ABDUCTIVE INFERENCE:
FROM PHILOSOPHICAL ANALYSIS TO NEURAL MECHANISMS

Paul Thagard
University of Waterloo

Thagard, P. (forthcoming). Abductive inference: From philosophical analysis to neural mechanisms. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Cognitive, mathematical, and neuroscientific approaches*. Cambridge: Cambridge University Press.

1. WHAT IS ABDUCTION?

In the 1890s, the great American philosopher C. S. Peirce (1931-1958) used the term "abduction" to refer to a kind of inference that involves the generation and evaluation of explanatory hypotheses. This term is much less familiar today than "deduction", which applies to inference from premises to a conclusion that has to be true if the premises are true. And it is much less familiar than "induction", which sometimes refers broadly to any kind of inference that introduces uncertainty, and sometimes refers narrowly to inference from examples to rules, which I will call "inductive generalization". Abduction is clearly a kind of induction in the broad sense, in that the generation of explanatory hypotheses is fraught with uncertainty. For example, if the sky suddenly turns dark outside my window, I may hypothesize that there is a solar eclipse, but many other explanations are possible, such as the arrival of an intense storm or even a huge spaceship.

Despite its inherent riskiness, abductive inference is an essential part of human mental life. When scientists produce theories that explain their data, they are engaging in abductive inference. For example, psychological theories about mental representations and processing are the result of abductions spurred by the need to explain the results of psychological experiments. In everyday life, abductive inference is ubiquitous, for example when people generate hypotheses to explain the behavior of others, as when I

infer that my son is in a bad mood to explain a curt response to a question. Detectives perform abductions routinely in order to make sense of the evidence left by criminal activity, just as automobile mechanics try to figure out what problems are responsible for a breakdown. Physicians practice abduction when they try to figure out what diseases might explain a patient's symptoms. Table 1 summarizes the kinds of abductive inference that occur in various domains, involving both targets that require explanation and the hypotheses that are generated to explain them. Abduction occurs in many other domains as well, for example religion where people hypothesize the existence of God in order to explain the design and existence of the world.

DOMAINS	TARGETS TO BE EXPLAINED	EXPLANATORY HYPOTHESES
science	experimental results	theories about structures and processes
medicine	symptoms	diseases
crime	evidence	culprits, motives
machines	operation, breakdowns	parts, interactions, flaws
social	behavior	mental states, traits

Table 1. Abductive inference in five domains, specifying what needs to be explained and the kinds of hypotheses that provide explanations.

The next section will briefly review the history of the investigation of abduction by philosophers and artificial intelligence researchers, and discuss its relative neglect by psychologists. First, however, I want to examine the nature of abduction and sketch what would be required for a full psychological theory of it. I then outline a neurocomputational theory of abductive inference that provides an account of some of the neural processes that enable minds to make abductive inference. Finally, I discuss the

more general implications of replacing logic-based philosophical analyses of human inference with theories of neural mechanisms.

Here are the typical stages in the mental process of abduction. First, we notice something puzzling that prompts us to generate an explanation. It would be pointless to waste mental resources on something ordinary or expected. For example, when my friends greet me with the normal "Hi", I do not react like the proverbial psychoanalyst who wondered "What can they mean by that?" In contrast, if a normally convivial friend responds to "Good morning" with "What's so good about it?", I will be prompted to wonder what is currently going on in my friend's life that might explain this negativity. Peirce noticed that abduction begins with puzzlement, but subsequent philosophers have ignored the fact that the initiation of this kind of inference is inherently emotional. Intense reactions such as surprise and astonishment are particularly strong spurs to abductive inference. Hence the emotional initiation of abductive inference needs to be part of any psychological or neurological theory of how it works. An event or general occurrence only becomes a target for explanation when it is sufficiently interesting and baffling. I know of no general experimental evidence for this claim, but Kunda, Miller, and Claire (1990) found that surprise triggered causal reasoning in cases of conceptual combination.

Second, the mind searches for possible hypotheses that could explain the target. Sometimes, the search is easily completed when there is a prepackaged hypothesis waiting to be applied. For example, if you know that your friend Alice gets stressed out whenever she has a deadline to meet, you might explain her grumpy behavior by the conjecture that she has a project due. In more deeply puzzling cases, the search for an

explanatory hypothesis may require a much longer search through memory, or even the use of analogy or other constructive processes to generate a highly novel hypothesis. This generation is what happens in science when a genuinely new theory needs to be developed.

Generation of a candidate explanatory hypothesis is the usual third stage of abductive inference. If one is epistemically lazy, abductive inference may end with the generation of a single candidate. But scientists and careful thinkers in general are aware of the perils of abductive inference, in particular that one should not accept an explanatory hypothesis unless it has been assessed with respect to competing hypotheses and all the available evidence. Philosophers call this fourth, evaluative stage of abductive reasoning *inference to the best explanation* (Harman, 1973; Thagard, 1988; Lipton, 2004). Ideally, the reasoner correctly decides that it is legitimate to infer a hypothesis because it really is the best explanation of all the available evidence. Thus generation of an explanatory hypothesis blends into its evaluation.

Just as abduction originates with an emotional reaction, it ends with one, because formation and acceptance of explanatory hypotheses usually produce positive emotions. Gopnik's (1998) comparison of explanations with orgasms is exaggerated, but it is nevertheless important that finding an explanation for something puzzling is often very satisfying. Hence we can mark emotional satisfaction as the final stage of abductive inference, as shown in Figure 1. This diagram will be flushed out substantially in later sections in terms of neurological processes. The result will be an account of abduction that goes far beyond the philosophical account that takes abduction to be a kind of inference of the form: q, if p then q, so maybe p. (Somebody once called this "modus

morons".) To foreshadow, the main differences include not only the crucial involvement of emotion, but also the allowance that both targets and hypotheses can be multimodal rather than purely verbal representations. Moreover, I will contend that the relation between a target and an explanatory hypothesis is that the target phenomenon is caused by the factors invoked by the hypothesis, and that people's understanding of causality is inherently non-verbal because it is rooted in visual and kinesthetic perception. Hence abduction, instead of looking like a feeble-minded cousin of the deductive principle modus ponens, is actually a far richer and more powerful form of thinking.

