A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation

MARIUS USHER

Abstract: A naturalistic scheme of primitive conceptual representations is proposed using the statistical measure of *mutual information*. It is argued that a concept represents, not the class of objects that caused its tokening, but the class of objects that *is most likely to have caused it* (had it been tokened), as specified by the statistical measure of *mutual information*. This solves the problem of misrepresentation which plagues causal accounts, by taking the representation relation to be determined via ordinal relationships between conditional probabilities. The scheme can deal with statistical biases and does not rely on arbitrary criteria. Implications for the theory of meaning and semantic content are addressed.

1. Introduction

Primitive/atomistic conceptual representations (or concepts) are central in empirical sciences like psychology and neuroscience, and in philosophical theories that address the structure of cognition. In psychology, concepts are thought to provide a fast and flexible way to categorize an uncountable set of possible objects within the external environment into a finite repertoire of categories. In neuroscience, neural assemblies are often said to *represent* classes of objects (such as faces) in virtue of their neural responses. In philosophy, representations in general (and concepts in particular) are used as a basis for naturalistic theories of meaning (Fodor, 1984; 1990; 1998; Millikan, 1998).

In Fodor's representational theory of mind (RTM), for example, concepts are mental particulars that function as atomic symbols with semantic properties. According to RTM, the truth value of a complex mental representation (such as 'the white cat is on the mat') is determined by the satisfaction of its atomic symbols (via their semantic relation to external objects) and by applying syntactic rules of computation that preserve the truth value. To avoid circularity,

I wish to thank Tim Crane, Martin Eimer, Michael Herrmann, Ariel Kernberg, Ernst Niebur and Nick Zangwill for very helpful comments and discussions, Keren Feingold for a careful reading of the manuscript and Samuel Guttenplan for very helpful editorial comments. Special acknowledgments are due to Chris Eliasmith for sharing access to his recent and very relevant work and for stimulating and challenging discussions and to Jerry Fodor for his challenging criticism and inspiring writing.

Address for correspondence: School of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, UK.
Email: m.usher@bbk.ac.uk

Mind & Language, Vol. 16 No. 3 June 2001, pp. 311–334.

the semantic content of symbols needs to be determined in a naturalistic way without relying on the notion of computation (as computation was defined in terms of preserving the truth value of symbols; Fodor, 1998) or on other semantic terms. This consideration, as well as the shared and public aspect of concepts, motivated Fodor to develop a *referential atomistic* theory of concepts, instead of an *inferential* one, where concepts are defined with regards to their relationships to each other. Although the topic is under debate (e.g., Fodor and Lepore, 1992, 1993; Churchland, 1993; 1998), a referential theory of meaning may be in a better position to explain the shared aspect of concepts (and the possibility for communication) in light of the fact that the relationships between concepts vary widely from person to person.[1]

One should also note that accepting a referential theory of the content of concepts does not mean that one has to accept *all* the assumptions of Fodor's *Language of Thought* (Fodor, 1975). One may, for example, hold that 99% of thought processes follow associationist (Hume, 1739) or constraint satisfaction principles (Rumelhart and McClelland, 1986). It is enough that syntactic computational processes can be used to evaluate the content of composite representations on the basis of their constituent symbols. In fact, one can even conceive of a referential connectionist theory of representation that does not satisfy Fodor's assumption on the compositionality and systematicity of mental representation,[2] but where the content of representations could still be determined by causal relations between the units of representation and properties of the environment. Such a causal component has been recently acknowledged by Churchland (1998) as one ingredient for his scheme of semantic content.[3] A causal referential component in a theory of content, as eloquently advocated by Fodor, is therefore, likely to provide a robust basis for a variety of cognitive theories.

The success of this enterprise, however, rests on the ability to provide a naturalistic account of how concepts represent. Unfortunately, however, all the naturalistic theories of reference run into difficulties (Cummins, 1989; Fodor, 1990; Crane, 1995). The major problem, often labeled as the problem of *intentionality*, is that it is not clear what makes a mind/brain representation *refer* to an external object (e.g., Searle, 1984). During the last 25 years, a number of theories of representation have been developed that rely on causal and informational principles to define the nature of mental representations (Stampe, 1977; Dretske, 1981, 1983). While these theories are successful in accounting

[1] Fodor also argues that a theory of conceptual content based on relationships between concepts will necessarily have a holistic aspect (each concept depends on almost every other concept) in a way that makes the theory useless (Fodor and Lepore, 1992; Fodor, 1998).

[2] The representation theory proposed here does not negate compositionality and, moreover, is consistent with (but does not rely on) it.

[3] Churchland (1998) assumes that both the causal relations of concepts to features of the environment as well as their relations to other concepts determine their semantic content. It is not so clear however, what is the differential role of the two factors in determining content.

for many representational properties, they fall short of accounting for misrepresentations. Other theories rely on adaptational principles in order to overcome the misrepresentation problem (Millikan, 1993; Papineau, 1987), but their success (in accounting for *all* cases of misrepresentation) is strongly disputed (Cummins, 1989; Fodor, 1990, section 3; Crane, 1995; Neander, 1995). Today, the issue of misrepresentation with respect to causal-informational theories of content (and the related *disjunction* problem; see next section) is one of the most debated and problematic topics.

The aim of this paper is to propose a naturalistic theory of reference on the basis of statistical measures such as *mutual information*. The theory takes as its assumption that concepts are used in an inherently probabilistic process of perceptual categorization (as discussed in the next section), and therefore any theory of representation that can account for cases of misrepresentation needs to take this probabilistic factor seriously. Using tools from Shannon's theory of *transmission of information*, I will argue that a statistical, informational/causal scheme can, if not provide a definitive account for the representational content of elementary concepts, at least enable significant steps in this direction.

The paper is organized as follows. In the next section I briefly discuss some of the difficulties encountered in causal theories of representation, focusing on the problem of misrepresentation. In section 3, I review some of the challenges raised against statistical approaches of representation, and then I describe a statistical competitive scheme that should be able to meet these challenges. In section 4, this scheme is applied to typical situations involving mispresprentations. Finally, the implications of this scheme for a theory of meaning are discussed.

2. Causal Theories of Content and the Misrepresentation-Disjunction Problem.

Causal principles seem to be the best candidate for a theory of mental representation (Fodor, 1990; Crane, 1995). After all, the act of perception is a causal process and its function is to provide an individual with knowledge about objects in its environment that causally affect it. Moreover, a theory of representation based on causation can solve a number of problems that are difficult for non-causal approaches. For example, resemblance-based theories (which define the representation relation in terms of the similarity between the symbol and its reference) and description-based approaches (where concepts are defined in terms of a list of properties) have difficulties in accounting for the asymmetry and the singularity of representation (Stampe, 1977; Fodor, 1990; Cummins, 1989; Crane, 1995).

