Relevance realization and the emerging framework in cognitive science

John Vervaeke[1,2], Timothy P. Lillicrap[3], Blake A. Richards[4]

1. Cognitive Science Program, University College, University of Toronto
2. Department of Psychology, University of Toronto
3. Centre for Neuroscience Studies, Queen's University
4. Department of Pharmacology, University of Oxford

Send correspondence to:

Email: tim@biomed.queensu.ca

Phone: (613)-767-9463

Timothy Lillicrap
Centre for Neuroscience Studies
Botterell Hall, Room 234
Queen's University
Kingston, Ontario
K7L 3N6

Abstract

We argue that an explanation of relevance realization is a pervasive problem within cognitive science, and that it is becoming the criterion of the cognitive in terms of which a new framework for doing cognitive science is emerging. We articulate that framework and then make use of it to provide the beginnings of a theory of relevance realization that incorporates many existing insights implicit within the contributing disciplines of cognitive science. We also introduce some theoretical and potentially technical innovations motivated by the articulation of those insights. Finally, we show how the explication of the framework and development of the theory help to clear up some important incompleteness and confusions within both Montague's work and Sperber and Wilson's theory of relevance.

Keywords:
Relevance, Constraints, Self-organization,
Opponent processing, Framework.

Introduction

There is a family of seemingly intractable problems in cognitive science. In each individual case, it is unclear how it might be resolved, but the problems are central and so cannot be ignored or marginalized. A well known example is the problem of combinatorial explosion which faced the general problem solving (GPS) framework of Newell and Simon [33]. We demonstrate here that this problem, and others, have remained intractable because of a theoretical circularity caused by the centrality of relevance to cognitive function. Attempts to deal with this circularity appear to have been hampered by what we think is a confusion about what it is we can scientifically explain. We will argue that one cannot have a theory of relevance itself because there is no stable, homogeneous class of entities which correspond to the term "relevance". However, we believe a theory of the mechanisms of how relevance can be realized is tractable. We will call this a theory of *relevance realization*. Our argument is analogous to the idea that one cannot have a theory of biological fitness, but one can have a theory of the mechanisms of natural selection that realize it. To draw this all together, we argue that perhaps the only way cognitive science can hope to circumvent this family of problems is to develop a non-circular theory of relevance realization.

In this essay, we begin by describing how these intractable problems in cognitive science lead to circular theories due to the issue of relevance. We then discuss an important methodological move: the attempt to circumvent

these problems through what Dennett calls "reverse engineering". But following an argument from Green, reverse engineering will require a criterion of the cognitive which does not rely upon folk-psychological intuitions about the nature of cognition; they themselves presuppose relevance realization and so will simply return us back to the intractable problems. We argue that the best way to avoid this threat is to make relevance realization the criterion of the cognitive.

If this is the case, we need some plausible account of what the mechanisms of relevance realization would look like. Of course, a complete specification of these mechanisms will require significant empirical work. The best that we can hope to provide here is a plausible account of the required structural principles for these mechanisms. To this end, we describe three important lower order constraints and a fourth higher order constraint used by cognitive agents which, when considered as dynamic, opponent processes, could help to produce a non-circular structural theory of relevance realization. Finally, we use our theory of relevance realization to critique previous theories of relevance (e.g. Sperber and Wilson's [47]) which we believe fall prey to a recursive regress.

Much of the machinery on which our theory runs has been borrowed directly from modern practice in the sub-disciplines of cognitive science (e.g. linguistics, machine learning, neuroscience, psychology). In this paper we attempt to draw connections between these various pieces of work, and in so doing, our aim is to contribute to what we view as the already emerging

framework in cognitive science.

The centrality of relevance realization to cognitive science.

The origins of cognitive science as a discipline are rooted in the research programs of the 20th century that sought to explain cognition in terms of computation and language, thus bringing together the disparate disciplines of cognitive psychology, artificial intelligence, and linguistics. Historically, some of the most central areas of research within this interdisciplinary framework have been problem solving, causal interaction with the world, categorization, induction, and communication [15, 44]. In each of these areas significant problems arose that prevented any single theory from gaining widespread acceptance. We will argue that these problems were related via relevance realization.

Relevance realization and problem solving.

The foundational framework for understanding problem solving for both cognitive psychology and AI is the general problem solving (GPS) framework of Newell and Simon [33]. One of the successes of this framework was to reveal an aspect of problem solving which initially was counter-intuitive, viz., combinatorial explosion [26, 32]. In order to understand combinatorial explosion, one needs to understand how problems were represented in the GPS framework. In this framework, a problem is represented by four elements: a representation of the initial state, a representation of the goal state, a rep-

resentation of all of the operators an agent can use to turn one state into another, and finally path constraints which disallow certain types of solutions [26, 32]. Taken together, these elements generate a problem space or search space which consists of all the possible sequences of states that the agent could take. A solution consists in finding the sequence of operations which will take the agent from the initial state to the goal state while obeying path constraints.

The GPS model was useful because it made apparent that for most problems which humans solve, the associated search spaces are vast and complex. For example, consider a typical chess game. On average, for each turn there are ≈ 30 legal operations you can perform, and there are typically 60 turns in a game. So the number of alternative sequences you would have to search in order to find a path from the initial state to the goal is F^D, where F is the number of operators and D is the number of turns. So in our chess example the number of pathways you would have to search would be 30^{60} which is a very large number. This number of paths is far too large for any conceivable computer to search exhaustively (consider for comparison that the number of electrons in the entire universe is estimated at $\approx 10^{79}$).

Nevertheless, humans successfully wend their way from initial states to goal states all the time (while respecting path constraints). How do they do it? How do they do a search through that space in a way that is non-exhaustive, but still intelligent? In practice, people make use of heuristic search in which large regions of the search space are not considered [33,

p. 96]. But typically models of such heuristics are hand crafted by AI practitioners for the problem of interest, and there are no general theoretical accounts of how to mechanistically generate heuristics powerful enough to produce human level competence.

A major failure of the GPS framework was that it relied on the assumption that problems form a well defined class and that most problems could easily be turned into well defined or formal problems [21]. It became clear that for many real world problems (e.g taking good notes during a lecture), the initial state, goal states, path constraints, and operators are either incomplete, vague, or missing. Such problems are known as 'ill defined problems'. Human being's ability to avoid combinatorial explosion stems from the way they can convert ill-defined problems into well-defined problems [26]. Certain actions are immediately ruled out via non-inclusion in the problem space during problem formulation, thereby making the search space far more tractable (e.g. no one even considers including the ambient temperature of the room while taking notes in a lecture). The key is our ability to zero in on the relevant information and the relevant structure of the information to perform the actions needed for good problem formulation.

However, to determine what is relevant to a problem also involves determining what is irrelevant! As such, the smaller search space which problem formulation affords us is only achievable if we initially consider the larger search space and segregate the relevant from the irrelevant. This leads us back into the problem of combinatorial explosion. We are caught in a nasty

circle here in which we need good problem formulation in order to deal with combinatorial explosion, and yet good problem formulation seems to be a combinatorially explosive problem. Only an account of how people realize what is relevant while avoiding combinatorial explosion can break through this theoretical circle.

Relevance realization and interaction with the environment.