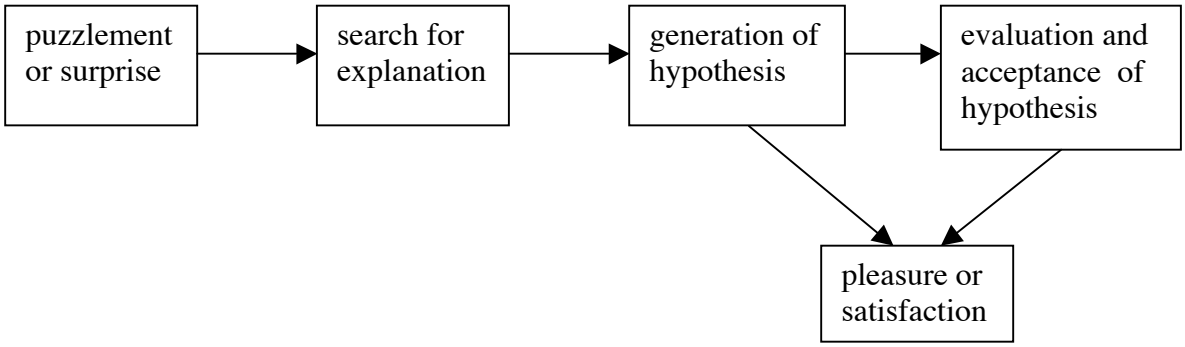

Figure 1. The process of abductive inference.

2. ABDUCTION IN PHILOSOPHY, ARTIFICIAL INTELLIGENCE, AND PSYCHOLOGY

Although the term "abduction" only emerged in the nineteenth century, philosophers and scientists have been aware of inference to explanatory hypotheses at least since the Renaissance (Blake, Ducasse, and Madden, 1960). Some thinkers have been skeptical that a hypothesis should be accepted *merely* on the basis of what it explains, for example Isaac Newton, John Herschel, August Comte, and John Stuart Mill.

But others, such as David Hartley, Joseph Priestley, and William Whewell have argued that such inference is a legitimate part of scientific reasoning. In the twentieth century, there are still philosophical skeptics about abduction (e.g van Fraassen, 1980), but many others contend that abduction, construed as inference to the best explanation, is an essential part of scientific and everyday reasoning (Magnani, 2001; Psillos, 1999; Thagard, 1992, 1999, 2000).

The generation of explanatory hypotheses has also interested philosophers concerned with how scientific discoveries are made. Hanson (1958) tried to develop Peircean ideas into a "logic" of discovery. More recently, Darden (1991), Magnani (2001), and Thagard (1988) have examined cognitive processes that are capable of producing new hypotheses, including scientific theories. The work of Shelley (1996) and Magnani (2001) shows how abduction can involve visual as well as verbal representations, which is important for the multimodal theory developed below.

In the field of artificial intelligence, the term "abduction" is usually applied to the evaluation of explanatory hypotheses, although it sometimes also includes processes of generating them (Charniak and McDermott, 1985; Josephson and Josephson, 1994). AI models of abductive inference have primarily been concerned with medical reasoning. For example, the RED system takes as input descriptions of cells and generates and evaluates hypotheses about clinically significant antibodies found in the cells (Josephson and Josephson, 1994). More recently, discussions of causal reasoning in terms of abduction have been eclipsed by discussions of Bayesian networks based on probability theory, but later I will describe limitations of purely probabilistic accounts of causality, explanation, and abduction. Abduction has also been a topic of interest for researchers

in logic programming (Flach and Kakas, 2000), but there are severe limitations to a characterization of abduction in terms of formal logic (Thagard and Shelley, 1997).

Some AI researchers have discussed the problem of generating explanatory hypotheses without using the term "abduction". The computational models of scientific discovery described by Langley, Simon, Bradshaw, and Zytkov (1987) are primarily concerned with the inductive generalization of laws from data, but they also discuss the generation of explanatory structure models in chemistry. Langley et al. (2004) describe an algorithm for "inducing explanatory process models" , but it is clear that their computational procedures for constructing models of biological mechanisms operate abductively rather than via inductive generalization.

Psychologists rarely use the terms "abduction" or "abductive inference", and very little experimental research has been done on the generation and acceptance of explanatory hypotheses. Much of the psychological literature on induction concerns a rather esoteric pattern of reasoning, categorical induction, in which people express a degree of confidence that a category has a predicate after being told that a related category has the predicate (Sloman and Lagnado, 2005). Here is an example:

Tigers have 38 chromosomes.

Do buffaloes have 38 chromosomes?

Another line of research involves inductive generalizations about the behavior of physical devices (Klahr, 2000). Dunbar (1997) has discussed the role of analogy and other kinds of reasoning in scientific thinking in real-world laboratories. Considerable research has investigated ways in which people's inductive inferences deviate from normative standards of probability theory (Gilovich, Griffin, and Kahneman (2002).