Stampe (1997), for example, characterizes the relation of representation between a representational symbol *R* and its representational object *S* by the requirement that *S* causes *R*, in such a way that properties of *S* are causally responsible for properties of *R*. This solves the problem of singularity (e.g., the concept 'this tiger' refers only to this particular tiger, and not just to any tiger, since only this tiger was causally responsible for its perception) as well as the asymmetry of representations (the concept 'tiger' represents tigers but not the opposite). Two main problems appear, however. First, according to this account one has difficulty in assigning content to tokens of a representation generated merely in thought (the thought about sheep is not always caused by sheep, as when, say, one tries to sleep). Second, it is clear that this scheme has difficulties accounting for the possibility of misrepresentation (Fodor, 1984; Cummins, 1989; Crane, 1995). If a symbol represents an object in virtue of being caused by it, then when an object is misrepresented (say when a dog is misperceived as a cat) the symbol is still linked to the cause (i.e., the dog). As discussed in detail by Fodor (1984), this generates two problems. Epistemologically, symbols are unintelligible unless one really knows what caused them (e.g., the content of the symbol 'cat' may correspond to a dog or to any other causal item (see discussion in Fodor, 1990, pp. 36–38)). Ontologically, according to Stampe's account the misrepresented dog-symbol represents its cause (i.e. the cat), leaving no room for misrepresentations.

A similar theory of representation was developed by Dretske (1981; 1983). Unlike Stampe's, however, Dretske's theory (as well as Fodor, 1990) is formulated in terms of counterfactually supporting causation (and therefore the actual history is irrelevant). In Dretske's terms, a symbol represents an item when it *carries information* about it. The use of information to determine the content of representation is essential for the approach I pursue here, which follows the framework first outlined by Dretske. There is, however, one essential difference. Although he defines representational content with regards to the information that a symbol carries about its cause, Dretske (1981; 1983) imposes a further restriction upon the requirement for representation. He demands that the conditional probability for the represented item, given the symbol, is unity. In other words, only certain information is accepted as the basis for a representational relation. This restriction, motivated by Dretske's concern with preserving the truth value of the representational relation on composite symbols and with preserving the transitivity of representation (but see criticism in Fodor (1990, pp. 57–58) and section 4.1 below), reduces, however, the statistical-informational measure to a logical one. As discussed in detail by Fodor, Dretske formulates the condition for *R* to represent *S*, by the requirement that the generalization *S and only S causes R is counterfactually supported* [or *S* (*and only S*) *causes R is a law*; Fodor, 1990, p. 57]. This makes Dretske's representation theory (Dretske, 1981; 1983) vulnerable to the same misrepresentation problem discussed above for Stampe's theory (Fodor, 1990; Cummins, 1989; Crane, 1995). A related and important way to describe the difficulties of these causal theories in accounting for representation is the disjunction problem. Since this is considered a major problem for causal theories of representation that I address here, I will provide a brief description, following closely Fodor's review (Fodor, 1990, ch. 3).

For misrepresentation to occur, the nomic relation between *R* and *S* must be imperfect (where *R* is the symbol that represents *S*). The idea for both Stampe and Dretske is to assume that while, for faithful representations, tokens of *R* are caused by tokens of *S*, for incorrect representations there are 'wild' *R* tokens that are caused in some other way (say by *T*). If that is the case, however, it would seem that what *R* represents is not only *S*, but rather the *disjunction* ($S \vee T$), which has a higher correlation with *R* than *S* alone has (Fodor, 1990, p. 40). As discussed by Fodor (1984), this problem is even more acute in Dretske's theory where the representation relation supports counterfactuals. Even if *T* never occurred so far, the mere possibility of its occurrence causing *R* is enough to include *T* within the content of *R* (Fodor, 1984; see also Dretske, 1983b, p. 17).

Most attempts to solve this problem within strictly causal theories have tried to distinguish between two types of situations, where type I are the ones which confer content, while type II do not. Dretske (1981; 1983), for example, relied on the distinction between learning (where conditions are assumed to be optimal) and retrieval phases (where they are not). This solution was strongly criticized by most representation theorists. First, there is no non-circular way to distinguish learning situations from other situations, as organisms appear to be able to learn to identify things without ever reaching perfection (Cummins, 1989, p. 68; Fodor, 1984; 1990). Second, because of the counterfactual-supporting property of Dretske's condition, it is not possible to rule out the causes of 'wild' tokens of a representation from its content domain, merely because they did not happen to occur during the learning period (Fodor, 1984).

Other approaches have tried to distinguish between type I and type II situations with regard to the notion of 'normalcy'. The challenge, though, is to provide a naturalistic account for 'normalcy' that does not rely on semantic features (see Fodor, 1990, ch. 2 and 3). In light of these difficulties, Fodor (1990) concluded that a naturalistic theory of representation (and referential content) that provides an account of misrepresentation without relying on a distinction between the two types of situations is needed. In the next section, I try to show how this can be achieved for a specific type of representation—conceptual (atomistic) representations—by returning to Dretske's (1981) original idea for determining the content of a representation with regards to the information it conveys about its causal factors (but without the restriction on conditional probability). This will require, however, a statistical framework (Oaksford and Chater, 1998) rather than a logic-based one.

3. A Statistical Scheme for Representational Content

Causal theories of representation are attractive because they make use of perception and categorization processes where mental representations are tokened. The problem is that in order to allow room for misrepresentation it needs to be the case that 'the conditions for the *truth* of a symbol dissociate from the conditions whose satisfaction determine what the symbol represents' (Fodor, 1990, p. 42). This is, however, what the causal theories cannot do: since causation is used to determine representational content it cannot also be used to dissociate content from truth.

The idea proposed here is that while causation is important for representation it provides too strong a condition. Note that the introduction of counterfactuals in the theories of Dretske (1981; 1983) and Fodor (1990) already provides a weaker condition, as it allows *R* to represent *S* even when *S* never happened. Dretske's additional requirement that the conditional probabilities are unity, however, nullify the full effects of this modification; when *R* is eventually tokened, then *S* needs to have caused its tokening (thus one still cannot have tokening of concepts 'offline' in thought). Fodor (1990) eliminated (and criticized) this conditional probability restriction within his causal theory of content, without, however, making use of the probabilistic framework for addressing the problem of misrepresentation. I will comment on Fodor's solution to this problem in the discussion section. The work presented here shall make full use of the probabilistic framework that provides statistical measures to define the reference of representations. The rationale behind the approach is that, since organisms perceive the world via a causal but probabilistic process, the causal object of a perceptual categorization can never be known with certainty. The rational strategy in this situation is to *estimate the most likely* object that could have caused the perceptual state (Bialek et al., 1991). The role of mental representations may therefore be to provide, not faithful access to the class of objects causing an act of perception, but rather a statistical inference (or hypothesis) of what type of object *could be* causally involved (see also Oaksford and Chater, 1998, for an introduction to a statistical framework to cognition and rationality that replaces the traditional logic-based one). In other words, when a concept is tokened, what is represented is not the type that caused the mental state but the type that is the *most likely* to have caused it. This is consistent with Dretske's original idea that mental symbols represent what they carry information about. A statistical measure for 'carrying information' is provided by Shannon's theory of the *transmission of information* (see e.g., Fano, 1961).

Before I describe the scheme for representational content based on this theory three preliminary issues need to be addressed. First I review the challenges that a statistical framework for representation needs to face, and which have often been used to dismiss it. Second, prior assumptions regarding the nature of concepts as tools for a categorization map of the world, are addressed. Third, two factors that make the categorization map intrinsically probabilistic are discussed.

