretation of the formal activities must be supplied from the outside, via an external set of conventions for interpreting input/output strings, halting configurations, initial conditions, etc.

6. Another example of the inability to recoup semantical value in formal terms is supplied by the well known problems associated with the attempt to derive intensional content from the full mathematical resources of higher-order modal logic. Even assuming the quite powerful machinery of type-theory, higher-order quantification and possible worlds semantics, Montague grammar and related programs in natural language semantics have been unable to define model-theoretic structures which possess the right general identity conditions to serve as the formal analogues of propositional content. This long standing failure to derive intensional content from logic and extensional set theory supplies one more piece of 'inductive' evidence to support a generalized version of Brentano's second premise.

7. 'Information based semantics' (a la Dretske, Barwise and Perry) supply yet another variation on the same basic problem. Even though the *term* 'information' occurs in various scientific contexts, such as computer science and electrical engineering, still the *concept* of information is just as unnaturalized as the philosophical notion of content. To the extent that information is mathematically well defined, it enjoys no semantical component, but rather is a purely quantitative affair dealt with in formal 'communication' theory (e.g. Shannon and Weaver). The intuitive *meaning* connoted by the term 'information' remains obscure, and tends to be either an unexplained mentalistic element

projected onto the syntax, whereby the intentional homunculus is never really discharged, or else is posited as something 'real' that exists independently of cognitive agents. This latter move constitutes a type of deviant Platonism, in which the Forms are no longer static and crystalline, but rather have taken on a liquid guise, and 'flow' through various abstract channels.

8. Searle's more fine grained version of the argument then appears to have the following structure, where the original formulation above is augmented with two further premises:

(1) the mind possesses intrinsic semantical content

(1.1) intrinsic semantical content is essentially linked to conscious presentations

(1.2) syntactical manipulations are not sufficient for conscious presentations

so,

(2') syntactical manipulations cannot capture intrinsic semantical content

therefore,

(3') the mind cannot be reduced to a mere syntactical system.

In this manner, premise (1.2) seems to transform the Chinese Room into a type of 'missing qualia' argument against computationalism, where the crucial load is no longer carried by metamathematical considerations, but rather by views on the origins of consciousness.

9. In itself this project already presupposes a vast simplification of the general philosophical issue, since it assumes that the intended interpretation is to be recovered from the fixed and pre-existing physical world, according to various aspects of its causal structure and history. And while this assumption is quite appropriate within scientific

confines, it is certainly not entailed by a purely *formal* approach to the mind. On the contrary, the closed character of syntactic systems makes them quite compatible with Descartes' solipsistic picture of the mind, wherein the very same thoughts and sensory impressions could occur in the absence of a material world to which they were related. According to Descartes there could be any number of different 'causal' circumstances correlated with the same mental event, and he therefore entertained a very radical version of the one-to-many problem, in which even a malignant demon could not be ruled out as a non-standard model.

10. In a similar vein, proponents of connectionism are wont to claim that the learning episodes whereby a network is trained provide a semantic grounding for their distributed processing states. But this won't work, for the same reasons that prevent 'swamp-man' type replicas with the same instantaneous structure but no past history from sharing wide mental states with their prototypes. Once a network has been trained and all the connection weights established, a duplicate system can then be constructed which will be indistinguishable from the 'grounded' network and will behave/operate in precisely the same manner, but which has had no prior interaction with its environment and hence no opportunity to ground its internal activities. In principle, connectionist computations are just as solipsistic as classical ones.

11. There is a certain degree of similarity between this view and Dennett's 'intentional stance'. Indeed, one could characterize the weak approach advocated herein as taking a 'representational stance', because it is operational and conventional rather than realist. However, it differs from Dennett's views on intentional systems in a number of ways. Perhaps most significantly, Dennett's position is *instrumentalist*: intentional

states are attributed on the basis of macroscopic behavior, and do not need to be underwritten by any knowledge of the inner workings of the system – they are justified merely if they yield an increase in predictive power. So Dennett's stance concerns how we get a theoretical handle on cognitive systems from the *outside*. In contrast, my position is about fine-grained interior details. Even if we knew *everything* about how processing was carried out and behavior produced, still nothing about this story would be enough to underwrite the traditional notion of 'aboutness' as a feature of internal processes and states. The representational stance is just a heuristic, a convenient way of abbreviating various systematic correlations between internal structures (ultimately described in microscopic detail) and external features of the environment.

12. In this way the traditional notion of intentionality is closely linked to ancient (and currently discredited) views on perception.

