THE SELF AS A SYSTEM OF MULTILEVEL INTERACTING MECHANISMS

Paul Thagard
Philosophy Department, University of Waterloo
June, 2012

Thagard, P. (forthcoming). The self as a system of multilevel interacting mechanisms. *Philosophical Psychology*.

ABSTRACT

This paper proposes an account of the self as a multilevel system consisting of social, individual, neural, and molecular mechanisms. It argues that the functioning of the self depends on causal relations between mechanisms operating at different levels. In place of reductionist and holistic approaches to cognitive science, I advocate a method of multilevel interacting mechanisms. This method is illustrated by showing how self-concepts operate at several different levels.

1. WHAT IS THE SELF?

The concept of the self has been theoretically important in philosophy, psychology, and related social sciences, including sociology, anthropology, and political science. The nature of the self is relevant to explaining many interesting phenomena, including self-consciousness, self-control, and self-esteem (see Thagard and Wood, forthcoming, for a discussion of more than eighty of them). Many other important ideas, including agency, autonomy, personhood, and responsibility, are tightly connected with these phenomena concerning the self, which I will call the *self-phenomena*.

Despite the centrality of such phenomena in social and clinical psychology, in philosophy of mind, and in related social sciences, there is currently no general, rigorous account of the self that can provide a principled, organized explanation of them. I propose that the self is best understood as a multilevel system, encompassing mechanisms that interact across four interconnected levels: social, individual, neural, and

molecular. Each of these levels can be understood as a subsystem consisting of environmental influences, component parts, interconnections between parts, and regular changes in the properties and relations of the parts. This approach rejects both the holistic view that higher levels are autonomous from lower levels and the individualistic view that higher levels can be entirely explained by mechanisms at lower levels. I use the term *multilevelism* to stand for the view that attention to multiple levels avoids the implausible assumptions and consequences of both individualistic reductionism and holistic antireductionism.

The idea of levels of explanation is common in cognitive science, for example in Simon (1962), Newell (1990), Churchland and Sejnowski (1992), Darden (2006), and Craver (2007). What I call multilevelism is similar to the *explanatory pluralism* of McCauley and Bechtel (2001), and to the *systemism* of Bunge (2003). It would be nicer to have a term as catchy as "holism" and "reductionism", but the Greek word for levels, *epipedos,* would yield the term *epipedism,* which sounds like a sexual perversion or skin disease.

My account of the self is radically different from most philosophical approaches, which tend to be either transcendental or deflationary. Transcendental views, held by philosophers such as Plato, Aquinas, Descartes, and Kant, take selves as supernatural entities – souls – that are not open to mechanistic explanation using the methods of natural science (Organ, 1987). At the other, deflationary extreme, some philosophers have been skeptical of the idea of the self as a determinate kind of thing, proposing instead that the self is just a bundle of perceptions (Hume, 1888), a convenient fiction amounting to a "center of narrative gravity" (Dennett, 1991), or simply a myth

(Metzinger, 2009). Similarly, postmodernist sociologists view selves as mere social constructions (Callero, 2003). In analytical, phenomenological, and Indian traditions, debates continue about whether the self is a substance, non-substance, or nothing at all (Gallagher, 2011; Siderits, Thompson, and Zahavi, 2011).

In contrast, social and clinical psychologists make substantial use of the concept of the self in their discussions of a wide range of phenomena (e.g., Baumeister, 1999; Alicke, Dunning, and Krueger, 2005). But they have largely shied away from the task of saying what selves are. The multilevel account is intended to fill this gap while avoiding the metaphysical extravagance of transcendental views and the explanatory impotence of deflationary ones. I follow in the tradition of James (1890) and Mead (1967) in taking a multifaceted approach that accommodates social, cognitive, and physiological aspects of the self, but provide far more detail about the nature of the relevant mechanisms and phenomena. Like Bechtel (2008), I adopt a mechanistic approach to the self, but stress the importance of integrating mechanisms that operate at multiple levels. I will illustrate the importance of such integration by showing how it applies to self-concepts.