Experimental research concerning causality has been concerned with topics different from the generation of causal explanations, such as how people distinguish genuine causes from spurious ones (Lien and Cheng, 2000) and how knowledge of the causal structure of categories supports the ability to infer the presence of unobserved features (Rehder and Burnett, 2005). Social psychologists have investigated the important abductive task of *attribution,* in which people generate explanations for the behavior of others (Kunda, 1999; Nisbett and Ross, 1980). Read and Marcus-Newhall tested the applicability of Thagard's (1992) computational theory of explanatory coherence to the evaluation of social explanations. Generally, however, psychologists have had little to say about the mental mechanisms by which new hypotheses are formed and evaluated. My review of interdisciplinary research on abductive inference has been very brief, for I want to move on to develop a new neurocomputational theory of abduction.

3. NEURAL STRUCTURES

The structure of abduction is roughly this:

There is a puzzling target T that needs explanation.

Hypothesis H potentially explains T.

So, H is plausible.

H is a better explanation of T and other phenomena than competing hypotheses.

So H is acceptable.

It would be psychologically unrealistic, however, to assume, as philosophers and AI researchers have tended to do, that T and H must be sentences or propositions (the meanings of sentences). A broader view of mental representation is required.

As already mentioned, abductive inference can be visual as well as verbal (Shelley, 1996; Magnani, 2001). For example, when I see a scratch along the side of my car, I can generate the mental image of grocery cart sliding into the car and producing the scratch. In this case both the target (the scratch) and the hypothesis (the collision) are visually represented. Other sensory modalities can also provide explanation targets. For example, in medical diagnosis the perception of a patient's symptoms can involve vision (rash), touch (swelling), sound (heart murmur), smell (infection), and even taste (salty, in patients with cystic fibrosis). An observant cook may similarly be prompted to generate hypotheses by various kinds of sensory experiments, asking such questions as "Why does the cheese have blue stuff on it?" (vision), "Why is the broccoli soggy?" (touch), "Why is the timer buzzing?" (hearing), "Why is the meat putrid?" (smell), and "Why is the soup so salty" (taste). Thus all of the senses can generate explanation targets that can initiate abductive inference.

It is an interesting question whether hypotheses can be represented using all sensory modalities. For vision the answer is obvious, as images and diagrams can clearly be used to represent events and structures that have causal effects. And the answer appears to be yes when one is explaining one's own behavior: I may recoil because something I touch feels slimy, or jump because of a loud noise, or frown because of a rotten smell, or gag because something tastes too salty. Hence in explaining my own behavior my mental image of the full range of examples of sensory experiences may have causal significance. Applying such explanations of the behavior of others requires projecting onto them the possession of sensory experiences that I think are like the ones that I have in similar situations. For example, when I see people wrinkle up their noses

in front of a garbage can, I may project onto them an experience similar to what I experience when I smell rotting garbage. In this case, my image of a smell is the olfactory representation of what I see as the cause of their behavior. Empathy works the same way, when I explain people's behavior in a particular situation by inferring that they are having the same kind of emotional experience that I have had in similar situations. For example, if a colleague with a recently rejected manuscript is frowning, I may empathize by remembering how annoyed I felt when a manuscript of mine was rejected, and my mental image projected onto the colleague constitutes a non-verbal representation that explains the frown. Of course, I may operate with verbal explanations as well, but these complement the empathetic ones. Hence there is reason to believe that abductive inference can be fully multimodal, in that both targets and hypotheses can have the full range of verbal and sensory representations. In addition to words, sights, sounds, smells, touches, and tastes, these can include emotional feelings, kinesthetic experiences such as feeling the body arranged in a certain way, and other feelings such as pain.

A narrowly verbal account of abduction such as those favored by logicians would clearly not be able to do justice to the multimodal character of abductive inference. But from a neurological perspective, there is no problem in postulating representations that operate in all relevant modalities. Let me define a neural structure as a complex <neurons, connections, spiking behaviors> that consists of a set of neurons, a set of synaptic connections among them, and a set of behaviors of individual neurons that specifies their patterns of spiking determined by the spiking behaviors of all those neurons to which they are connected. If the neurons are thought of as a dynamical system governed by a set of differential equations, then the spiking behaviors can be

thought of as the state space of the system. In contrast to standard connectionist models of neural networks, it is important to specify the behavior of neurons as more than just patterns of activation, because there is evidence that spiking patterns can be both neurologically and computationally important (Eliasmith and Anderson, 2003; Maass and Bishop, 1999; Rieke et al., 1997).

On the plausible assumption that all mental representations are brain structures, we can conjecture that verbal and sensory representations are neural structures of the sort just described. Hence we can reconceptualize abduction neurologically as a process in which one neural structure representing the explanatory target generates another neural structure that constitutes a hypothesis. Two major problems need to be solved in order to construct a neurological model of abduction consistent with the flow chart presented earlier in figure 1. The first is how to characterize the emotional inputs and outputs to the abductive process, that is how to mark the target as puzzling and the hypothesis as satisfying. The second is how to represent the explanatory relation between the neurally represented target and hypothesis.

The first problem can be dealt with by supposing that emotions are also neural structures in the sense just defined. I do not mean to suggest that for each emotion there is a constrained set of neurons that encodes it. Emotions involve complex interaction among sensory processes involving bodily states, and cognitive processes involving appraisal of a person's situation (see Thagard, 2005, ch. 10). Hence the neural structure corresponding to an emotional experience is not a neuronal group situated in a particular brain area, but a complex of neurons distributed across multiple brain areas.