3.1 Challenges Facing a Statistical Scheme for Representation

A number of arguments have been put forward to demonstrate that statistical measures cannot offer a valid account of representation. One such argument was already mentioned. Accordingly, if the representation relation is to be explained by a correlation/frequency measure, one is faced with the fact that a representation symbol covaries more often (and therefore more reliably) with a class of disjunctive stimuli that includes its reference than with the class corresponding to its reference alone (Fodor, 1990; Hutto, 1999). A similar argument involves the observation that, when the expectation for (or the salience and importance of) an item is high, the representation frequency may be high while the occurrence frequency of the represented item is low. Millikan (1989) provides the example that, although the representation of danger is very often tokened (for a vulnerable animal), the actual occurrence of a predator in its immediate environment may be quite rare (so that given the tokening of danger representation it is more likely that there is no predator than that there is one).

A second argument is related to the concern that an appeal to statistical measures will result in arbitrary criteria: 'are average conditions those which obtain in at least 50% of the occasions, or is it 90%? . . . But the notion of semantic content is surely not relative' (Millikan, 1989). A similar argument (the 'redox' principle) has been put forward by Dretske (1981, 1983) to support the claim that informational content cannot rely on uncertain messages that are likely to require arbitrary criteria for their disambiguation. More generally, it is said that there is no way in which statistical measures can set up the 'standard' for correct representation. To anticipate, I will try to show that in fact a statistical approach can meet all these challenges.

3.2 Conceptual Maps

Since I focus on primitive conceptual representations, I will start by defining the terms and assumptions underlying my scheme. These definitions follow quite closely recent work by Fodor (1998) and by Millikan (1998) on the nature of concepts. I take concepts to correspond to categories (Fodor, 1998) that are used to refer to objects in the external environment. In this sense concepts provide a many-to-one map of the world (many world-items map onto the same concept). Millikan (1998) uses the Aristotelian term 'substances' to characterize the type of items that concepts refer to. The examples she examines, and which I also focus on, include what Millikan calls real kinds (cat, chair), individuals (Mama, Bill Clinton) and stuffs or ordinary substances (milk, gold). Moreover, Millikan (1998) argues that concepts are to be individuated by the capacity to identify exemplars rather than by a description of their properties (i.e., the concept 'cat' is individuated by the ability to tell cats from non-cats, rather than by being able to list properties such as fur, meowing, etc.). These properties are, according to Millikan, secondary to the referential character of concepts. They are acquired later during human development (infants acquire linguistic representation of substances, such as 'animal' and 'food' much earlier than they acquire linguistic terms for their properties), and the ability to identify a substance is needed in order to confer to it a set of properties.[4] This approach is also consistent with Fodor's (1998) referential atomism.

Here I adopt Millikan's characterization of concepts, but I label their referents as 'objects' (instead of 'substances') to avoid the possible confusion with ordinary substances (or stuffs, such as 'gold').[5] Two properties of this approach are particularly important for the theory of representation presented here. First, concepts are symbols with conditions of satisfaction within acts of perception. For example, the concept 'cat' is satisfied when tokened in perception if it is caused by seeing a cat. This allows one to address the simplest and most widely discussed cases of misrepresentation, where a concept is tokened during an act of perception in response to an object that corresponds to another concept (e.g., when perceiving a small dog in a dark night as a cat; Fodor, 1990; Crane, 1995). A second important feature of this scheme is that the representational map is limited to apply only to objects as defined above. It thus excludes *events* such as 'a glimpse of a dog in the dark'. Such events are situation dependent and do not satisfy the requirement of being an object (or substance). To put it differently, the use of concepts relies on a prior ontological assumption that the world can be individuated with regards to objects (and not just with regards to events).[6] Accordingly, the content of a concept is a set of objects and not a set of events (or situational properties).

3.3 Probabilistic Factors in Perception

Two factors make the categorization map between objects in the world and conceptual states, probabilistic. The first factor is related to intrinsically noisy information processing in the central nervous system. Both human performance and single neurons show highly variable (probabilistic) behavior even under identical external stimulation.[7] For example, in the neuroscience literature, the issue of variability in the discharge patterns of nerve cells has recently been the focus of intensive research (Softky and Koch, 1993; Shadlen and Newsome 1994; Usher et al., 1994). This aspect of information processing is

[4] This does not mean that the identification process is insensitive to a set of sensory properties that the perceptual object generates. Such sensory properties, however, need not be part of the conceptual repertoire of the user, and may not play a role in lexically based definitions.

[5] No perfect term for Aristotelian-substances exists in common language. The term objects, as used here, should be taken in a broad way to apply also to real kinds, to individuals and to ordinary substances, like 'gold'. This term is preferred here for being more accessible to non-philosophy cognitive scientists.

[6] Likely reasons for human and animals to operate upon this assumption in their formation of conceptual maps, are adaptationally evolved perceptual mechanisms such as *object constancy* and *Gestalt* principles, which divide the world in terms of whole objects (dogs but not dogs-at-night). Exceptions to this such as, evening-star/morning-star, exist, however they take place only when object-binding fails.

[7] This is clear for brief or degraded stimuli but is also present under the most optimal sensory conditions as quantified by the variability in behavioral measures such as response latencies.

consistent with the computational framework of Parallel Distributed Processes (PDP, Rumelhart, McClelland, and the PDP Research Group, 1986, ch. 5–7; McClelland, 1991; Movellan and McClelland, 1995; Usher and McClelland, 2001) and Attractor Neural Networks (ANN, Hopfield, 1982; Hinton and Sejnowski, 1983, 1986; Amit, 1989). In particular, McClelland (1991) and Movellan and McClelland (1995) have shown that *only with intrinsic variability* incorporated can a classical PDP model (the interactive activation model) account for empirical data patterns involving effects of context on perceptual categorization (Massaro, 1989). The implication of this intrinsic noise factor is that, even in *identical* stimulus conditions, the perceptual process of categorization is probabilistic.

The second factor is conceptually more interesting. The categorization process, as presented above, involves a (many to one) map between objects and concepts. Objects, however, are not directly presented in perception. They are projected onto the senses via perceptual stimuli. This projection process depends on specific situations (light, distance, occlusion, modality, etc.). Since such situations are orthogonal to the map (between objects and concepts) they introduce an additional random factor that the perceptual system needs to take into account when selecting the most likely *object* responsible for a perceptual representation.

3.4 A Statistical Competitive Scheme based on Mutual Information

As discussed above, a statistical measure for content based on a measure of correlation is not enough. First, such a measure would be biased by the occurrence frequency of objects (Millikan, 1989). Second, if the categorization process is truly probabilistic, then any object in the world may causally token a symbol with some low probability. A simple correlation measure would then imply that the symbol has as its content the whole world; that would trivially obtain the maximal correlation with the symbol.

Fortunately, Shannon's theory of *transmission of information* (see, e.g., Fano, 1961) provides more refined statistical measures to characterize the information a symbol in a receiving system carries about states of the environment (here denoting objects that affect perception). The relevant measure of this theory for a theory of representation is the *mutual information* (*MI*) between a symbol (or representation), R_i, and objects or states of the environment, S_j (the indexes i,j enumerate various representation states and objects). The mutual information can be simply expressed in terms of conditional probabilities between objects and representations:

$$MI_{ij} = \log \frac{P(R_i|S_i)}{P(R_i)} = \log \frac{P(R_i,\ S_j)}{P(R_i)\ P(S_j)} = \log \frac{P(S_j|R_i)}{P(S_j)} \tag{1}$$

This identity is based on Bayes' law of probability and shows that the mutual information can be computed either from the matrix of joint probabilities (middle term) or from the matrices of conditional probabilities (left and right terms; the left term corresponds to the probability conditioned on objects, S_i, and normalized with the frequency of symbols, R_i, while the right term corresponds to the converse).[8] Since in what follows we only make use of ordinal relations in *MI*, and since the logarithm is a monotonic function, we can eliminate it (or more formally, rely on exp(*MI*) which provides the same expression but without the logarithm).