The necessity of breaking this circle is especially apparent when we consider how agents take action in the environment while intelligently dealing with unintended side effects. This is best illustrated with an example from Dennett [8] about a robot trying to acquire its food in a very basic manner. Suppose we have a robot designed to retrieve batteries as its food source and then transport those batteries to a location where they can be consumed. Also, suppose our robot comes upon a wagon upon which there is the battery, but unfortunately there is also a bomb on the wagon. The robot correctly deduces that if it pulls the wagon then the battery will come along as an intended effect. However, an unintended side effect is that the bomb comes along and destroys the robot. As its designers we attempt to remedy this situation by having the robot deduce not only the intended effects of its actions but also potential side effects. When we test our improved robot in a repeat of the original situation we find it stopped and endlessly calculating for the simple reason that the number of potential side effects it can consider is indefinitely large. We seek to remedy this by having the robot form a list of potentially

relevant side effects. This part of our thought experiment requires that we have some theory of the properties of information that renders it relevant. However difficult this assumption is let us grant it for the sake of continuing the thought experiment in order to further see what it reveals. Once again we find our robot stopped endlessly calculating. This is because it is creating two lists. One of potentially relevant side effects and one of irrelevant potential side effects, and because each list is again indefinitely large, the calculation cannot be completed.

As Dennett's example illustrates, for an agent to to take action in even relatively simple circumstances it must somehow intelligently ignore a great deal of information, but this seems paradoxical. This requires zeroing in on the relevant information while not even considering most of the irrelevant information. It requires putting a frame around one's cognition. This is therefore a generalized version of the frame problem [36], indicating that we have on our hands an essential problem that extends well beyond this case.

Relevance realization and categorization.

The same dilemmas plague another another central component of research in cognitive science, namely categorization. The importance researchers have placed on categorization has been in part motivated by the acceptance that there are deep connections between what is required for problem formulation and what is required for categorization of novel information [21]. Categorization is the process by which we create classes which support powerful

inductive generalizations that are relative to the features held in common by the members. In other words, the members of a category are only identical in so far as they contribute to the inductive generalizations. Yet, as Goodman famously noted, any two objects can be infinitely similar or dissimilar [16], which presents any formal theory of categorization with a seemingly insurmountable computational task similar to the combinatorial explosion faced by problem solvers. Thus, the immediate issue is how we zero in on the relevantly shared properties which will be useful for inductive generalizations. We think that this issue is encountered by schema [38], script [39], and stereotype or prototype theories [21], as well as more recent theories which deal with context sensitivity in categorization [14]. Each one of these families of theories involves an implicit theory of relevance realization that is presumed to solve this computational nightmare, wherein relevance is usually specified in terms of one static property of information such as frequency, or invariance, or prototypicality, etc. However, it quickly became apparent that the attempt to capture relevance in this manner fails. For example, Medin, in an influential review article noted the following about prototype theory [29]:

"Prototype theories imply constraints that are not observed in human categorization, predict insensitivity to information that people readily use, and fail to reflect the context sensitivity that is evident in human categorization. Rather than getting at the character of human conceptual representation, prototypes appear to be more of a caricature of it." (p. 1472)

Medin's point is that prototypicality does not capture the relevance realization needed for categorization, and it is therefore failing in a deep way. A much more complex and dynamic account of relevance realization is needed. (see [30] for a more recent review that argues that the dynamic complexity of the information integration within concept formation and use may require a "psychometaphysics" in which concepts are embedded in theoretical and explanatory projects. These are projects that would clearly require sophisticated relevance realization.)

Despite the empirical merit possessed by these implicit theories of relevance realization they have been inadequate in generating good models of categorization [29, 30]. We suggest that this is because relevant information cannot simply be always identical to frequent, or invariant, or prototypical information, because relevance is context sensitive. Dissenting voices such as Barsalou [2] have presented similar arguments, and argued for making the context sensitive application of information central to what it is to be a concept. Indeed, he claimed that "a concept can be viewed as an agent-dependent instruction manual that delivers specialized packages of inferences to guide an agent's interactions with particular category members in specific situations" [2, p. 626]. The circle that looms here, of course, is that at any given time a context could contain an infinite number of variables or predicates regarding the environment. A much more sophisticated account of relevance realization is needed, which breaks through this threatening circle by dynamically integrating features such as frequency, invariance, etc.,

rather than following a strict, context insensitive rule.

It is interesting to note that as it became apparent that script, schema, and stereotype theories suffered from "tunnel vision" [21, 56, 51], researchers tried to understand intelligence in terms of how we are embodied and embedded in the world, i.e. how we operate in the environment [59], in order to get at the missing contextual sensitivity. On this view, if we can understand how agents utilize the world around them that would provide foundations for how categorization and problem solving take place. Before an agent can generate facts or formulate problems, information needs to be processed in terms of how it is relevant for successful action in the environment. Thus, at the core of any embodiment thesis is the idea that we are acting in the world. Of course, *intelligent* agents need to be able to successfully couple their actions to future effects in the world. However, once you try to have a reliable tracking between cause and effect you face a deep problem, viz., the problem of dealing with side effects as discussed above. We are unaware of any research program within the embodied cognition movement that can solve this problem.

Relevance realization and rationality.

There is a tempting philosophical strategy in the face of such problems within psychology and artificial intelligence. Most theories within these disciplines presuppose a background normativity, specifically, that cognitive processes should be rational in nature. Perhaps one could alleviate the problems of

circularity that we have highlighted if we explicate the normativity of rationality presupposed in much of cognitive science. Many philosophers have undertaken this project [4, 20, 49, 43]. We will examine two whose work we consider representative in nature and whose work highlights the centrality of relevance realization to rational induction.

Cherniak [4] has influentially argued that you cannot say that to be rational is to simply be logical because many algorithms couched in logic lead to combinatorial explosion. He argues instead that to say that we are rational means that we've zeroed in on some of the relevant subset of logical inferences for the task at hand. He therefore calls attention to the fact that relevance realization is central to the issue of rationality in general. Cherniak attempts to explain relevance realization in terms of memory compartmentalization. Cherniak's idea is that a system can avoid a combinatorially explosive search through memory for relevant information by dividing memory into compartments. The system only searches one compartment at a time. If the compartments are labelled in some fashion, i.e., if they are content addressable, then the search goes right to the relevant compartment and the search is thereby constrained to just that compartment. However, Chiappe & Vervaeke [5] have argued that this account presupposes in a vicious way the very thing it is an attempt to explain because the formation and use of such compartments requires the ability to determine how things are relevant to each other. We will go a step further and suggest that *any* explanation of relevance realization in terms of pre-existing memory organization

will presuppose the very thing it is attempting to explain. This is because any content addressable information organization scheme requires successful categorization of the information, and as we have outlined above, categorization itself requires relevance realization! Naturally, memory organization can play a role within an account of relevance realization, but it cannot be the foundational role assigned to it by Cherniak, or to theories like his.

In a fashion similar to Cherniak, Putnam [34] argues that inductive inferences cannot rely on an invariant syntactic formalization. This is because the success of any particular inductive logic is relative to the environment in which it is operational. For example, one could have an environment which is noisy, and so requires cautious induction, but if the environment contains little noise, it is beneficial to act less cautiously. Putnam's idea is that inductive logics can vary in a context dependant fashion, e.g. by changing a caution parameter or the mechanism for generating inferences, but still arrive at similarly rational results. To extend the example: one person can use a specific inductive logic in some environment and inductively conclude some particular belief while another person could be using a different inductive logic in a different epistemic environment, and produce the same inductively justified belief, though the underlying computational processing / states would be different. Of course, to implement this sort of environmentally dependent inference a system must be able to determine the context of the inductive inference, and as we have already noted, such an ability presupposes relevance realization. Thus, Putnam's refinement of the normativity of

rationality returns to the problem of relevance realization, just as Cherniak's did. We would also like to note that Putnam's argument highlights a point we will develop later, viz., that the attempt to solve such problems through an invariant syntax of inference seems to be seriously misplaced.