Now the question becomes: what is the relation between the neural structure for the explanation target and the neural structure for an emotion such as puzzlement? There are two possibly compatible answers involving different aspects of neural structure: connections and spiking behavior. The target neural structure may be interrelated with an emotional neural structure because some of the neurons in the first structure have direct or indirect synaptic connections with the second structure. Hence the two neural structures are part of a larger neural structure. Moreover, these interconnections may establish temporal coordination between the two neural structures, so that the spiking behavior of the target neural structure is synchronized or approximately coordinated with the spiking behavior of emotional neural structure. Through one or both of these means – physical connectivity and temporal behavior – the brain manages to mark the target as puzzling and in need of explanation. Similarly, the hypothesis can be represented by a neural structure which operates in any verbal or sensory modality, and which can be associated with a neural structure corresponding to the emotional experience of satisfaction or pleasure. Thus part of the flow chart in figure 1 translates into the following neurological diagram, figure 2. What remains to be investigated is the relation of explanation and inference marked by the thin arrow. We need an account of explanation as a neurological process.

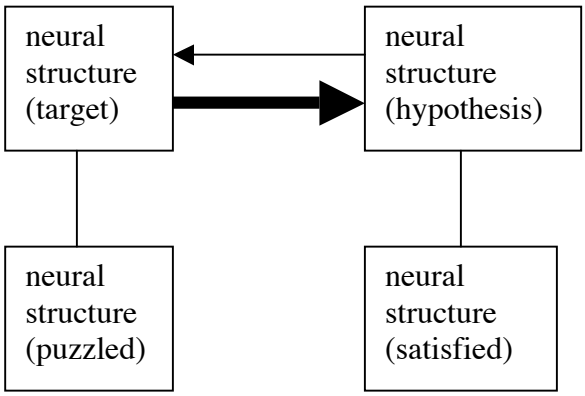

Figure 2. Abduction as a neural process. The plain lines indicate associations based on synaptic connectivity and temporal coordination, and the thin arrow indicates an explanatory relation to be clarified in the next section. The thick arrow indicates an inferential relation to be clarified in section 5.

4. EXPLANATION AND CAUSALITY

Abductive inference is from a target to a hypothesis that explains it, but what is explanation? In philosophy and cognitive science, there have been at least six approaches to the topic of explanation (Thagard, 1992, pp. 118ff.). Explanations have been viewed as deductive arguments, statistical relations, schema applications, analogical comparisons, causal relations, and linguistic acts. All of these approaches have illuminated some aspects of the practice of explanation in science and everyday life, but the core of explanation, I would argue, is causality: a hypothesis explains a target if it provides a causal account of what needs to be explained. In medicine, a disease explains symptoms because the abnormal biological state that constitutes the disease produces the symptoms through biological mechanisms (Thagard, 1999). In ordinary social life, attributing a mental state such as an emotion explains a person's behavior because the

mental state is the assumed cause of the behavior. In science, a theory explains phenomena such as the motions of the planets by providing the causes of such phenomena. Deductive, statistical, schematic, analogical, and linguistic aspects of explanation can all be shown to be subordinate to the fundamental causal aspect of explanation.

Hence the problem of describing the explanatory relation between hypotheses and targets is largely the problem of describing the causal relation between what neural structures for hypotheses represent and what neural structures for targets represent. To put it most simply, abduction becomes: T, H causes T, so maybe H, where H and T are not propositions but what is represented by neural structures. Hence it is crucial to give a philosophically and psychologically plausible account of causality.

The philosophical literature on causality is huge, but here is a quick summary of extant positions on the nature of causality:

1. Eliminativist: Causality is an outmoded notion no longer needed for scientific discourse.

2. Universalist: Causation is a relation of constant conjunction: the effect always occur when the cause occurs.

3. Probabilistic: Causation is a matter of probability: the effect is more probable given the cause than otherwise.

4. Causal powers: the cause has a power to produce the effect.

Each of these positions is problematic. The eliminativist position runs afoul of the fact that talk of causal mechanisms still abounds in science. The universalist position is untenable because most causal relations are statistical rather than constant: infection by a

mycobacterium causes tuberculosis, but many people infected by it never develop the disease. Similarly, the probabilistic position cannot easily distinguish between cases where a cause actually makes an effect more likely, as in infection by the mycobacterium, and cases where the effect is made more likely by some other cause. For example, the probability that people have tuberculosis given that they take the drug Isoniazid is much greater than the probability that they have tuberculosis, but this is because Isoniazid is a commonly prescribed treatment for the disease, not because it causes the disease. (See Pearl, 2000, for a broad and deep discussion of causality and probability.) People have an intuitive sense of the difference between causal and purely statistical relations. In part, this arises from an understanding of how mechanisms connect causes with their effects: see section 7 below. But it also arises from a natural perceptual inclination to see certain kinds of occurrences as causally related to each other. This inclination does not depend on the postulation of occult causal powers that causes must have in relation to their effects, but on fundamental features of our perceptual systems.

Evidence that causal relations can be perceived comes from three kinds of psychological evidence: cognitive, developmental, and neurological. Michotte (1963) performed a large array of experiments with visual stimuli that suggest that adults have a direct impression of cause-effect relations: when people see an image of one ball moving into another that begins to move, the first ball is perceived to cause the second one to move. Such reactions are even found in young infants: Leslie and Keeble (1987) provided experimental evidence that even 27-week old infants perceive a causal relationship. Baillargeon, Kotovsky, and Needham (1995) report that infants as young as 2.5 months expect a stationery object to be displaced when it is hit by a moving object;

by around 6 months, infants believe that the distance traveled by the stationery object is proportional to the size of the moving object. Thus at a very primitive stage of verbal development children seem to have some understanding of causality based on their visual and tactile experiences. According to Mandler (2004), infants' very early ability to perceive causal relations need not be innate, but could arise from a more general ability to extract meaning from perceptual relationships. Whether or not it is innate, infants clearly have an ability to extract causal information that develops long before any verbal ability.