The measure of mutual information alone, however, is not enough to determine the content of representational symbols. After all, any symbol carries some amount of information about any object (unless there is absolutely zero correlation). What is needed is some procedure to determine which one of the objects in the environment the symbol represents, given the amount of information it carries about them. Two complimentary approaches are outlined below, differing in the way they determine the representational map: from the world to representations—an externally based approach; or from representation to the world—an internally based approach, but alike in making use of a comparative type of scheme.

External-based scheme. According to this approach, the classes of objects in the world (e.g., cats, dogs, gold, etc) are considered to be given as part of the structure of the world.[9] The aim of the scheme is to determine the representation states that correspond to these classes of objects and to account for the representational relation. Basically, this is similar to the pragmatic way in which neuroscientists proceed to determine the neural representation an animal has for objects within a category, say cats. Schematically, the animal is presented with a representative sample of cats and one determines the neural structure that shows the best response (in terms of activation) *in average over all the sample*, and *relative* to responses it generates for items of other classes (not cats). (For a philosophically friendly presentation and criticism of this approach, see Eliasmith, 2000). This can be formalized in term of *MI*:

C1. R_i is a representation for the class S_i, (only) when S_i, *can cause* R_i *and the mutual information between* R_i *and* S_i *is larger than the mutual information between* R_i *and objects of any* other *types,* S_j, *in average over exemplars and situations.*

Using the second term of Eq. 1, this can formulated as:

[8] The formula can be understood following the intuition that no mutual information exists when the two systems are independent (i.e., uncorrelated, and thus $P(R_i, S_j) = P(R_i)P(S_j)$, resulting in $\log \left(\frac{P(R_i, S_j)}{P(R_i)\ P(S_j)}\right) = 0$, and that MI_{ij} should increase with the deviation from independence (measured by $\frac{P(R_i, S_j)}{P(R_i)\ P(S_j)}$).

[9] I.e., there is a set of properties that can, in principle, be formulated in scientific terms and that characterizes these classes. I do not address here use-dependent objects, such as *doorknobs* (Fodor, 1998), which are likely to be culture/use dependent.

$$MI(R_i,\ S_i) = \frac{P(R_i|S_i)}{P(R_i)} > \frac{P(R_i|S_j)}{P(R_i)} = MI(R_i,\ S_j);\ \textit{for all}\ j \neq i \tag{2}$$

which, because of the identical denominator, can be reduced to a simple expression that involves *forward* conditional probabilities (i.e., conditioned on objects):

$$P(R_i|S_i) > P(R_i|S_j);\ \textit{for all}\ j \neq i \tag{3}$$

The definition **C1** (Eq. 3) provides a necessary condition for R_i to be a representation of the class S_i. For example, for symbol R to represent cats, it is necessary that $P(R|cats) > P(R|dogs$, etc). Thus R represents cats if R is 'cats' but not if R is 'dogs', 'animals' or 'gold', etc. As defined above, however, **C1** does not provide a sufficient condition for R to represent 'cats'; subclasses of 'cat' (e.g., 'siamese') will also satisfy Eq. 3 [$P(\text{'}siamese\text{'}|cats) > P(\text{'}siamese\text{'}|dogs, etc)$]. To provide a sufficient condition, **C1** can be strengthened to:

C1*. *The difference (or contrast) between the mutual information that R_i carries about S_i and the information it carries on* other *type of objects, is higher than for other representation states that satisfy* **C1**.

Clearly, $P(\text{'}cat\text{'}/cats) > P(\text{'}siamese\text{'}/cats)$ (when computed over a representative sample of cats) and therefore 'cat' but not 'siamese' represents cats.[10] Definition **C1** (Eq. 3) can be interpreted to imply that a necessary condition for the symbol R to have S as its content is that R has a higher probability to be tokened by exemplars of the category S (in average over exemplars and situations) than by exemplars of other categories. Interestingly, Eq. 3 does *not* restrict the relation between the forward conditional probabilities comparing different concepts (R_i, R_j) for a specific category of objects, S_i; it is still possible (but not necessary) that:

$$P(R_i|S_i) < P(R_j|S_i);\ j \neq i. \tag{4}$$

The example below shows the matrix of *forward* conditional probabilities, $P(R_i|S_j)$, for a 2-case categorization, corresponding to Millikan's case of the rare high-significance predator (see Appendix).

	R1	R2
S1	.8	.2
S2	.6	.4

[10] In neuroscience terms, a necessary and sufficient condition for R to be a representation of the class S can be obtained by averaging the neural activity patterns of responses, for samples of objects of type S. It is unlikely that two different classes (or a class and a non-representative subclass) will result in the same pattern. In the case of subsets (cats/siamese) one possibility is that the broader concept will pick a neural representation consisting of a subset of neurons that participate in the representation of the specific concept.

In this case, Eq. 3 is satisfied (because when comparing over objects, i.e., over columns of the matrix, the diagonal elements have the highest values), however, Eq. 4 is not (comparing over representations, i.e., over rows, the diagonal elements do not always have the highest values [.6 > .4]). This does not prevent R_2 from representing S_2 since what matters according to **C1** is Eq. 3 and not Eq. 4.

Internal-based scheme. This approach is closer to the task an individual (human or animal) faces and takes into consideration the limited resources of the individual (the animal does not have fully representative samples of objects and their objective probabilities at its disposal). To the animal what is given is the representation tokened, and the class of objects that may have caused it needs to be estimated (see also Eliasmith, 2000, for a detailed discussion of this approach and its advantages). This can be formulated as follows:

C2. For a given representation state, R_i, its content corresponds to *the set of objects, S_i, which can cause R_i and for which the mutual information between S_i and R_i is larger than the mutual information between S_i and any other representational state, R_j* (where $j \neq i$ and the relations are estimated when objects are presented under all situations).

In order to find whether S_i belongs to the representational extent of R_i, one needs to compare the mutual information matrix element M_{ii} to all the other elements M_{ji} (keeping the second index of the matrix, i which denotes objects, fixed). Using the right term in Equation 1 (where the denominator $P(S_i)$ is the same for all elements compared, and the logarithm can be ignored) this can be simply formulated (using Eq. 1) in terms of *backward* conditional probabilities, as:

$$P(S_i|R_i) > P(S_i|R_j), \textit{ for all } j. \tag{5}$$

This relation involves *backward* conditional probabilities of external objects given representation states (see Eliasmith (2000), for an illustration of the difference between this type of conditional probability and the traditional *forward* type [where the objects are given]). A similar condition can be formulated (using the left term in Equation 1) in terms of the *forward* conditional probabilities (of representations given external states), but requires a normalization by the representation frequencies:

$$\frac{P(R_i|S_i)}{P(R_i)} > \frac{P(R_j|S_i)}{P(R_j)}, \textit{ for all } j. \tag{6}$$

Only in the special case where the probabilities of representations are uniform, i.e. $P(R_i) = P(R_i)$, Equation 6 simplifies to Eq. 4, which as we saw, does not need to be satisfied for a representational relation to hold (when the probabilities of symbols are not uniform).