2. MULTILEVEL SYSTEMS

In order to identify a self as a multilevel system, we need to characterize the systems, levels, and mechanisms that constitute selves. My characterization synthesizes and adapts ideas developed by philosophers of science, particularly Bunge (2003) and Bechtel (2008). We can define a system as a structure, <Environment, Parts, Interconnections, and Changes>, EPIC for short. Here the parts are the objects (entities) that compose the system. To take a simple example, a bicycle is composed of such parts as the frame, wheels, handlebars, chain, and pedals. The environment is the collection of

items that act on the parts, which for a bicycle would include people who push on the pedals, roads that interact with the wheels, and air molecules that provide wind resistance to the handlebars. The interconnections are the relations among the parts, especially the bonds that tie them together. In a bicycle, key relations include the physical connections between the chain and the wheels and between the handlebars and the frame. Finally, the changes are the processes that make the system behave as it does, for example the turning of the bicycle's chain and wheels.

The self cannot be easily decomposed into a single EPIC system. Even a bicycle can be understood at multiple physical levels—for example, with the wheel decomposing into various parts such as the hub, the rim, the tube, and spokes, each of which consist of molecules, which consist of atoms, which consist of sub-atomic particles, which may consist of quarks or multidimensional strings. For most purposes, it suffices to consider bicycles at the single level of observable parts such as wheels and pedals in interaction with each other, although an engineer attempting to optimize performance may have reason to work at lower levels, as when nanotechnology is used to design extremely light racing bikes.

To characterize multilevel systems, we can generalize the EPIC idea and think of a multilevel system as consisting of a series of quadruples, with the structure:

$$<E_1, P_1, I_1, C_1>$$
$$<E_2, P_2, I_2, C_2>$$
$$\ldots$$
$$<E_n, P_n, I_n, C_n>.$$

At each level, there is a subsystem consisting of the relevant environment, parts, interconnections, and changes. A later section lays out the relations between environments, parts, interconnections, and changes at different levels.

What are the most important levels for understanding selves? The answer to this question depends on what mechanisms are needed to explain the many interesting self-phenomena. I conjecture that there are four main subsystems required for such explanations, operating at social, individual, neural, and molecular levels, which are the levels that can be used to explain emotions, consciousness, and other important aspects of thinking (Thagard 2006, 2010a). To spell out the claim that the self is a multilevel system, we need to describe each of the four levels, specifying their parts, interconnections, environment, and changes.

3. LEVELS OF THE SELF

3.1. The Social Self

At the most familiar social level, the set of parts consists of individual persons. Even at this level, there is a hierarchy of additional sublevels of social organization, such as families, neighborhoods, regions, nations, and states, just as at the neural level there are additional levels of organization such as populations of neurons and brain areas. The social parts are influenced by an environment that includes all the objects that people causally interact with, including natural objects such as rocks and lightning bolts, artifacts such as houses and cars, and social organizations such as teams and governments. The interconnections at the social level consists of the myriad relations among people, ranging from mundane perceptual ones such as a person being able to recognize another, to deeper bonds such as being in love, to ones involving several persons, such as belonging to the same sports team. Finally, the changes at the social level consist of the many processes of human interaction, ranging from talking to playing games to sexual intercourse. Humans are social animals (Aronson, 2003).

3.2 The Individual Self

At the individual level, the self consists of personal behaviors and the many mental representations that people apply to themselves and others. The most common representations are personality concepts, such as *kind, mean, cheerful, morose, adventurous, cautious, agreeable, hostile, sociable, unfriendly,* and hundreds of others. People use such concepts to form rule-like beliefs about individuals, such as that a friend is optimistic, as well as about social groups, such as that Canadians are courteous. Behaviors are properties of individuals, but mental representations can be considered as parts of them if one adopts an information-processing rather than a commonsense view of the mind.