Recent work using functional magnetic resonance imaging has investigated brain mechanisms underlying perceptual causality (Fugelsang et al., 2005). Participants imaged while viewing causal events had higher levels of relative activation in the right middle frontal gyrus and the right inferior parietal lobule compared to those viewing non-causal events. The evidence that specific brain structures are involved in extracting causal structure from the world fits well with cognitive and developmental evidence that adults and children are able to perceive causal relations, without making inferences based on universality, probability, or causal powers. It is therefore plausible that people's intuitive grasp of causality, which enables them to understand the distinction between causal relations and mere co-occurrence, arises very early from perceptual experience. Of course, as people acquire more knowledge, they are able to expand this understanding of causality far beyond perception, enabling them to infer that invisible germs cause disease symptoms. But this extended understanding of causality is still based on the perceptual experience of one event making another happen, and does not depend on a mysterious, metaphysical conception of objects possessing causal powers.

For discussion of the role of causality in induction, see Bob Rehder's chapter in this volume.

Now we can start to flesh out in neurological terms what constitutes the relation between a target and an explanatory hypothesis. Mandler (2004) argues that CAUSED-MOTION is an *image schema*, an abstract, non-propositional, spatial representation that expresses primitive meanings. Lakoff (1987) and others have proposed that such non-verbal representations are the basis for language and other forms of cognition. Feldman and Narayan (2004) have described how image schemas can be implemented in artificial neural systems. I will assume that there is a neurally encoded image schema that establishes the required causal relation that ties together the neural structure of a hypothesis and the neural structure of the target that it explains. We would then have a neural representation of the explanatory, causal relation between hypotheses and targets. This relation provides the abductive basis for the inferential process described in the next section.

The model of abductive inference sketched in figures 1 and 2 has been implemented in a computer simulation that shows in detail how neural processes can generate emotional initiation and causal reasoning (Thagard and Litt, forthcoming). The details are too technical to present here, but the simulation is important because it shows how causal and emotional information distributed over thousands of artificial neurons can produce a simple form of abductive inference.

5. INFERENCE

On the standard philosophical view, inference is the movement from one or more propositions taken to be true to another proposition that follows from them deductively or

inductively. Here a proposition is assumed to be an abstract entity, the meaning content of a sentence. Belief and other mental states such as doubt, desire, and fear are all propositional attitudes, that is, relations between persons and propositions. An inference is much like an argument, which is the verbal description of a set of sentential premises that provide the basis for accepting a conclusion. Most philosophical and computational accounts of abductive inference have assumed this kind of linguistic picture of belief and inference.

There are many problems with this view. It postulates the existence of an infinite number of propositions, including an infinite number that will never be expressed by any uttered sentence. These are abstract entities whose existence is utterly mysterious. Just as mysterious is the relation between persons and propositions, for what is the connection between a person's body or brain and such abstract entities? The notion of a proposition dates back at least to Renaissance times when almost everyone assumed that persons were essentially non-corporeal souls, which could have some non-material relation to abstract propositions. But the current ascendancy of investigation of mental states and operations in terms of brain structures and processes makes talk of abstract propositions as antiquated as theories about souls or disease-causing humors. Moreover, philosophical theories of propositional belief have generated large numbers of insoluble puzzles, such as how it can be that a person can believe that Lewis Carroll wrote *Alice in Wonderland,* but not that Charles Dodgson did, when the beliefs seem to have the same content because Carroll and Dodgson are the same person.

Implicit in my account of abductive inference is a radically different account of belief that abandons the mysterious notion of a proposition in favor of biologically

realistic ideas about neural structures. In short, beliefs are neural structures consisting of neurons, connections, and spiking behavior; and so are all the other mental states that philosophers have characterized as propositional attitudes, including doubt, desire, and fear. This view does away with the metaphysical notion of a proposition, but does not eliminate standard mental concepts such as belief and desire, which are, however, radically reconstrued in terms of structures and processes in the brain (cf. Churchland, 1989).

This view of mental operations makes possible an account of the nature of inference that is dramatically different from the standard account that takes inference to operate on propositions the same way that argument operates on sentences. First, it allows for non-verbal representations from all sensory modalities to be involved in inference. Second, it allows inferences to be holistic in ways that arguments are not, in that they can simultaneously take into account a large amount of information before producing a conclusion. How this works computationally is shown by connectionist computational models such as my ECHO model of explanatory coherence (Thagard 1992). ECHO is not nearly as neurologically realistic as the current context requires, since it uses localist artificial neurons very different from the groups of spiking neurons that I have been discussing, but it at least shows how parallel activity can lead to holistic conclusions.

What then is inference? Most generally, inference is a kind of transformation of neural structures, but obviously not all such transformations count as inference. We need to isolate a subclass of neural structures that are representations, that is ones that stand for something real or imagined in the world. Roughly, a neural structure is a

representation if its connections and spiking behavior enable it to relate to perceptual input and/or the behavior of other neural structures in such a way that it can be construed as standing for something else such as a thing or concept. This is a bit vague, but is broad enough to cover both cases where neural structures stand for concrete things in the world, e.g. George W. Bush, and for general categories that may or may not have any members, e.g. unicorns. Note that one neural structure may constitute many representations, because different spiking behaviors may correspond to different things. This capability is a feature of all distributed representations.

Accordingly, we can characterize inference as the *transformation of representational neural structures*. Inference involving sentences is a special case of such transformation where the relevant neural structures correspond to sentences. My broader account has the great advantage of allowing thinking that uses visual and other sensory representations to count as inference as well. From this perspective, abduction is the transformation of representational neural structures that produces neural structures that provide causal explanations.

This neural view of inference is open to many philosophical objections. It remains to be shown that neural structures have sufficient syntactic and semantic complexity to qualify as sentential representations. The syntactic problem is potentially solved by theories of neural combinatorics such as the tensor product theory of Smolensky (1990) which show how vectors representing neural activity can be combined in ways that capture syntactic structure. The semantic problem is potentially solved by providing more detail about how neural structures can relate to the world and to each

other (Eliasmith, 2005). But much more needs to be said about what enables a neural structure to constitute a meaningful representation.