To summarize, the question an individual faces is estimating which of the concepts in its repertoire best matches an object in term of mutual information. This more individualistic (internal-based) approach is different from the more normative (external based) approach, where the competitive process of selection takes place among samples of objects for each concept. The difference between these two perspectives to representation have been discussed in detail by Eliasmith (2000), who labeled them (in the context of neuroscience) as the *animal* vs *observer* perspectives.

An additional feature of the internal-based approach is that it provides an account for the *best estimate* (or best-exemplar/prototype) of a concept, which is a central part of psychological phenomena. This can be done as in Eq. 3, i.e. $P(R_i|S_i) > P(R_i|S_j);\ j \neq i$, where S_i is interpreted, not as a representative sample of exemplars for the concept (that were given in the external-based approach), but as the best estimate of an object (corresponding to a prototype), which given the tokening of the concept R_i, the individual can consider most likely to have tokened it, relative to all other objects of that category or of other categories.

An important issue to consider is whether the normative external-based approach is consistent with the internal-based approach, which is at the individual's disposal. We saw that the external-based approach does not require the satisfaction of Eq. 4 (i.e., that in comparing among representation states for a given object, the conditional probabilities do not need to be in favor of the diagonal element). It can be demonstrated, however, that even when this happens, the inclusion of the denominator (as in Eq. 6) restores the dominance of the diagonal element as required for the internal-based approach (this resolves Millikan's challenge, as discussed in the next section and in the Appendix).

Finally, one additional qualification needs to be made to the procedure **C2**. As presented so far, conceptual content is determined via a competition process between all concepts for a perceptual object in terms of MI. This procedure may face a problem with concepts that form sub/super-ordinate hierarchy (e.g., animal/dog/poodle). As defined in **C2**, if an object has higher MI with the poodle concept, it will be excluded from the content domain of the superordinate concept 'dog'.[11] Clearly, however, this happens because of the excessive competition (between concepts) in the scheme which does not correctly reflect the competitive process that takes place within the cognitive system (where a stimulus can be categorized as 'dog' and 'animal' simultaneously). To correct for this, condition **C2** can be modified so that the competition is restricted: when determining the content of concept A, the competition is limited to concepts that are not subordinate to A and do not compete with it.

[11] I wish to thank to Chris Eliasmith for bringing this to my attention.

4. Using the Scheme

Here I will try to show how this competitive statistical scheme can be used to account for representational content in simple situations that are considered problematic for causal theories. In particular I will show that it can meet the challenges raised against statistical theories (e.g., of not relying on arbitrary criteria and of being able to cope with expectation and frequency bias effects) providing an account for paradigmatic cases of misrepresentation.

4.1 No Arbitrary Criteria

I examine here an important argument (Dretske's *redox* principle) previously used to highlight the problems of relying on arbitrary criteria of probability in informational theories of representation, and I will try to show that the present scheme can stand its challenge.

The *redox* principle (Dretske, 1981, 1983) is supposed to show that the transitivity of informational content (A has the content that B, and B has the content that C, implies that A has the content that C) is inconsistent with conditional probabilities between symbols and contents of any *arbitrary* value smaller than unity (this argument was one of the major justifications Dretske (1983) brought to support the restriction of conditional probability.[12]) Transitivity would indeed be violated if one defines the condition for A to represent B by the requirement that the latter causes the former with a conditional probability larger than some arbitrary threshold (say, .9). (This follows from simple laws of probability of independent events, $P(A|C) = P(A|B)P(B|C)$; if both $P(A|B)$ and $P(B|C)$ are higher than the threshold, say .91, $P(A|C)$ is lower than the threshold).

The reason why the scheme presented here (e.g., **C2**), does not suffer from this problem (thus satisfying transitivity) is that it does not rely on an arbitrary criterion value to determine content of representation, but rather on a *comparative* measure that involves *ordinal* relations between values of conditional probabilities. To illustrate this in a day-to-day setup, consider the situation where someone is told by another person that she detected the content A in her observation of an event (for example that person-A took part in a bank robbery). What this implies is a causal chain of events, going from the object A itself (the person recognized), to its representation in the first observer, A' (the belief that person-A was involved), and ending with the representation in the second observer A''. According to the present approach, A' has the content of A even when the conditional probability for A' to have been caused by A, $P(A|A')$, is lower than one, if this probability is higher than the probability for *any other* representational symbol, B' (say corresponding to the belief that person-B was involved), to have been caused by A, $P(A|A') > P(A|B')$

[12] See, also, Fodor, 1990, p. 58, for a criticism of the compositionality argument.

(see Eq. 5. This corresponds to the requirement that the probability for the A-belief to have been caused by the A-person is larger than the probability for the B-belief to have been caused by the same A-person, which means that the A-belief carries more information on A than any other belief, such as the B-belief). Transitivity of informational content would therefore imply that the same *ordinal* relationship between the relative probabilities is preserved when the A', B' representations are replaced with the A'', B'' representations. This can be the case despite the fact that all the conditional probabilities decreased. Therefore, receiving a noisy message of another noisy message may still provide informational content, as intelligence agencies definitely know. Moreover, noisy messages are the rule in the perceptual life of animals and humans who need to make the best of it.

4.2 Misrepresentations and the Disjunction Problem

There are a number of situations where misrepresentations happen. Most typically discussed is a situation where, because of degraded sensory input, an object is misrepresented for another one (say, a small dog at night is misrepresented as a cat) (Cummins, 1989; Hutto, 1999; Fodor, 1990; Crane, 1995). In the strictly causal theories, such a situation would have forced one to accept the unreasonable statement that the 'cat' token represents, in fact, a dog. This is definitely not the case here. According to the normative external-based approach, **C1**, the symbol 'cat' represents the category of objects that (via a representative sample) best tokens it (in terms of mutual information) relative to (representative samples of) objects in other categories. It does not matter that in a specific case, the symbol was tokened by a dog; what matters is only the general regularity. The disjunction problem does not arise also, as the content of the representational symbol is *not* the object that caused it, or the 'set that has the maximal correlation with it', but is limited to items belonging to the *class* whose exemplars obtain the highest *MI* to the representation symbol. In fact the symbol 'cat' (or its neural substrate), according to this approach, was picked by searching for the best response to cats only (on average over exemplars and conditions). Notice that, according to the assumption implicit to the categorization map, what is represented are objects only and not objects-under-a-situation. A dog-at-night, is still a dog and in order to find the concept that represents it, the same dog needs to be presented under all conditions (night, day, etc.). Similarly, according to **C2**, in average under all conditions, the small dog is more likely to be tokened as a dog, rather than as a cat (especially when looked upon in daylight and when barking). Therefore, it belongs to the content of the symbol where most dogs are mapped to, i.e., to the content of 'dog'. It having caused the tokening of the 'cat' symbol is a case of misrepresentation.[13]

[13] In his comments on a previous version of this paper, Jerry Fodor formulated the following objection to my solution of the misrepresentation-disjunction problem. Assume that the

A more problematic situation may arise, however, if the (mis)perceived dog is really perverse; under all possible scrutiny and under the best sensory situations, it somehow seems more 'catish' than most cats (maybe it was surgically modified to appear as a cat: it even mews). In this situation, indeed, the representational scheme **C2** will assign it to the content of the 'cat' representation and the misrepresentation is unexplained. However, such a misrepresentation is, by assumption, beyond the power of the agent to detect it and therefore it does not pose the same kind of problem. Indeed, according to **C2**, the content of the 'cat' representation contains the disjunction of 'cat' and all other items that are *absolutely indistinguishable* from cats. A similar disjunction is also accepted by Fodor in his reply to Baker's problem in his 'theory of content' (Fodor, 1990, pp. 103–104; see also Eliasmith (2000), p. 80 for a discussion of *allowable* disjunctions).