Relevance realization and communication.

Similar problems are faced by research into language and communication. Although a great deal of the modern generative linguistics program is focused on purely syntactic questions [7], it is undeniable that an important aspect of our linguistic cognition is the pragmatic aspect. Any attempt to explain the nature of language must account for the relationship between language and communication, and any theory which fails to do this is a failed theory of language. Indeed, a central insight from seminal philosophers like Austin and Grice about language is that we are often doing more than just making statements [1, 19]. For Austin, we are often performing actions instead of stating things, and for Grice we are often conveying more information than we are stating.

Grice developed this theory of conveyance in his theory of conversational implicature. What comes out of this work is that you cannot fold conversational implicature into semantics because you would overload the lexicon. Instead of folding implicature into a lexicon, it must be something that is continually worked out between individuals. Grice's insight is that 'working out' is rational cooperation, and is thus constrained by four maxims: 1.

quantity, 2. quality, 3. manner, and 4. relevance. Thus, Grice's work gives relevance a central role in communication.

It is possible to take this even further, though. Sperber and Wilson [47] argue that the first three of these maxims all collapse into the fourth maxim of being relevant. They argue along the following lines, the maxim of quantity is just 'provide the relevant amount of information', the maxim of manner just ends up being 'use the relevant format'. The maxim of quality is a little less clear because it requires truthfulness, and truthfulness is not as easily seen to boil down to relevance. But, the maxim of quality cannot be the rule 'convey all that is possibly true about what you are thinking,' as we have already seen in our discussion of Cherniak. So complete truth cannot be the normative standard that people imply. Therefore, what people must be sharing are the relevant truths. We are now presented with the same computational dilemma we encountered before: there are an infinite number of truths, and *prima facie* it would be impossible to segregate the relevant from the irrelevant in a computationally tractable manner.

Thus, theories of pragmatics suffer the same problems with relevance realization that theories of problem solving, categorization, and action do. This time though, Sperber and Wilson explicitly identify this problem and attempt to generate a theory of relevance. We think that is not a sufficient account, but their account will play an important role in the framework we lay out later in this paper.

To summarize, we have demonstrated that fundamental research streams

in cognitive science have encountered essential problems related to computational intractability that have frustrated any attempts at formal theories of problem solving, categorization, action, induction, and communication. We have also shown that these problems form a family of interdependence, with the problems from one area leading to difficulties in others. Thus, we've not only shown the pervasiveness of relevance realization but also that it's theoretical appearance is systematic.

Reverse engineering and the emerging criterion of the cognitive.

In the face of these daunting problems, one may attempt to get around these thorny conceptual/theoretical issues concerning relevance realization via 'reverse engineering' [8]. Under this methodology, attempts to directly understand cognition are abandoned, and research is instead focused onto designing an intelligent machine. The hope is that if this research program is successful we will understand various cognitive phenomena in virtue of having designed a machine that exhibits the phenomena. Such principles therefore become one's theoretical account of the cognitive processes that generate the intelligent behaviour. Hence, artificial intelligence would allow us to work backwards into theories of things like categorization, communication, etc., while avoiding the theoretical circularities we have encountered in the other direction.

However, we are assuming here that we will know when a machine is in fact intelligent. The obvious methodological question is how to establish

this identity. Turing famously proposed just such a test [53]. As Fodor [11] points out, the Turing test has an important methodological principle at work; namely that the test screens off certain factors of comparison. However, this principle tacitly makes use of our assumptions about what constitutes intelligence. These assumptions will of course be shaped by our current folk psychology and/or our explicit scientific psychology.

In 1994, Green [17], pointed out that unless we have some agreement about how to pick out cognitive phenomena (i.e. a criterion of the cognitive), we will never be able to determine the correct factors for comparison. Green reviews some of the common criteria. None of these are are widely agreed upon, i.e. they cannot be used to powerfully pick out examples of cognitive processes. For example, the view that seemed to be achieving success in the mid 1980s was that cognitive processes are inferential processes operating on syntactic representations [35]. But this fell under heavy criticism from the connectionists [37, 18, 46, 45]. The original formulation of the Turing test biases one towards paying attention to inferential and language-like features of cognition. A connectionist would be very unhappy with any version of a test for the cognitive which only pays attention to these factors. Without a powerfully applicable criterion of the cognitive, the interpretation of simulations, and hence the whole A.I. project, will remain seriously in question.

We propose that the systematic importance of relevance realization to cognitive processes makes it the obvious choice for a criterion of the cog-

nitive. Put succinctly: *any attempt to engineer an intelligent system must ultimately focus on the development of a system that can realize relevance.* In fact, we believe that a great deal of current work in machine learning and theoretical neuroscience should be viewed in these terms. Examples include current work in categorization [24, 25, 55], optimal control [52, 40, 28], and reinforcement learning [50, 10, 9]. In each of these cases a significant focus of the research program is the development of systems that can cope with the computational intractability rooted in the need for relevance realization. Indeed, the goal of such research could be said to be the development of systems that can determine relevant features, controls, or actions for problems encountered in the real world. Thus, we put forward that such a criterion of the cognitive is already emerging [59]. Unlike in the past where generally the criterion was intuitively generated from our folk psychology, current work focuses on methods to solve the problems that were encountered in the past when applying such intuitive criteria. Therefore, the methodological move is to base one's criterion of the cognitive on whatever facilitates solving these difficult technical problems. As such, an explicitly developed theory of relevance realization will help the development and application of new techniques and theories, and ultimately, our understanding of cognition[1].

[1]It may be helpful to contrast what we believe to be the central framing metaphor for the criterion which held prominence until the mid 1980s with the new emerging metaphor. The classical metaphor is that cognition is essentially computation and that the brain is essentially a computer. The idea is that information is organized around inferential, and syntactic relations which are isomorphic with a linear causal order. This program is implemented on a static hardware, and because the hardware is static it is almost completely irrelevant to the software. The development of the hardware is also irrelevant

Towards a theory of relevance realization.

Before we begin to articulate even a cursory theory of relevance realization we feel it is important to identify a framework that will avoid circularities. This is because theoretical work on relevance will lead to regress both due to its privileged position within cognition, as we have shown above, and due to its position within the practice of science itself. In this section we describe how this leads to circularities and we identify three guidelines for theorists that will help avoid them. The first guideline is that a theory of relevance is impossible, and instead a theory of self-organizing mechanisms for relevance realization is what's required. The second is that a theory of the mechanisms of relevance realization must not be representational or syntactic, but economic. The third is that a theory of relevance realization cannot rely on a completely general purpose learning algorithm, but must involve competition between multiple competing learning strategies.

---

except insofar as it is a process for producing the mature hardware that can run the software. The computer is a stable logic machine.

In contrast, we might call the new metaphor the Logos (capturing both the sense of logistics and the Greek sense of the term as: making information belong together) Multi-Machine (LMM). Here, information is organized in terms of economic properties (which we will discuss in greater detail later) and relevance relation, viz., how information can be economically integrated together to support successful interaction with the world. This is isomorphic not with a linear causal order, but with a circular causal order of a self-organizing dynamic system. This system is instantiated in a plastic neural network. The brain is a Multi-Machine: a machine which can make itself into new kinds of machine such that it not only learns but increases its capacity for learning. In this new model there is no clear line between the 'hardware' and 'software' since both influence each other as they run. Thus the developmental history of the hardware is always relevant to the explanation of cognition. Note also that one of the things that an LMM can develop is a computer, i.e., one of the machines within the multi-machine of the brain can be a computer.