Another objection to my account of neural structures and inference is that it requires some way of specifying groups of neurons that are part of identifiable neural structures. In the worse case, one might be forced to conclude that there is only one neural structure in the brain, consisting of *all* the neurons with all their connections and spiking behaviors. This problem becomes especially acute in cases of inference that involve multiple kinds of verbal, perceptual, and emotional representation, which require that multiple brain areas be involved. In practice, however, the problem of isolating neural structures does not seem to be insurmountable. Neuroscientists often talk of groups, populations, or assemblies of neurons that are identifiable subsets of the 100 billion or so neurons that constitute an entire brain. Groups are identifiable because they have far more connections with each other than they do with neurons in other parts of the brain. Hence even though there is a great deal of interconnectivity in the brain, we can still identify groups of neurons with high degrees of connection to each other and spiking behaviors that enable them to constitute representations. So the view that inference is transformation of neural structures does not devolve into the much less precise claim that inference is just brain transformation.

Another philosophical objection to the neural theory of inference is that it is unduly narrow in that does not apply to inferences by robots or by extraterrestrial beings with brains radically different from ours. My response is that I am only concerned here to provide a special theory of human inference, and will leave the problem of developing a general theory of inference for the occasion when we actually encounter robots or aliens

that are intelligent enough that we want to count what they do as inference. The general theory would consist of a broader account of a representational structure, <parts, relations, behaviors>, analogous to the <neurons, connections, spiking behaviors> of humans and other terrestrial animals. It is an open question what the degree of similarity will be between the mental mechanisms of human and non-human thinkers, if there are any of the latter.

6. EMOTIONAL INITIATION

I described earlier how abductive inference is initiated by emotional reactions such as surprise and puzzlement, but other forms of inference also have affective origins. The influence of emotions on decision making has often been noted (Damasio, 1994; Mellers et al., 1999; Thagard, 2001, 2006). But less attention has been paid to the fact that inferences about what to do are usually initiated by either positive or negative emotions. Decisions are sometimes prompted by negative emotions such as fear: if I am afraid that something bad will happen, I may be spurred to decide what to do about it. For example, a person who is worried about being fired may decide to look for other jobs. It would be easy to generate examples of cases where other negative emotions such as anger, sadness, envy, and guilt lead people to begin a process of deliberation that leads to a practical inference. More positively, emotions such as happiness can lead to decisions, as when someone thinks about how much fun it would be to have a winter vacation and begins to collect travel information that will produce a decision about where to go. Just as people do not make abductive inferences unless there is some emotional reason for them to look for explanations, people do not make inference about what to do unless negative or positive emotional reactions to their current situation indicate that action is

required. In some cases, the emotions may be tied to specific perceptual states, for example when hunger initiates a decision about what to eat.

Deductive inference might be thought to be impervious to emotional influences, but there is neurological evidence that even deduction can be influenced by brain areas such as the ventromedial prefrontal cortex that are known to be involved in emotion (Houdé et al., 2001; Goel and Dolan, 2003). It is hard to say whether deduction is initiated by emotion, because I think it is rarely initiated at all outside the context of mathematical reasoning: Readers should ask themselves when was the last time they made a deductive inference. But perhaps deduction is sometimes initiated by puzzlement, as when one wonders whether an object has a property and then retrieves from memory a rule that says that all objects of this type have the property in question. This kind of inference may be so automatic, however, that we never become aware of the making of the inference or any emotional content of it.

Analogical inference often involves emotional content, especially when it is used in persuasion to transfer negative affect from a source to a target (Blanchette and Dunbar, 2001; Thagard and Shelley, 2001). For example, comparing a political leader to Hitler is a common way of motivating people to dislike the leader. Such persuasive analogies are often motivated by emotional reactions such as dislike of a person or policy. Because I dislike the leader, I compare him or her to Hitler in order to lead you to dislike the leader also. Practical analogical inferences are prompted by emotions in the same way that other decisions are: I want to go on vacation, and remember I had a good time at a resort before, and decide to go to a similar resort. Analogical abductions in which an

explanatory hypothesis is formed by analogy to a previous explanation are prompted by the same emotional reactions (surprise, puzzlement) as other abductive inferences.

Are inductive generalizations initiated by emotional reactions? At the least, emotion serves to focus on what is worth the mental effort to think about enough to form a generalization. As a social example, if I have no interest in Albanians I will probably not bother to form a stereotype that generalizes about them, whereas if I strongly like or dislike them I will be much more inclined to generalize about their positive or negative features. I conjecture that most inductive generalizations occur when there is some emotion-related interest in the category about which a rule is formed.

It is unfortunate that no one has collected a corpus that records the kinds of inferences that ordinary people make every day. I conjecture that such a corpus would reveal that most people make a large number of practical inferences when decisions are required, but a relatively small number of inductive and deductive inferences. I predict that deduction is very rare unless people are engaged in mathematical work, and that inductive inferences are not very frequent either. A carefully collected corpus would display, I think, only the occasional inductive generalization or analogical inference, and almost none of the categorical inductions studied by many experimental psychologists. Abductive inferences generating causal explanations of puzzling occurrences would be more common, I conjecture, but not nearly as common as practical inferences generating decisions. If the inference corpus also recorded the situations that prompt the making of inferences, it would also provide the basis for testing my claim that most inference, including practical, abductive, inductive, analogical, and deductive, is initiated by

emotions. For further discussion of the relation between deduction and induction, see the chapters in this volume by Oaksford and by Rips and Asmuth.