13. This is not to say that externalism offers a naturalistic account of public language semantics. Although it bypasses one of the major theoretical faults of the traditional approach, by taking meanings out of the head, still the notion of wide content presupposes various mentalistic and intentional primitives. A theory of public language semantics is an idealized, normative affair rather than straightforwardly naturalistic or descriptive.

14. Many connectionists buy into the mainstream view that representational content is essential to the mind, and seek ways of incorporating this idea into the architecture of artificial neural networks. A comparatively recent such attempt (Churchland, P.M. 1998, Laasko, A. and G. Cottrell 2000, O'Brien, G. and J. Opie 2001) uses cluster analysis to locate 'vehicles' of representational content. I would argue that

such attempts suffer from exactly the same built-in tension that afflicts the LoT model; namely, the purported content for which the clusters serve as vehicles does no work. Just as in the classical case, the postulation of content within the connectionist framework is gratuitous, because it plays no role in the cognitive manipulation of inputs to yield salient outputs.

15. The broad position expressed here makes no assumptions about what type of computational architecture is appropriate – this is a question for scientific research. The discussion has taken the classical model as a basic illustration, but the key points about processing structure/syntax versus content obviously generalize.

16. From this perspective, there is no cleanly detached mental level characterized by intrinsic 'aboutness', and there is no genuinely autonomous psychological domain. Instead, the applicability of the term 'cognitive system' becomes a matter of degree, determined by the nature and sophistication of its input/output capabilities and the richness of its internal processing structures. In turn, abstract processing structures are ultimately seen as a level of description for neurophysiological activities and events.

References:

Akins, K. (1996), 'Of Sensory Systems and the 'Aboutness' of Mental States', *The Journal of Philosophy* 93 (7), pp. 337-372.

Barwise, J. and J. Perry (1983), *Situations and Attitudes*. MIT Press.

Brentano, F. (1874), *Psychology from an Empirical Standpoint*.

Brooks, R. (1991), 'Intelligence without Representation', in J. Haugeland, ed., *Mind Design II*, MIT Press, pp. 395-420.

Churchland, P.M. (1998), 'Conceptual Similarity Across Sensory Diversity: The Fodor/Lepore Challenge Answered', *Journal of Philosophy* 95(1), pp. 5-32.

Dennett, D. (1987), *The Intentional Stance*, MIT Press.

Dowty, D., R. Wall and S. Peters (1981), *Introduction to Montague Semantics*. Reidel.

Dretske, F. (1981), *Knowledge and the Flow of Information*. MIT Press.

Dretske, F. (1995), *Naturalizing the Mind*. MIT Press.

Fodor, J. (1978), *The Language of Thought*. Thomas Y. Crowell.

Fodor, J. (1980) 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology', *Behavioral and Brain Sciences* 3, pp. 63-73.

Fodor, J. (1987), *Psychosemantics*, MIT Press.

Harnad, S. (1990), 'The Symbol Grounding Problem', *Physica D*, 42, pp. 346-355.

Harnad, S. (1991), 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1, pp. 43-54

Laasko, A. and G. Cottrell (2000), 'Content and Cluster Analysis: Assessing Representational Similarity in Neural Nets', *Philosophical Psychology* 13 (1), pp. 47-76.

Millikan, R. (1984), *Language, Thought and Other Biological Categories*. MIT Press.

O'Brien, G. and J. Opie (2001), 'Connectionist Vehicles, Structural Resemblance, and the Phenomenal Mind', *Communication and Cognition* 34, pp. 13-38.

Putnam, H. (1975) 'The Meaning of `Meaning'', in *Mind, Language and Reality*, Cambridge University Press, pp. 215-271.

Putnam, H. (1988), *Representation and Reality*. MIT Press.

Pylyshyn, Z. (1984), *Computation and Cognition*. MIT Press.

Searle, J. (1980), 'Minds, Brains and Programs', *Behavioral and Brain Sciences* 3, pp. 417-424, quoted in M. Boden, ed., *The Philosophy of Artificial Intelligence*, Oxford University Press, 1990.

Searle, J. (1990), 'Consciousness, Explanatory Inversion and Cognitive Science', *Behavioral and Brain Sciences*, 13, pp. 585-596.

Searle, J. (1992), *The Rediscovery of the Mind*. MIT Press.

Shannon, C. and W. Weaver (1949), *The Mathematical Theory of Communication*, University of Illinois Press.

Stich, S. (1999), 'Autonomous Psychology and the Belief-Desire Thesis', in W. Lycan, ed., *Mind and Cognition*, 2nd edition, Blackwell Publishers, pp. 259-270.

Van Gelder, T., (1996), 'Dynamics and Cognition', in J. Haugeland, ed., *Mind Design II*, MIT Press, pp. 421-450.