4.3 Expectation and Frequency Bias Effects

Let us now address, specifically, the example raised by Millikan (1989), where, for a vulnerable animal, the representation frequency of 'danger' (e.g., in relation to a predator) may be high, despite the fact that the occurrence frequency of that predator may be very low. To illustrate this, assume that there are two types of items in the animal's environment: deer (not dangerous) and tigers (dangerous). Assume also that tigers are very rare relative to deer (in a fraction of 1/9). Since *not-missing* a tiger is of crucial importance to the animal (much more than mistaking a deer for a tiger), the animal is biased in its categorization in favor of tigers: it has a high chance (.6) of mistaking a deer for a tiger and much lower chance (.2) of doing the opposite. [P(*'deer'*|*tigers*) = .2, while P(*'tiger'*|*deer*) = .6; i.e., whether a

representation 'cat' is (always) caused by a dog in condition *C* (say, dog in dark night or dog at exactly this place and time, etc). Since *C* is, by assumption, sufficient for a dog to cause 'cat', the probability that a dog in *C* will cause 'cat' is 1. Thus it would follow that 'cat' means *cat or dog-in-C*. This objection can be addressed at two levels. First, according to the probabilistic scheme presented, it is not clear that any *C* exists that gets the stipulated probability of 1 (for a dog to token 'cat'.) This is surely the case for degraded sensory situations (such as dark nights), where noise in the neural information processing (which may be ultimately dependent on quantum fluctuations) will make the process, essentially probabilistic. (Notice that the inclusion of timing in the definition of C is inconsistent with the idea of probability, as it makes C unique and unrepeatable). Another candidate for dog-in-C (tokening cats with probability 1) could be proposed to be something like: 'dog when I happen to think of a cat'. This is, however, a subject dependent property and not an objective world property, as needed for a naturalistic theory of concepts; according to the statistical theory presented here, one needs to average out over situational contexts (e.g., thinking of cats or of dogs prior to the act of perception and categorization). Second, a constraint imposed in this theory is that concepts map objects (or substances) and not events or situations. This constraint, satisfied by adaptationally evolved perceptual mechanisms such as *object constancy*, requires that only *whole* items (dogs but not dogs-at-night) are used to define conceptual content via their *MI* with representational symbols. Accordingly, dogs do not belong to the content of 'cat' representations since they have a higher *MI* with 'dog' than with 'cat' or any other concept that satisfies the constraints described above.

tiger or a deer are present, the animal is more likely to categorize it as 'tiger' (than as 'deer').] The matrix of conditional probability for this case was shown in section 3.4 (see also the Appendix).

The three matrices corresponding to the joint probabilities, $P(R_i, S_j)$, the *forward* conditional probabilities, $P(R_i|S_j)$ (shown in section 3.4), and the *backward* conditional probabilities, $P(S_i)|R_j)$, which can be formulated in terms of nromalized *forward* conditional probabilities, $P(R_i|S_j)P(R_i)$, are shown in the Appendix. The way in which the scheme **C1** accounts for representations in this case was addressed in section 3. Consider here the account obtained according to **C2**. The relevant measure (Eq. 6) is the normalized matrix (see Appendix). It is shown there that within each row of this matrix, the largest value is the one on the diagonal (unlike in the unnormalized matrix). Therefore, despite the bias in expectation, the mutual information scheme, **C2**, provides an account for the fact that 'deer' represents deer, despite the fact that deer are more likely to token 'tiger'.

This is a general result that holds in any situation where the probability of each representation given the correct object $P(R_i, S_j)$ is higher than the probability of that representation given an incorrect object ($P(R_i, S_j)$ (see Eliasmith, 2000, Appendix). Thus if the probability for deer to token 'deer' is larger than the probability of tigers to token 'deer' (.4 vs .2 in the example above) and if the probability for tigers to token 'tiger' is higher than the probability of deer to token 'tiger' (.8 vs .6 in the example), this is enough to confer on the two symbols the ability to represent the corresponding items. While these two requirements correspond to the psychological generalization that the animal has the ability to discriminate the two items, the relation between probability for deer to token 'deer' and its probability to token 'tiger' depends on decision biases and is irrelevant for the representational status of the symbols. Notice also that the frequency of objects in each class, $P(S_j)$ does not figure at all in this calculation and therefore it doesn't matter that there are more deer than tigers. The representational relations depend on counterfactuals as reflected by the conditional probabilities and therefore the actual frequency of objects in the external world does not matter.

4.4 Generalization to Socially Shared Concepts

The account for conceptual content offered under **C2** is mainly intended to cover the way an individual makes use of concepts to interpret likely causes of perceptual experience. This approach, however, has a straightforward generalization to social groups, where the same principle applies. The content of a concept shared by the group can be determined, to refer to objects in the world that have the highest mutual information with the tokening of the concept *in the group* relative to all other concepts (Eqs. 5, 6). The only difference is that here the tokening of a concept involves a group agreement[14] (see also Eliasmith, 2000, p. 79).

[14] This can be done either by a majority rule or, more effectively, by nominating experts.

5. Discussion

A fundamental issue for the theory of meaning and semantic content, which triggered a strong debate in cognitive science, is: 'What is the content of a symbol and what accounts for semantic relations?' In the absence of a definitive answer the program of the Cognitive Science threatens to collapse into a theory of manipulation of ungrounded, meaningless symbols (Harnad, 1990; Searle, 1980, 1984). The field is divided between referential theories where the meaning of a symbol is determined by its causal relations with items in the world (i.e., reference; Fodor, 1987, 1990, 1998; Dretske, 1981, 1983; Millikan, 1998) and inferential theories where meaning is defined mainly with regards to conceptual interrelations (Harman, 1982). Recently, a new version of an inferential theory for meaning was proposed by Paul Churchland (1993) on the basis of connectionist network theory. In this approach, semantic relations are determined by the similarity among neural activity patterns in a semantic space.[15] In a more recent work, however, Churchland included causal relations between representations and 'stable and objective macrofeatures of the external environment' as a component (a referential one) in his theory of meaning (Churchland, 1998, p. 8). The approach suggested here may provide a way to bring these two frameworks closer together.

The statistical scheme for conceptual content proposed here is close to Fodor's in a number of ways. First, as in Fodor's theory, concepts are mental particulars which provide a categorization of the external environment. Second, the content of concepts is provided by a causal referential relation to things in the world (and not by inter-conceptual relations). Third, as in Fodor's theory, *meaning is not identical with causation*. For Fodor (1990, pp. 90–91) this involves a dissociation between meaning and information. Since I rely on a statistical theory of information this second separation is not needed, because information is interpreted according to the contrastive/competitive procedure discussed earlier. When a concept is tokened, the information it conveys is about the class of items it carries the most information about, and not about what caused it in a singular case.

In particular, the external-based scheme **C1** is formally very close to Fodor's. In Fodor's theory of *asymmetric causation*, a concept R has the meaning A, if A can cause R and in addition every other type of item B that causes R is asymmetrically dependent on the fact that A causes R.[16] (where the asymmetry is defined to mean that if the A-to-R causal relation is broken, so will be the B-to-R relation, but not the opposite). In fact, the external-based scheme provides exactly such an asymmetric relation. According to it, R represents A if: R carries information

[15] It is beyond the scope of this paper to provide even a brief description of this debate, but see Fodor and Lepore (1992, 1993); Churchland (1993).