7. MECHANISMS

Some psychological research on inductive inference has pointed to the tendency of people to assume the presence of underlying mechanisms associated with categories of things in the world (Rehder and Burnett, 2005; Ahn, Kalish, Medin, and Gelman, 1995). Psychologists have had little to say about what mechanisms are, or how people use representations of mechanisms in their inferences. In contrast, philosophers of science have been productively addressing this issue, and the point of this section is to show the relevance of this understanding of mechanisms to the problem of abductive inference. The relevance is double, in that the neural structures I have been describing are clearly mechanisms, and abductive inferences often involve the generation or application of new hypotheses about mechanisms.

Machamer, Darden, and Craver (2000, p. 3) characterize mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." A mechanism can also be described as a system whose behavior produces a phenomenon in virtue of organized component parts performing coordinated component operations (Bechtel and Abrahamson, 2005). I think these ways of describing mechanisms are mostly equivalent, and offer my own terminological variant of a mechanism as consisting of a complex <objects, relations, changes>, consisting of a set of objects (entities, parts) that have properties and physical relations to each other that cause changes to the properties of the objects and changes to the relations the objects have to each other and to the world. For example, a bicycle is a

mechanism consisting of wheels, pedals, and other parts that are connected to each other, with regular changes to these parts and their relation to the world arising from external inputs such as a person pedaling and the internal organization of the machine. Similarly, a neural structure <neurons, connections, spiking behaviors> is clearly a mechanism where the objects are neurons, their relations are synaptic connections, and their changes are spiking behaviors. I conjecture that whenever people think of categories in terms of underlying mechanisms they have something like the pattern of <objects, relations, changes> in mind. In human minds, mechanisms can be represented verbally, as in "the pedal is bolted to the frame", but visual and kinesthetic representations are also used in science and everyday thinking (Thagard, 2003). Neural structures as well as the inferences that transform them are clearly mechanisms, and mechanisms can be mentally represented by combinations of different sorts of neural structures.

I doubt that *all* abductive inference is based on representation of mechanisms, because abduction only requires a single causal relation, not full knowledge of a mechanism. If all I know about electric lights is that when you push the switch, the light comes on, then I can abduce from the fact that the light is on that someone pushed the switch. But such knowledge hardly constitutes awareness of a mechanisms because I know nothing about any interacting system of parts. However, when I do know a lot about an underlying mechanism, my abductive inferences can be much richer and more plausible. For example, an electrician who knows much about the objects that constitute a house's electrical systems (wires, switches, fuses, etc.) is in a much better position to explain the normal occurrences and breakdowns of the system. Similarly, a physician who is familiar with the biological mechanisms that operate in a patient's body can

generate diagnoses about what might have gone wrong to produce various symptoms. The general structure of mechanism-based abductive inference is therefore:

> Mechanism <objects, relations, changes> is behaving in unexpected ways.
>
> So maybe there are unusual properties or relations of objects that are responsible for this behavior.

Mechanism-based abduction differs from the simple sort in that people making inferences can rely on a whole collection of causal relations among the relevant objects, not just a particular causal relation.

So far, I have been discussing abduction *from* mechanisms, in which representations of a mechanism is used to suggest an explanatory hypothesis about what is happening to the objects in it. But abductive inference is even more important for generating knowledge about how the mechanism works, especially in cases where its operation is not fully observable. With a bicycle, I can look at the pedals and figure out how they move the chains and wheels, but much of scientific theorizing consists of generating new ideas about unobservable mechanisms. For example, medical researchers develop mechanistic models of how the metabolic system works in order to explain the origins of diseases such as diabetes. Often, such theorizing requires postulation of objects, relations, and changes that are not directly observed. In such cases, knowledge about mechanisms cannot be obtained by inductive generalization of the sort that works with bicycles, but depends on abductive inference in which causal patterns are hypothesized rather than observed. This kind of abduction *to* mechanisms is obviously much more difficult and creative than abduction from already understood

mechanisms. Often it involves analogical inference in which a mechanism is constructed by comparing the target to be explained to another similar target for which a mechanism is already understood. In this case, a mechanism <objects, relations, changes> is constructed by mapping from a similar one. For more on analogical discovery, see Holyoak and Thagard (1995, ch. 8).

In order to show in more detail how abductive inference can be both true and from mechanisms, it would be desirable to apply the neurocomputational model of abductive inference developed by Thagard and Litt (forthcoming). That model has the representational resources to encode complex objects and relations, but has not yet been applied to temporal phenomena involving change. Hence neural modeling of inferences about mechanisms is a problem for future research.

8. CONCLUSION

In sum, abduction is multimodal in that can operate on a full range of perceptual as well as verbal representations. It also involves emotional reactions, both as input to mark a target as worthy of explanation and as output to signal satisfaction with an inferred hypothesis. Representations are neural structures consisting of neurons, neuronal connections, and spiking behaviors. In abduction, the relation between hypotheses and targets is causal explanation, where causality is rooted in perceptual experience. Inference is transformation of representational neural structures. Such structures are mechanisms, and abductive inference sometimes applies knowledge of mechanisms and more rarely and valuably generates new hypotheses about mechanisms.

Much remains to be done to flesh out this account. Particularly needed is a concrete model of how abduction could be performed in a system of spiking neurons of

the sort investigated by Eliasmith and Anderson (2003) and Wagar and Thagard (2004). The former reference contains valuable ideas about neural representation and transformation, while the latter is useful for ideas about how cognition and emotion can interact. Thagard and Litt (forthcoming) combines these ideas to provide a fuller account of the neural mechanisms that enable people to perform abductive inference. Moving the study of abduction from the domain of philosophical analysis to the realm of neurological mechanisms has made it possible to combine logical aspects of abductive inference with multimodal aspects of representation and emotional aspects of cognitive processing. We can look forward to further abductions about abduction.