The importance of a theory of a self-organizing relevance realization mechanism

If relevance realization is to be the criterion of the cognitive, a naive assumption might be that you have to come up with an account of relevance with which you can pick out all relevant things in order to find generalizations over the class. In this sense, you may try to come up with a theory of what relevance *is*. This is the tactic that has been employed by other researchers to date (cite Sperber & Wilson, etc.). However, we believe that this is a fundamental mistake.

To begin with, consider by analogy the role played by 'fitness' in the theory of evolution. A common confusion regarding fitness is that what makes a creature fit is the possession of one or more of the defining features of fitness. Thus, people sometimes misconceive of evolution as the designing of particular features like speed, intelligence, acute vision, etc. Of course, in reality the class of all *possible* organisms which are fit is completely heterogeneous, unstable, and dependent on context. As such, we can make no systematic inductive generalizations about the class of fit organisms. What evolutionary theory provides is not an account of the biological features that define fitness, but a mechanism by which fitness is realized in a contextually sensitive manner. Therefore, by strong analogy with the centrality of natural selection in biology, we do not really want a theory of relevance. We want instead a theory that articulates a mechanism for how relevance is realized in a contextually sensitive manner. We assert that the need for approaching

relevance in this way is a direct result of circularities inherent in attempting to build a theory of a phenomenon which underpins all the other phenomena within the discipline. This is especially the case for relevance due to the fact that the practice of science itself relies on these cognitive phenomena which are dependant on it. This has previously been highlighted by Chiappe & Vervaeke [5]. We briefly review the issue here.

Any scientific statement has to be protected by provisos which keep it from being trivially falsified. For example, we say that "sugar is soluble", even though someone might for instance freeze the water just as the sugar was added, etc. The list of such "falsifications" is obviously long and heterogeneous but they are considered irrelevant to the scientific statement. It looks like we have run into a circularity here: the very articulation of a theory of relevance would require an implicit identification of what is and isn't relevant regarding the application of the statement. This runs us into a kind of chicken and egg problem for cognition.

In a similar fashion, any scientific statement is going to rely upon pragmatic conveyance for its interpretation and understanding. But attempts to render pragmatic conveyances into semantic statements meets the problem that a stipulation of these conveyances relies upon implicatures. Normally, we rely upon people's ability to realize relevant implications and implicatures in order to make communication practicable (see Grice / Sperber and Wilson above). But, if our goal is to design a theory of relevance itself it should not include an assumption of the relevant implications for a cognitive agent.

Whatever else a scientific statement needs to be, it needs to be communicable since science is essentially a community enterprise, so this difficulty with conveyance cannot be ignored.

Finally, any scientific explanatory statement(s) involves inference to the best explanation which requires reference to the contrast class (the set of explanations it is better than). Of course, given the under-determination of a theory by its data, the number of alternative yet acceptable explanations is infinite. Typically the abductive selection that allows scientists to ignore these alternatives is usually explained in terms of concepts such as simplicity and similarity. Yet, this relies heavily upon relevance since it will depend on the relevant features of the explanations for determining simplicity and similarity. So a theoretical explanatory statement about the nature of relevance would rely again upon the very thing we are attempting to explain.

For all of these examples we end up with a seemingly vicious infinite regress. Importantly though, there is an implicit foundationalism in this discussion; that is, that the only way to stop this infinite regress is to find some 'urrelevance' around which the chain of explanation is designed. However, there is a ready alternative to such foundationalism - an evolutionary coherentism in which relevance realization is a dynamic, self designing, self-organizing process that is not equated with an immutable identity. In the same fashion that evolutionary theory dissolves the chicken and egg problem, we believe that a non-definitional theory that proposes self-organizing mechanisms for relevance realization will be able to dissolve the threat of vacuous

or cyclic explanations in cognitive science. We need to radically give up an "intelligent design" framework for understanding the generation of relevance realization by cognitive systems.

The requirement for an economic model of relevance realization

Even if we resolve the theoretical regress by moving to a self-organizing, non-definitional model, there is a further difficulty remaining which was described by Vervaeke [57]. It seems *prima facie* that relevance realization is going to be relative to the interests and goals of an organism, and interests and goals are about future states of affairs. To direct behaviour towards some future state of affairs requires some representation of that state of affairs. However, representations are aspectual by nature since one does not represent all of the aspects or features of a thing. This means that one needs relevance realization to generate good representations because one has to pick which aspects are relevant to represent. This is a patently vicious explanatory circle. Note that this is a problem in our attempts to explain the brain's behaviour; it is obviously not a problem with the brain's functioning.

So, the somewhat unintuitive move is to consider that theories of relevance realization should be pitched at a sub-representational level to avoid this theoretical conundrum. What this means is that the theory can only initially make use of completely immanent properties and relations of the information available to the brain. The prototypical response to this issue in cognitive science has been to drop to the logical/syntactic level of ex-

planation since the formal properties are completely self-contained and only become representation when they have been assigned content. This is the theoretical framework behind the many variants of computation functionalism.

One of the founding figures of computational functionalism, Jerry Fodor [12], has recently made an argument that one cannot capture relevance in the syntactic structure of tokens within a formal system. Fodor's argument is that notions like relevance or centrality or importance are all aspects of cognitive commitment, i.e., how much a system cares about something and devotes its resources to it. Cognitive commitment is an economic issue, and as such it is both globally defined and contextually sensitive. This means that it cannot be captured in the syntax of a token since the syntax by its very logical nature does not consider economic issues, must be locally defined, and operates in a contextually invariant manner. As Fodor puts it, syntax is locally defined but relevance is globally defined and therefore cannot run off of syntax alone. Fodor rightly regards this as a devastating problem for computational functionalism because very many important processes cannot be captured by syntax. This means that any viable theory of relevance realization is going to have to make use of sub-syntactic (e.g. using vectoral representations) properties of information whose operations are locally defined but have global effects. This tact is becoming relatively widespread in cognitive science, with examples surfacing in the work of Hinton [24], and Gärdenfors [14].

One possible way to meet this demand is to turn to the economic properties of information and action. We notice that in economies, decisions are made locally (e.g. you buy milk at the grocery store), but these local decisions contribute to the global organization, and of course, the global organization constrains future local processing. This is all done without centralized control, i.e, it is self-governing. This is exactly the sort of mechanism which Fodor was looking for.

The economic approach uses internal measures of cost (e.g. metabolic), and reward (e.g. dopaminergic). The reader may worry that we are invoking a pre-established harmony between the internal running of the cognitive economy and successful behaviour in the world. The worry of course is that pre-established harmonies often presuppose a god-like designer. However, we have no such worry since we can confidently presume that, to the extent that it exists, evolution has worked out this pre-established harmony.

Obviously, the fact that the cognitive economy is internal, does not mean it's causally isolated. Like all economies the cognitive economy has imports and exports. It imports sensory data, and exports motor commands to the world. The exports have effects in the world which to some degree constrain the future imports. Just as we can intelligently talk about the American economy as a distinct entity that is nevertheless causally interwoven with other economies we can talk about the internal cognitive economy even though it is causally interwoven with the world. The important relation to discover then, the pre-established harmony, is the balance of internal economic variables

whose processing results in successful interaction with the world.