[16] I use here for illustration a simplified version of his theory (Fodor, 1990, p. 91) but see Fodor (1990, p. 93) for a more precise formulation.

about *A*, and for any *B* that *R* carries information about, this information is lower than for *B*. The main difference between this scheme and Fodor's is that while in **C1** the contrast is made in terms of a statistical (but objective) measure of mutual information, for Fodor's it is mediated by a metaphysical (or logical) contrast. A large number of theorists have found Fodor's asymmetric causation theory untransparent for not providing an explanation of what determines the asymmetric dependency (Cummins, 1989; Crane, 1995; Hutto, 1999; Eliasmith, 2000). Whether or not it may be possible to obtain a bootstrapping of content on the basis of logics and metaphysical considerations alone,[17] the scheme proposed here provides a simple naturalistic explanation to support Fodor's theory of *informational atomism*.

In addition to the external-based scheme, which may be thought of as a normative one, I have also proposed an individual-based scheme **C2**. I have argued that this scheme is, in fact, the relevant one for describing how concepts are used within a noisy environment. Unlike other referential theories (e.g., Fodor, 1990), it provides, not only an account for the content of conceptual categories, but also for their best exemplars (or prototype). A similar theory (also based on the statistical measure of mutual information) was recently developed by Eliasmith (2000), showing that it can account for another important characteristic of both psychological and neurobiological representations—their variable degree of *goodness*. The view that a statistical framework needs to replace the traditional logic-based approach for cognition and rationality has recently been strongly supported (Oaksford and Chater, 1998) within the domain of inferences and problem solving.

Despite the attractive features of the statistical approach presented, there are still a number of issues that require further scrutiny. For example, according to **C2**; every possible object can be assigned to the content of some concept representation (because of the competitive rule). If the object is totally unfamiliar, this may be an undesired feature. A number of ways may be used in order to deal with this problem. First, it is not unplausible that one of the concepts corresponds to a 'don't know' situation (see, e.g., Chappell and Humphrey, 1994, for a neural network model used to explain data in memory literature which relies on this assumption). Second, it is plausible that an unfamiliar item is categorized by a relatively general concept (say animate/non-animate) when categorization under specific concepts is not possible. Future work is needed to refine this theory providing a complete account of the representation of hierarchies of sub/super-ordinate concepts.

[17] See e.g., Fodor's discussion on pp. 125–129, 1990, that examines how God could determine what a mental symbol in a person's brain means. In the example, God sees that the symbol 'c' is caused by cats, but also that it is caused sometimes by shoes. To know that 'c' means cats, God needs to examine the counterfactuals. He will look in worlds where 'c' is not caused by cats. Fodor admits that in such worlds 'c' may still be caused by shoes. The asymmetric dependency arises because, if that happened, 'c' would not mean cats anymore. This sounds to me a bit too much like a logical bootstrapping.

Finally, I conclude by examining implications of this scheme for a theory of meaning with regards to the referential/inferential debate. One of the appeals of a referential theory of meaning is that it allows a simple explanation for the shared aspect of concepts and their role in communication in light of the strong interpersonal variability in conceptual relations (Fodor, and Lepore, 1992; Fodor, 1998; Millikan, 1998). For Fodor (1998), moreover, the meaning of a symbol is exhausted by its reference (i.e., synonyms have the same meaning if they have the same reference). In the theory presented here, the content of a concept (**C2**) is determined by causal relations to objects in the world via a probabilistic process of categorization. While this determines content in a referential way (as Fodor likes), the content of the categories are likely to develop gradually. Accordingly, infants have only a few broad (and maybe disjunctive) conceptual categories (the concept 'sheep' may represent both sheep and goats) which gradually get refined to a larger and more precise repertoire (Mandler, Bauer, and McDonough, 1991; Mandler and McDonough, 1998). Nevertheless, the variation in conceptual content for an individual person gets fixed within a few years of life to an almost common standard. This explains how, despite the interpersonal variability in the structure of conceptual interrelations, the public character of concepts is maintained via the shared reference to items in the external environment (Fodor, 1998).

Even if one accepts, however, that the primary meaning of semantic symbols is determined by their reference, it seems likely that a second component of meaning (which perhaps depends on the former) is needed to explain the ways in which concepts function in associative processes (this is exactly the part of cognition Fodor is less interested in). Such associative processes, as measured in semantic priming (e.g., Meyer and Schvaneveldt, 1971) are mainly driven by interconceptual relations, typically explained by *spreading of activation* in semantic networks (Collins and Loftus, 1975; Anderson, 1983). In the scheme I presented (see footnote 10, and Usher and Niebur, 1999; Herrmann, Ruppin and Usher, 1993), as well as in Churchland's (1998) recent version of 'space semantics', this component is driven by similarities between neural patterns of activity (see also Laasko and Cottrell, 2000). Interestingly, while Fodor has combated connectionism for a long time on the issue of compositionality of concepts (e.g., Fodor and McLaughlin, 1990), it may be possible that a connectionist type of approach, the *Attractor Neural Networks* (ANN; Hopfield, 1982; Anderson et al., 1977; Amit, 1989; Chappell and Humphrey, 1994) could provide a basis for both a referential theory of atomic symbols[18] (that can be used by a symbolic system) and for their associational powers in terms of similarity of neural patterns (Herrmann et al. 1993; Plaut and Booth, 2000).

[18] ANN are a good candidate for this because: i) they perform a *robust* multi-state categorization, ii) they can support sustained states of activations even in absence of input corresponding to concept activation in thought.

Appendix

1. Consistency of C1 and C2. According to **C1**:

$$
\begin{aligned}
&P(R_1|S_1) > P(R_1|S_2)\\
&P(R_2|S_2) > P(R_2|S_1) \qquad (7)
\end{aligned}
$$

Even if $P\ (R_1|S_1) > P(R_2|S_1)$ is not satisfied, one can check that the normalized terms (that are required by the MI formula (Eq. 6) satisfy condition **C2**

$$\frac{P(R_1|S_1)}{P(R_1)} > \frac{P(R_2|S_1)}{P(R_2)} \qquad (8)$$

Assuming that the frequencies of objects are equal $P(S_1) = P\ (S_2)$, the probabilities for the representations can be calculated: $P(R_1) = P(R_1|S_1) + P\ (R_1|S_2)$; $P(R_2) = P(R_2|S_1) + P(R_2|S_2)$ Inserting this in Equation 8 one obtains: $\dfrac{1}{1+\dfrac{P(R_1|S_2)}{P(R_1|S_1)}} > \dfrac{1}{1+\dfrac{P(R_2|S_2)}{P(R_2|S_1)}}$ which is satisfied by inserting the relations from Eq. 7 (condition **C1**).