Acknowledgements: I am grateful to Jennifer Asmuth, Aidan Feeney, Abninder Litt, and Douglas Medin for helpful comments on an earlier draft. Funding for research was provided by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299-352.

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 79-116). Oxford: Clarendon Press.

Bechtel, W., & Abrahamsen, A. A. (2005). Explanation: A mechanistic alternative. *Studies in History and Philosophy of Biology and Biomedical Sciences, 36*, 421-441.

Blake, R. M., Ducasse, C. J., & Madden, E. H. (1960). *Theories of scientific method: The renaissance through the nineteenth century*. Seattle: University of Washington Press.

Blanchette, I., & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition, 29*, 730-735.

Charniak, E., & McDermott, D. (1985). Introduction to artificial intelligence. In. Reading, MA: Addison-Wesley.

Churchland, P. M. (1989). *A neurocomputational perspective*. Cambridge, MA: MIT Press.

Damasio, A. R. (1994). *Descartes' error*. New York: G. P. Putnam's Sons.

Darden, L. (1991). *Theory change in science: Strategies from Mendelian genetics*. Oxford: Oxford University Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461-493). Washington: American Psychological Association.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (Vol. 1035-1054). Amsterdam: Elsevier.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Feldman, J., & Narayan, S. (2004). Embodied meaning in a neural theory of language. *Brain and Language, 89*, 385-392.

Flach, P. A., & Kakas, A. C. (Eds.). (2000). *Abduction and induction: Essays on their relation and integration*. Dordrecht: Kluwer.

Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., & Dunbar, K. N. (2005). Brain mechanisms underlying perceptual causality. *Cognitive brain research, 24*, 41-47.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Goel, V., & Dolan, R. J. (2003). Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning. *NeuroImage, 20*, 2314-2321.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*, 101-118.

Hanson, N. R. (1958). *Patterns of discovery*. Cambridge: Cambridge University Press.

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press/Bradford Books.

Houdé, O., Zago, L., Crivello, F., Moutier, F., Pineau, S., Mazoyer, B., et al. (2001). Access to deductive logic depends on a right ventromedial prefrontal area devoted to emotion and feeling: Evidence from a training paradigm. *NeuroImage, 14*, 1486-1492.

Josephson, J. R., & Josephson, S. G. (Eds.). (1994). *Abductive inference: Computation, philosophy, technology*. Cambridge: Cambridge University Press.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Kunda, Z. (1999). *Social cognition*. Cambridge, MA: MIT Press.

Kunda, Z., Miller, D., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science, 14*, 551-577.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Langley, P., Shrager, J., Asgharbeygi, N., Bay, S., & Pohorille, A. (2004). Inducing explanatory process models from biological time series. In *Proceedings of the ninth workshop on intelligent data analysis and data mining*. Stanford, CA.

Langley, P., Simon, H., Bradshaw, G., & Zytkow, J. (1987). *Scientific discovery*. Cambridge, MA: MIT Press/Bradford Books.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition, 25*, 265-288.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40*, 87-137.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.

Maass, W., & Bishop, C. M. (Eds.). (1999). *Pulsed neural networks*. Cambridge, MA: MIT Press.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Magnani, L. (2001). *Abduction, reason, and science: Processes of discovery and explanation*. New York: Kluwer/Plenum.

Mandler, J. M. (2004). *The foundations of mind: Origins of conceptual thought*. Oxford: Oxford University Press.

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General, 128*, 332-345.

Michotte, A. (1963). *The perception of causality* (T. R. Miles & E. Miles, Trans.). London: Methuen.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, N. J.: Prentice Hall.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Peirce, C. S. (1931-1958). *Collected papers*. Cambridge, MA: Harvard University Press.

Psillos, S. (1999). *Scientific realism: How science tracks the truth*. London: Routledge.

Read, S., & Marcus-Newhall, A. (1993). The role of explanatory coherence in the construction of social explanations. *Journal of Personality and Social Psychology, 65*, 429-447.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264-314.

Rieke, F., Warland, D., de Ruyter van Steveninick, R. R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Shelley, C. P. (1996). Visual abductive reasoning in archaeology. *Philosophy of Science, 63*, 278-301.

Sloman, S. A., & Lagnado, D. (2005). The problem of induction. In R. Morrison & K. Holyoak, J. (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 95-116). Cambridge: Cambridge University Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*, 159-217.

Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press/Bradford Books.

Thagard, P. (1992). *Conceptual revolutions*. Princeton: Princeton University Press.

Thagard, P. (1999). *How scientists explain disease*. Princeton: Princeton University Press.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. (2003). Pathways to biomedical discovery. *Philosophy of Science, 70*, 235-254.

Thagard, P. (2005). *Mind: Introduction to cognitive science* (2nd ed.). Cambridge, MA: MIT Press.

Thagard, P. (forthcoming). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Thagard, P., & Litt, A. (forthcoming). Models of scientific explanation. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge: Cambridge University Press.

Thagard, P., & Shelley, C. P. (1997). Abductive reasoning: Logic, visual thinking, and coherence. In M. L. Dalla Chiara, K. Doets, D. Mundici & J. van Benthem (Eds.), *Logic and Scientific Methods* (pp. 413-427). Dordrecht: Kluwer.

Thagard, P., & Shelley, C. P. (2001). Emotional analogies and analogical inference. In D. Gentner, K. H. Holyoak & B. K. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 335-362). Cambridge, MA: MIT Press.

van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

Wagar, B. M., & Thagard, P. (2004). Spiking Phineas Gage: A neurocomputational theory of cognitive-affective integration in decision making. *Psychological Review, 111*, 67-79.