The impossibility of a general learning algorithm solution

Relevance realization must be a set of pervasive constraints on processing rather than a specific machine for realizing relevance. The reason for this is that if relevance realization took place within a specialized mechanism, the mechanism itself would merely confront the frame problem. The problems of how to decide what to ship there, and how it might avoid the combinatorial explosion of doing so, would immediately and continually arise. This sort of a move only shifts the problem. It does nothing to solve it. Thus, there cannot be a relevance realization mechanism in any straightforward sense. As we've said we need to give up any intelligent design framework for understanding relevance realization.

One area where this point can be brought to bear is in the field of meta-learning. This is an exciting new field of research that exemplifies many of the features of the emerging framework we are explicating. In meta-learning systems one has basic learners that apply specific learning algorithms/strategies to current problems. In addition to this one has a higher order meta-learner whose job it is to assess the applicability (an important concept as we shall see) of various different learning strategies across time so as to improve the selection and operation of the base learning strategies. This makes the system a self-adaptive learner. However, one conceptual problem which seems to confront this approach is explained well in a recent review by Vilata and

Drissi [58]. They point out that the advantage of meta-learning over standard learning is that the bias (the set of assumptions in the algorithm that restricts and structures the problem space) in the base level learning algorithm is no longer fixed *a priori* but is dynamically modified by the meta-learner. This addresses one of the central issues of relevance realization in problem solving. However, the meta-learner itself must have a learning algorithm with such an a priori fixed bias. One can address this by having a meta-meta-learner, but an obvious infinite regress now ensues. This conceptual problem seems to involve an imposition of an intelligent design framework in which there must be a specific machine that ultimately decides how relevance is assigned. We argue that the solution to such problems is to make the meta-learning self-organizing and immanent to the learning. We outline below how we think this might be done.

It is important to note though that a meta-learner cannot simply be a general purpose, higher-order learning algorithm. There were hopes in the early machine learning literature that a sort of general learning algorithm could be found that would do well on all problems [3, 60, 54]. But, just as the GPS framework met insurmountable problems, the hopes of a general purpose learning algorithm were eventually dashed as well:

> Although the human brain is sometimes cited as an existence proof of a general-purpose learning algorithm, appearances can be deceiving: the so-called no-free-lunch theorems [Wolpert, 1996], as well as Vapnik's necessary and sufficient conditions for consis-

> tency [Vapnik, 1998, see], clearly show that there is no such thing as a completely general learning algorithm. All practical learning algorithms are associated with some sort of explicit or implicit prior or bias that favours some functions over others. Since a quest for a completely general learning method is doomed to failure, one is reduced to searching for learning models that are well suited for a particular type of task.

Given the no-free-lunch theorems - which demonstrates that all learning algorithms are inherently tuned to some subset of problems - one important tact a cognitive system can use is to adopt strategies that are complementary in that they have goals that are in a trade-off relationship. The system can then use opponent processing in order to continually redesign the learning strategy it is using. Opponent processing is a powerful way to have self-organization implement a heuristic solution to the no-free-lunch restriction by having a continual competitive trade-off between two or more complementary strategies. In biology there is evidence at many levels of analysis that evolution has centred upon this as a solution (e.g. the control and maintenance of homeostasis in the body by the autonomic nervous system which plays the sympathetic and parasympathetic systems against one another). In this way nature creates mechanisms that can strategy shift in a completely self-organizing manner and also immanent to the processing. This means that we need an account of how processing is constrained to operate in this manner. The meta-learners speciality is thus the problem of balancing a se-

ries of constraints. In the remainder of this paper we speculate about the nature of some of these constraints.

In the next three sections, we identify three important interactional problems that cognitive agents face and attempt to specify three corresponding internal economic processes that involve competition between opposing goals (see Table 1). We believe that the mechanism behind relevance realization is ultimately the process that enables the brain to balance these competing goals. Thus, *we argue that relevance is never explicitly calculated by the brain at all*, but the high-level phenomena of relevance realization emerges from the brain's attempt to dynamically balance its economic requirements. The following characterization of these economic processes is still in its infancy, but we hope that this initial sketch will inspire other researchers, and possibly lay the foundation for a much more detailed account of the mechanisms

underlying this core component of cognition.

Table #1:

Internal Economic Properties:	External Interactional Properties:
Cognitive Scope Compression $\leftrightarrow$ Particularization $\Delta w_{ij} = -\eta \frac{\partial J(\cdot)}{\partial w_{ij}} - \alpha w_{ij}$	**Applicability** General Purpose $\leftrightarrow$ Special Purpose
Cognitive Tempering TD Learning $\leftrightarrow$ Inhibition of Return $V(s_t) \leftarrow V(s_t) + \alpha \left[r_{t+1} \gamma V(s_{t+1}) - V(s_t) \right]$	**Projectability** Exploiting $\leftrightarrow$ Exploring
Cognitive Prioratization Cost function #1 $\leftrightarrow$ Cost function #2 $J_{1,2}(\beta, \alpha) = \frac{1}{\beta} J_1(\alpha, \cdot) + J_2(\alpha, \cdot)$	**Flexible Gambling** Focusing $\leftrightarrow$ Diversifying

The applicability problem: Cognitive Scope.

Let's say that you are trying to engineer processes to control your cognitive economy in order to maximize some sort of future-discounted reward. One of the engineering problems which confront you at the interactional level is whether to go for general purpose or special purpose machinery. The question is what internal economic processing can you use to optimize your *applicability* as a machine in the world. The point - stemming from optimal control and the no-free-lunch theorems - seems to be that hard commitment to either of these strategies (general purpose or specific) is undesirable. The engineer-

ing trick is to make use of opponent processing so that one is continuously dynamically designing and re-designing the right kind of tool for the task at hand (i.e. the sort of tool which will allow you to maximize reward signals for the task). How though does one get *internal economic* processing to trade off between these interactional styles of processing (general purpose versus special purpose)?

One formal example is typified by an engineering solution for training neural networks. When updating the weights in a neural network, one typically follows the negative of the gradient (or estimate of the gradient) of some performance metric, $J(\cdot)$ with respect to the weights, e.g.:

$$\Delta w_{ij} = -\eta \frac{\partial J(\cdot)}{\partial w_{ij}} \tag{1}$$

where, η, is a learning rate and, w_{ij} is a synaptic weight in the network which connects neuron i with neuron j. If the weight update term consists solely of this negative gradient term, networks will tend to over-fit the data that it trains on. A common solution to this problem is to introduce a 'weight decay', or regularization term into the weight update so that,

$$\Delta w_{ij} = -\eta \frac{\partial J(\cdot)}{\partial w_{ij}} - \alpha w_{ij} \tag{2}$$

where, α, is the magnitude of weight decay applied. Such a term penalizes large weights and pushes them towards zero [22]. This has the effect of simplifying the network which in turn makes it better at generalizing from

its training experience. The ratio of the gradient and weight decay terms will determine the extent to which a network focuses on generalization at the expense of performance on data it has trained with.

A network which is making use of weight decay is opponent processing compression against particularization (i.e. explaining the data well). This opponent processing we refer to as *cognitive scope* - trying to capture both the spatial analogy of general versus specific and the perceptual analogy of effectiveness of perception as in "microscope" and "telescope". Our claim is that a brain constrained to internally processing cognitive scope tracks the opponent processing between general purpose and special purpose machines and thereby optimizes it's applicability of information for action within the world.