2. An example with bias in frequencies and expectations. Assume that there are two types of external items: deer and tigers, with probabilities, $P(Tiger) = .1$ and $P(deer) = .9$, and that the *forward* conditional probability matrix, $P(R_i|S_j)/P(R_i)$, is:

	'tiger'	'deer'
Tiger	.8	.2
Deer	.6	.4

The joint probability matrix, $P(R_i,S_j)$, is:

	'tiger'	'deer'
Tiger	.08	.02
Deer	.54	.36

From this the *backward* conditional probability matrix, computed from the normalized forward matrix, $P(R_i|S_j)/P(R_i)$, is:

	'tiger'	'deer'
Tiger	1.29	.52
Deer	.96	1.05

It can be also shown that the dominance of the diagonal terms over rows of the matrix is not dependent on the external item frequencies (.1, .9) or on the specific values of conditional probabilities (.8 vs .6 and .4 vs .2) assumed here. See also Eliasmith (2000, Appendix) for a general proof.

School of Psychology
Birkbeck College, University of London

References

Amit, D.J. 1989: *Modeling brain function: The world of attractor neural networks*. Cambridge, UK: Cambridge University Press.

Anderson, J.A., Silverstein, J.W., Ritz, S.A. and Jones, R.S. 1977: Distinctive features, categorical perception and probability learning: Some applications of a neural model. *Psychological Review*, 84, 413–451.

Anderson, J.R. 1983: *Cognitive Psychology and its Implications*. New York: W.H. Freeman and Company.

Bialek, W., Rieke, F., VanSteveninck, R.R.D. and Warland, D. 1991: Reading a neural code. *Science*, 202, 1854–1857.

Chappell, M. and Humphrey, M.S. 1994: An autoassociative neural network for sparse representations — analysis and application to models of recognition and cued recall. *Psychological Review*, 101, 103–128.

Churchland, P.M. 1993: Fodor and Lepore: state-space semantics and meaning holism. *Philosophy and Phenomenological Research*, LIII, 667–672.

Churchland, P.M. 1998: Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered. *Journal of Philosophy*, XCV, 5–32.

Collins, A.M. and Loftus, E.F. 1975: A spreading activation theory of semantic processing. *Psychological Review*, 82, 407–428.

Crane, T. 1995: *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. London: Penguin.

Cummins, R. 1989: *Meaning and Mental Representation*. Cambridge: Mass., London, MIT Press.

Dretske, F. 1981: *Knowledge and the Flow of Information*. Cambridge Mass.: MIT Press.

Dretske, F. 1983: Precis of knowledge and information flow. *The Behavioral and Brain Sciences*, 6, 55–90.

Dretske, F. 1983b: The epistemology of belief. *Synthese*, 55, 3–19.

Eliasmith, C. 2000: *How Neurons Mean: A Neurocomputational Theory of Representational Content*. Ph.D. dissertation, Dept. of Philosophy, Washington University in St. Louis.

Fano, R.M. 1961: *Transmission of information*. Cambridge Mass.: MIT Press.

Fodor, J.A. 1975: *The Language of Thought*. New York: Crowell.

Fodor, J. 1984: Semantics, Wisconsin style. *Synthese*, 59, 231–250.

Fodor, J. 1990: *A Theory of Content and Other Essays*. Cambridge, Mass.: MIT Press.

Fodor, J. 1998: *Concepts: Where the Cognitive Theory Went Wrong*. Oxford: Clarendon Press.

Fodor, J.A. and Lepore, E. 1992: *Holism: a shopper's guide*. Oxford: Blackwell.

Fodor, J.A. and Lepore, E. 1993: Reply to Churchland. *Philosophy and Phenomenological Research*, LIII, 679–682.

Fodor, J.A. and McLoughlin, B. 1990: Connectionism and the problem of systematicity: why Smolensky's solution does not work. *Cognition*, 35, 183–204.

Harman, G. 1982: Conceptual role semantics. *Notre Dame Journal of Formal Logic*, 23, 242–256.

Harnad, S. 1990: The symbol grounding problem. *Physica D*, 42, 335–346.

Herrmann M., Ruppin E. and Usher M. 1993: On the Dynamic Activation of Memory *Biological Cybernetics*, 68, 455–463.

Hinton, G.E. and Sejnowski, T.J. 1983: Optimal perceptual inference. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 448–453.

Hinton, G.E. and Sejnowski, T.J. 1986: Learning and relearning in Boltzmann machines. In Rumelhart, D.L., McClelland, J.L., and the PDP Research Group, (1986), pp. 282–317.

Hume, D. 1739: *A Treatise of Human Nature*. Ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1888 (First Edition, 1739).

Hopfield, J.J. 1982: Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, 79, 2554–2558.

Hutto, D.D. 1999: *The Presence of Mind*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Laasko, A. and Cottrell, G. 2000: Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology*, 13, 47–76.

Mandler, J.M., Bauer, P.J. and McDonough, L. 1991: Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology*, 23, 263–298.

Mandler, J.M. and McDonough, I. 1998: On developing a knowledge base in infancy. *Developmental Psychology*, 34, 1274–1288.

Massaro, D.W. 1989: Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, 21, 398–421.

McClelland, J.L. 1991: Stochastic interactive activation and the effect of context on perception. *Cognitive Psychology*, 23, 1–44.

Meyer, D.E. and Schvaneveldt, R.W. 1971: Facilitation in recognition of pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227–234.

Millikan, R.G. 1989: Biosemantics. *Journal of Philosophy*, 86, 281–297.

Millikan, R.G. 1993: *White Queen Psychology and Other Essays for Alice*. Cambridge: Cambridge University Press.

Millikan, R.G. 1998: A common structure for concepts of individuals, stuffs, and real kinds; more mamma, more milk and more mouse. *Behavioral and Brain Sciences*, 9, 55–100.

Movellan, J.R. and McClelland, J.L. 1995: *Stochastic interactive activation, Morton's Law, and optimal pattern recognition* (Technical Report PDP.CNS.95.4). Pittsburgh, PA 15213: Department of Psychology, Carnegie Mellon University.

Neander, K. 1995: Misrepresenting and malfunctioning. *Philosophical Studies*, 79, 109–141.

Oaksford, M. and Chater, N. 1998: *Rationality in an Uncertain World*, Psychology Press Ltd., Hove.

Papineau, D. 1987: *Reality and representation*. Oxford: Basil Blackwell.

Plaut, D.C. and Booth, J.R. 2000: Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, 107, 786–823.

Rumelhart, D.L., McClelland, J.L. and the PDP Research Group 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 1. Cambridge, MA: MIT Press.

Searle, J.R. 1980: Minds, brains and programs. *Behavioral and Brain Sciences*, 3, 417–457.

Searle, J.R. 1984: *Minds, Brains and Science*. London: British Broadcasting Corporation.

Shadlen, M.N. and Newsome, W.T. 1994: Noise, neural codes and cortical T organization. *Current Opinion in Neurobiology*, 4, 569–579.

Softky, W.R. and Koch, C. 1993: The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience*, 13, 334–350.

Stampe, D. 1977: Towards a causal theory of linguistic representation. In P. French, D. Euhling and H. Wettstein (eds.), *Midwest studies in philosophy*, 2, 42–63. Minneapolis: University of Minneapolis Press.

Usher, M., Stemmler, M., Koch, C. and Olami, Z. 1994: Network amplification of local fluctuations causes high rate variability, fractal firing patterns and oscillatory local field potentials. *Neural Computation*, 6, 795–836.

Usher, M. and McClelland, J.L. 2001: The time course of perceptual choice: The Leaky Competing Accumulator Model. *Psychological Review*, in press.

Usher, M. and Niebur, E. 1999: Mental representations: a computational-neuroscience scheme. In *Understanding Representation in the Cognitive Science*. A. Riegler and M. Peschl. (eds.), Plenum Press, pp. 135–143.