Here we are broadening Turing's insight, that you can track rationality in the world by having a device that just internally pays attention to the logical syntax of the information. However, given what we've said about rationality, we think that Turing's conception of rationality as just the logical management of inference is insufficient. A lot of rationality has to do with the realization of relevance [13]. Our point is that we can track this aspect of behavioural relevance realization by internally processing cognitive scope. This extension makes use of Hinton's insight, which we term internalization (in contrast to representation), and which is typified in such methods as the wake/sleep learning algorithm [23]. In such learning algorithms the way one gets neural networks to engage in unsupervised learning, i.e. learning

where the target value in the world is not known, is by having the network relate to itself in a completely internal fashion. It relates to itself in an internal fashion such that it treats itself as a micro-environment[2] in which the procedural abilities to track environmental variables can be trained. These procedural abilities are then turned on the world in an attempt obtain the actual information from the world which can then be used to improve the micro-environment so that the whole system bootstraps itself up into effective interaction with the world. This internalized training of cognitive scope is what we mean by saying that the internal processing of economic variables, such as cognitive scope, track the interactional properties such as determining applicability.

So, part of what constitutes relevance realization is when cognitive scope tracks applicability so that the cognitive organism is continually re-designing itself as some trade-off between a general purpose and a special purpose machine. Note also, how we have specified a general constraint on information processing that is completely immanent to such processing.

[2]The original internal environment is nothing more than statistical patterns in the data stored within a neural network. However, such networks are very good at picking up very complex statistical patterns that pick up correlational and causal patterns. These complex structures within the data can serve as a virtual world upon which to train procedural abilities. It is important to remember that at first the structures of this internal "world" need not be an accurate representation of the world. Internalization only needs a demanding informational environment in which to train the skills for picking up complex information from complex environments. Of course, as an internal environment the target value is known to the system and can be used in correction. The rate of practice can be adjusted, and variation can also be introduced, all to improve the procedural learning.

The projectability problem: Cognitive Tempering.

Another engineering problem that faces individuals in the world is whether to go for exploitative machinery or exploratory machinery (see Table 1). An exploitative machine is one that tends to stick with it's currently known actions and to simply select from those actions the one that it thinks will have the highest payoff now. In contrast, the exploratory machine will forgo immediately available payoffs to look for the possibility of contexts/actions which will produce higher payoffs later on.

Neither one of these strategies pursued exclusively is a good overall problem solving strategy. The problem with being exclusively exploitative is that there is a good chance that there are much higher payoffs elsewhere and this therefore constitutes a significant opportunity cost for the explorer. You do not want to have a machine that only looks for information which facilitates exploitation in the immediate future since other information may help build towards larger future payoffs.

On the other hand, a machine which spends almost all of its time exploring for opportunity runs the risk of losing available payoff as it searches for better payoff. In a similar fashion to the opponent processing mentioned for general-purpose versus special purpose, we would want a machine that is dynamically moving between these strategies. We want a system that is optimizing for the interactional property that we call *projectability*, which is the dynamic balance between exploiting the here-and-now and exploring the there-and-

then. The system has to setup a projective relationship between the actual here-and-now with the possible there-and-thens. How then do we construct a machine which is able to optimize for projectability solely in terms of internal economic properties?

One promising way of doing this is to couple powerful reinforcement algorithms such eligibility trace temporal-difference (TD) learning [50] with an temporally decaying inhibition of return trace. In TD learning (Table 1 shows the value function update equation for the most basic form of TD learning as discussed by Sutton & Barto (1998)), a memory trace (e.g. an eligibility trace) of recent actions performed by a machine may kept. When reward (or punishment) is encountered, the actions in the trace are credited with having brought about the reward (or punishment), and so are reinforced (or weakened). Those actions which were performed most recently are reinforced most strongly, while those which happened long ago are reinforced only a little. This technique of apportioning reward to actions backward in time allows an agent to learn to perform actions for payoffs in the distant future.

TD learning reinforces action patterns (i.e. reinforcement of return) which tend to have you return to states/actions temporally associated with high reward (this is the element which pushes you to be exploitative). On the other hand, to explore in an intelligent fashion, the machine can lay down an inhibition of return trace a temporally decaying memory trace which indicates to the machine not to return (in a soft way) to states/actions it

has recently seen and thereby promotes exploration [27]. This trace would be traded off against the reinforced action patterns to find an intelligent way to trade off exploitation with exploration.

This internal processing of the opposition between inhibition of return and reinforcement of return we call *cognitive tempering*. We call it cognitive tempering because we are trying to capture the metallurgic sense of neither being too flexible nor too inflexible, and also the root word "temp" - having to do with time. What we are proposing is that cognitive tempering can be trained to track projectability. A system which is constrained for working out cognitive tempering will find good projectability in the world.

The problem of flexibly gambling: Cognitive Prioritization.

A third interactional problem faced by cognitive agents is how to gamble flexibly in the face of ambiguous information (see Table 1). Ambiguity, whether caused by the introduction of noise (e.g. perceptual) or in overlap of the entities in the environment which can generate the same information, means that one is always gambling with the commitment of one's cognitive resources. Then the interactional issue is how one is to wager one's cognitive resources in the world. Betting should be flexible because, for instance, the scarcity of one's internal reserves ought to cause a trade between focusing and diversifying as betting strategies. For example, if an agent is very thirsty it will tend to gamble all of it's efforts on getting water, i.e. it will focus it's wagers on this project. As soon as thirst is satiated though, the agent will

begin to pursue a diversity of problems. As we will discuss in the following section, Montague points out that agents ought to "care" differentially about the environment because they run on batteries [31].

What are the internal economic constraints that can track these interactional trade offs? In contrast to the other two economic constraints, this one is much more conjectural in nature. What we intend to do here is to provide an argument for the plausibility of being able to produce a mathematical formalization of this constraint. Whereas the internal economic constraints we call cognitive scope and cognitive tempering had to do with how cost functions might be heuristically optimized, *cognitive prioritization* has to do with the structure and prioritization of cost functions. In short, one of the things which allows agents to be truly successful in the world is adapt not only their behaviour to suite a given task, but also to adapt the sorts of *tasks* they are interested in optimizing.

In order to make clear what we mean, consider the following example: an animal has two basic operational goals in its life; one is to find food to sustain its energy stores, and the second is to avoid being food for another larger animal. All other projects are asymmetrically dependent on these ongoing projects[3]. Both of these external goals are tracked internally by two cost functions, $J_1(\cdot)$ and $J_2(\cdot)$ respectively. One of these cost functions, J_1,

[3]In a personal communication, Zachary Irving pointed out that didactically it may help to think that, instead of just having two cost functions - food and predator avoidance - you had ten (e.g. mate seeking, water seeking, sleep, young rearing etc.). If this was the case, then when beta gets very low, the a learning system looks very specifically at food seeking, at the cost of all those other activities that are potentially relevant.

is an internal metric which tracks how well your doing at maintaining your energy reserves. The second, J_2, is an internal metric which tracks how well you are doing at predator avoidance. Suppose that J_1 is multiplied by a leaky integrator function which tracks the level of satiation for the animal (it integrates acquisition of resources and leaks as they are depleted).

Now, suppose that the animal is therefore interested in what we will call a joint cost function, $J_{1,2}$:

$$J_{1,2}(\beta, \alpha) = \frac{1}{\beta} J_1(\alpha, \cdot) + J_2(\alpha, \cdot) \tag{3}$$

where, β, is the leaky integrator which indicates the level of satiation, and α is a vector containing the animal's adjustable parameters governing action, J_2 is a relatively constant cost function which resolves ambiguity in this manner: it emphasizes misses over mistakes; that is, it is much more important to not miss the predators approach than it is to mistake something for a predator that is not one. On the other hand J_1 will tend to resolve ambiguity by emphasizing mistakes over misses: that is, it is initially more important not to mistake poisonous or inedible material for food than to miss food. But, as energy resources deplete, the system gives more and more emphasis to not missing food opportunities than to mistakenly eating poison or inedibles. When satiation, β, is low (e.g. $\frac{1}{\beta}$ is large), J_1 becomes dominant and can put pressure on $J_{1,2}$ to focus resource investment into the food acquisition project.

Of course, we are not at all committed to there only being two cost functions - there are likely many which are flexibly trade off against each other. The example presented above is merely illustrative of how cognitive prioritization may operate as a constraint within relevance realization. Also, the approach we have taken here easily permits the addition and balance of other cost functions.

Interaction between the three constrains.

We think that the three internal economic constraints, cognitive scope (CS), cognitive tempering (CT), and cognitive prioritization (CP), are all mutually constraining within an internal economic arena. In addition, we think that there are higher order constraints on this process of running the internal cognitive economy. We think there is a selective constraint to be as efficient as possible in this economy. But there is an opponent constraint to be resilient in your processing. The main problem is if you just push for efficiency, you can lose a lot of latent preadaptive functions which may turn out to have long term value to you. So, you do not want to downsize too much and become brittle in your ability to handle environmental perturbations - you

want some overlap, redundancy, and variation in your processing.

Table #2:

	Efficiency (Selection)	
Compression-Generalization	TD Learning-Exploiting	-Focusing
↕	↕	↕
Particularization-Specialization	Inhibition of Return-Exploring	-Diversifying
	Resiliency (Variation)	

Resiliency introduces variation into the economy and efficiency introduces selection - they are opponent to one another and so the whole system will tend to evolve. This is very similar to Siegler's idea that cognitive development shows significant parallels to the process of evolution [42]. Thus, relevance realization is continually evolving. It is continuously self-adaptively self-designing. This is cashed out neurologically in the developmental complexification [41] (a dialectic of integration for efficiency and diversification for resilience) of the brain's functionality. Putting these ideas together implies that cognition is inherently developmental in nature, rather than development just being the peripheral issue of how cognition emerges. Thus, we may speculate that developmental psychology ought be seen as central to cognitive psychology.

The constraint of efficiency is specifically discussed by Montague [31] as being important to getting cognitive systems to "care" about information, i.e., find information relevant. According to Montague, such caring will make

it possible for cognitive systems to choose what information to pay attention to and which actions to perform. His basic argument is that because organisms run on energy reserves, what he calls "batteries," all of the cognitive processing of real world organisms is constrained to be as efficient as possible. Although he does discuss some interesting ideas about internal communication and modelling, he does not explicate how in general the constraint for efficiency is implemented in all processing.

In contrast, Sperber and Wilson [47] much more explicitly develop such an account. According to Sperber and Wilson information is relevant to the degree to which it trades off between the maximization of cognitive effect and the minimization of cognitive effort. Relevance is a kind of cognitive profit, and information is more relevant if it is more efficiently obtained, i.e. more effect for less effort. We think that there are very important insights in this approach. There is an emphasis on economic properties that are internally specified, and there is use (at least implicitly) of self-organization through opponent processing.

However, we do think that there are important problems with the attempt to equate relevance with efficiency. First is that since relevance is defined as efficiency it is not possible according to Sperber and Wilson to be inefficient in processing and realize relevance. Since it is plausible that the brain also pursues resiliency it may often processing information in a manner that is currently inefficient so that it does not loss the ability to repair, re-learn, or re-design itself in the future. We suggest rather than efficiency defining

relevance it should be thought of as a higher order constraint operating in an opponent fashion with the higher order constraint of resiliency. This opponent processing instantiates the goal of making cognition continually evolve. Naturally saying that relevance realization is cognitive evolution is irredeemably vague. In our proposal the evolution is specified in terms of the interaction of higher order constraints which are further specified in terms of the interaction of lower order constraints which are further specified in terms of opponent processing in and between cost functions operating in an economic manner. Though of course, there is still an enormous amount of work to do, both theoretically and experimentally, to build on the sketch we have given here.

The second problem for Sperber and Wilson's theory is the problem of how the efficiency is specified in terms of economic properties. Although there is great insight on Sperber and Wilson's part about the role of economic properties, there is nevertheless confusion in the formalization of these properties. Sperber and Wilson confuse the economic level with both the syntactic and semantic levels of processing. For example, they discuss relevance processing in terms of logical inferential relations which are clearly at the syntactic level, and they discuss cognitive effect in terms of belief revision which is clearly pitched at the semantic/representational level. This intrusion of the syntactic and semantic levels subjects their theory to the kinds of circularities we noted earlier. This conclusion was clearly pointed out by Chiappe and Kukla [6]. They note that for Sperber and Wilson the calculation of cognitive profit

is constrained by the information that is active in the current context. We would argue that this is a dim recognition of the importance of interactional properties like the applicability of processing that need to be further specified and incorporated significantly into the theory. In any case, Chiappe and Kukla point out that Sperber and Wilson need to explicate how the initial context is updated and developed. Sperber and Wilson do this in terms other potential contexts from memory which the organism can select. According to the theory the organism needs to select a context that maximizes relevance (understood as cognitive profit.) This is, of course, circular in that determining which context from all potential contexts will do this is the kind of problem for which relevance realization is needed. Sperber and Wilson attempt to address this by arguing that contexts (in memory) have accessibility relations to each other by means of which the benefits and costs of moving between contexts can be calculated according to cognitive profit. Note that there is an insight here into relevance realization constraints operating at different levels of analysis. However, please also note that the explore versus exploit problem has been transferred inside in that as the cognitive system moves between candidate contexts in memory it must decide if it should stop at any currently valuable context and exploit it or keep exploring for a potentially more valuable context. While this is problematic for Sperber and Wilson's explanation it does suggest a way in which memory organization could serve as an internal environment for the further internalization and training, i.e., bootstrapping, of cognitive tempering.

Yet such memory organization cannot be foundational as Chiappe and Kukla point out. As we noted earlier, memory organization itself presupposes relevance realization and so presupposes the very process it is being used to explain. Chiappe and Kukla point out that Sperber and Wilson have no account of memory organization and so the whole account is circular in nature. Sperber and Wilson [48] inadequately address this by arguing that these foundational problems are probably addressed by some combination of modularity (special purpose machinery) and something like a blind hill-climbing algorithm (general purpose machinery). There is a lot of confusion in this answer between interactional properties and economic properties, and about how these interactional properties are organized in processing, and how this processing is realized in a completely internal economic fashion. There is no clear discussion of how relevance realization could develop and bootstrap itself up into sophisticated relevance realization. All of these things need to be carefully teased apart and the relations between them worked out if the charge of circularity is going to be dissolved. We believe our current theory begins to do just this, although much more work needs to be done.

Conclusion

We have argued that relevance realization is a pervasive problem within cognitive science and a new framework for doing cognitive science is emerging in which relevance realization is the criterion of the cognitive. As such, we believe that the explication and explanation of cognition will ultimately be

in terms of processes of relevance realization. Further, we have argued that this framework is beginning to discover and grapple with the theoretical and technical tools required to address questions concerning the mechanisms of relevance realization in the brain. We have sketched what we believe are the crucial points in such an explanation of relevance realization. Although we are no doubt wrong in detail, we believe that we have shown that it is very plausible that the correct answer will be turn out to be relevantly similar to the one we have presented here.

Acknowledgements

We would like to acknowledge the help of Zachary C. Irving, Najam Tirmizi, Leo Ferraro, Vladislav Sekulic, and Alex Lamey.

References

[1] J.L. Austin. *How to do things with words*. Oxford University Press, Oxford, England, 1962.

[2] Lawrence Barsalou. *Situated conceptualization*, chapter 28, pages 619–650. Elsevier, Amsterdam, 2005.

[3] Yoshua Bengio and Yann LeCun. *Large-Scale Kernel Machines*, chapter Scaling learning algorithms towards AI, pages 321–358. MIT Press, 2007.

[4] C. Cherniak. *Minimal Rationality*. MIT Press, Cambridge, MA, 1986.

[5] D. Chiappe and J. Vervaeke. Fodor, cherniak, and the naturalization of rationality. *Theory and Psychology*, 7(6):799–821, 1997.

[6] D.L. Chiappe and A. Kukla. Context selection and the frame problem. *Behavioral and Brain Sciences*, 19(3):529–530, 1996.

[7] Noam Chomsky. *The Minimalist Program*. The MIT Press, Cambridge, MA, 3rd edition, 1997.

[8] D. Dennett. *Cognitive Wheels: The Frame Problem of AI*, chapter The Robot's Dilema: The Frame Problem in Artificial Intelligence. Greenwood Publishing Group Inc., Westport, CT. USA, 1987.

[9] K. Doya, K. Samejima, K. Katagiri, and M. Kawato. Multiple model-based reinforcement learning. *Neural Computation*, 14:1347–1369, 2002.

[10] Kenji Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245, 2000.

[11] J. Fodor. *Psychological Explanation: An Introduction to the Philosophy of Psychology*. Random House, New York, 1968.

[12] J. Fodor. How the mind works: what we still don't know. *Daedalus*, (Summer):86–94, 2006.

[13] M French, Robert. Subcognition and the limits of the turing test. *Mind*, 99(393):53–65, 1990.

[14] P. Gärdenfors. *Conceptual Spaces: The geometry of thought*. MIT Press, Cambridge, MA, 2000.

[15] Lila R. Gleitman and Mark Liberman. *An Invitation to Cognitive Science: Volume 1: Language*. MIT Press, Cambridge, 2nd edition, 1995.

[16] N. Goodman. *"Seven strictures on similarity"*. Bobbs-Merrill, New York, 1972.

[17] C. Green. Fodor, functions, physics and fantasyland: Is ai a mickey mouse discipline? *Journal of Experimental Theoretical Artificial Intellilgence*, (8):95–106, 1996.

[18] Christopher D. Green and John Vervaeke. *What kind of an explanation, if any, is a connectionist net?*, pages 201–208. Captus, North York, ON, 1996.

[19] H.P. Grice. *Studies in the way of words*. Harvard University Press, Cambridge MA, 1989.

[20] Gilbert Harman. *Change in View: Principles of Reasoning*. MIT Press, Cambridge, 1988.

[21] J. Haugeland. *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge, MA, 1985.

[22] G. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, pages 5–13. ACM Press, New York, NY, 1993.

[23] G.E. Hinton, P. Dayan, Brendan J. Frey, and Radford M. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

[24] G.E. Hinton, S. Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

[25] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

[26] Keith J. Holyoak. *Problem Solving*, chapter 8. MIT Press, Cambridge, 2nd edition, 1995.

[27] R.M Klein and W. MacInnes. Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, 10(4):346–362, 1999.

[28] Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. *1st International Conference on Informatics in Control, Automation and Robotics*, 1:222–229, 2004.

[29] Douglas Medin. Concepts and conceptual structures. *American Psychologist*, 44:1469–1481, 1989.

[30] Douglas Medin and Lance Rips. *Concepts and Categories: Memory, Meaning and Metaphysics*, chapter 3. Cambridge University Press, New York, 2005.

[31] R. Montague. *Your Brain is (Almost) Perfect*. Plume, New York, USA, 2006.

[32] A. Newell and H. Simon. *Human problem solving*. Princeton University, Inglewood Cliffs, NJ, 1972.

[33] Allen Newell and Herbert A. Simon. *Computer Science as Empirical Inquiry: Symbols and Search*, chapter 4. MIT Press, Cambridge, MA, 1997.

[34] H. Putnam. *Representation and Reality*. MIT Press, Cambridge, MA., 1991.

[35] Zenon Pylyshyn. *Computation and Cognition: Toward a foundation for cognitive science.* MIT Press, Cambridge, 1984.

[36] Z.W. Pylyshyn. *The robot's dilemma: The frame problem in artificial intelligence.* Ablex Publishing Corporation, NJ, 1988.

[37] W. Ramsey, S.P. Stich, and J. Garon. *Connectionism, eliminativism, and the future of folk psychology*, pages 199–228. Lawrence Erlbaum, Hillsdale, NJ, 1991.

[38] D.E. Rumelhart and A. Ortony. *The representation of knowledge in memory.* Lawrence Erlbaum Associates Inc., Hillsdale, N.J., 1977.

[39] R.C. Schank and R.P. Abelson. *Scripts, plans, goals and understanding.* Lawrence Erlbaum Associates Inc., Hillsdale, N.J., 1977.

[40] Yury P. Shimansky, Tao Kang, and Jiping He. A novel model of motor learning capable of developing an optimal movement control law online from scratch. *Biological Cybernetics*, 90:133–145, 2004.

[41] D.J. Siegel. Toward and interpersonal neurobiology of the developing mind:attachment relationships, "mindsight" and neural integration. *Infant Mental Health Journal*, 22:67–94, 2001.

[42] Robert S. Siegler and Christopher Shipley. *Variation, Selection, and Cognitive Change*, chapter 2, pages 31–76. Erlbaum, Hillsdale, NJ, 1995.

[43] Herbert A. Simon. *Reasoning in Human Affairs*. Standford University Press, San Francisco, 1983.

[44] Edward Smith and Daniel Osherson. *An Invitation to Cognitive Science: Volume 3: Thinking*. MIT Press, Cambridge, 2nd edition, 1995.

[45] P. Smolensky. On the proper treatment of connectionism. *Behavioural and Brain Sciences*, 11:1–73, 1988.

[46] P. Smolensky. *Connectionism, constituency, and the language of thought*, pages 201–227. Blackwell, Oxford, 1991.

[47] D. Sperber and D. Wilson. *Relevance: Communication and Cogniton*. Blackwell, Oxford, 2nd edition, 1995.

[48] D. Sperber and D. Wilson. Fodor's frame problem and relevance theory. *Behavioral and Brain Sciences*, 19(3):530–532, 1996.

[49] Stephen Stich. *The Fragmentation of Reason: Preface to a Pragmatic Theory of cognitive evaluation*. MIT Press, Cambridge, 1993.

[50] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. The MIT Press, Cambridge, Massachusetts, 1998.

[51] Evan Thompson. *Mind in life: biology, phenomenology, and the sciences of the mind*. Belknap Publishing, London, 2007.

[52] Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 16(11):1226–1235, November 2002.

[53] Alan M Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

[54] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.

[55] Vladimir N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[56] Francisco Varela, Evan T. Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive science and human experience*. The MIT Press, Cambridge, MA, 1991.

[57] J. Vervaeke. *The Naturalistic Imperative in Cognitive Science*. University of Toronto, Toronto, Canada, ph.d. thesis edition, 1997.

[58] R. Vilata and Y. Drissi. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, (18):77–95, 2002.

[59] M. Wheeler. *Reconstructing the Cognitive World*. MIT Press, Cambridge, MA., 2005.

[60] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transaction on Evolutionary Computation*, 1(1):67–82, 1997.