There are at least three different ways of talking about mental representations, found in everyday conversation, philosophical discourse, and current psychological theories. In everyday conversation, people speak of mental states such as beliefs, emotions, concepts, and ideas in ways tied to dualist notions that mind is a non-material, supernatural substance. In contrast, my concern is with developing a scientific, evidence-based theory of the self, so I will pay no further attention to everyday concepts of mental entities that derive from unreflective introspections and theistic metaphysics.

Nor will I pay much attention to current philosophical theories of mental representation that view beliefs as propositional attitudes, which are supposed to be relations between persons and abstract entities (propositions) that are the meanings (content) of sentences. The doctrine of mental states as propositional attitudes has been critiqued elsewhere (Churchland, 2007; Thagard 2008, 2010a). From the perspectives of folk psychology and standard philosophy of mind, it is odd to describe mental

representations such as concepts and beliefs as *parts* of people. More commonly, concepts and beliefs are spoken about as if they are *possessions* of people, and the philosophical idea of propositional attitudes understands mental representations as relations between people and abstract entities. Some philosophers claim that to speak otherwise of mental representations is to commit a category mistake.

This objection, however, is scientifically naïve, because the point of theoretical development is to change concepts, not to stick with ordinary ones. Folk psychology has no more claim to truth than folk physics, chemistry, and biology, all of which have long since been superseded by scientific ideas. Since the 1960s, cognitive psychology has developed new, information-processing conceptions of concepts and other mental representations, by analogy to structures and processes used in computers. On this analogy, at least at a crude level, concept and beliefs are like the data structures (e.g. strings, lists, objects, arrays, etc.) that are part of a computer program, which is part of a running computer. Analogously, mental representations can be parts of people, in a way that is even more obviously true from the perspective of the neural level to be discussed below. Cognitive psychology abounds with ideas about what kinds of computational structures might be found in the mind. For example, there are diverse theories about concepts (e.g. Murphy, 2002), and processing theories about non-supernatural propositions (e.g. Anderson, 1983).

Thus, at the individual level, the self consists of a subsystem where the parts are mental representations such as concepts, schemas, beliefs, attitudes, propositions, rules, situations, analogies, images, and so on – all the kinds of representations found in textbooks in cognitive science (e.g. Thagard, 2005). The environment for these parts

consists of all the objects in the world that can be inputs to and outputs from mental processes, including objects in the world and other people. The interconnections of a system of mental representations consists of the relations between them, particularly the bonds that hold them together. For examples, concepts are organized by kind and part-whole relations: the concept *bicycle* is related to concepts *machine* and *wheel,* because a bicycle is a kind of machine and its parts include wheels. Beliefs have concepts as parts, as when people put the concepts *bicycle* and *heavy* together to form the belief that bicycles are healthy.

Folk psychology can tell us nothing about the processes that cause the interactions of mental representations, and philosophical psychology has only limited theories of inference such as ones based on deductive logic. But cognitive psychology over the past 40 years has developed rich ideas about mental processing that apply to a wide range of mental representations, from concepts, to rules, to images. For example, theories of spreading activation among concepts explain many interesting phenomena about memory and language such as priming effects. Rule-based thinking has been modeled by processing systems such as ACT that provide detailed accounts of inferential mechanisms (Anderson, 2007). These theories and their attendant computational models generate mappings from the properties that apply to mental representations at one time and the properties that apply at a later time. Thus, cognitive psychology provides accounts of the processes by which concepts, rules, and other mental representations change over time. Increasingly, cognitive theories are being tied to neural processes.

3.3 The Neural Self

Characterizing the neural subsystem is relatively straightforward. The most important parts of the brain are neurons, which are cells that also exist in related parts of the nervous system such as the spine. The interconnections of the neural system are largely determined by the excitatory and inhibitory synaptic connections between neurons, although glial cells in the brain and hormonal processes are also relevant (Thagard, 2006, ch. 7). The environment of the neural system is better described at a smaller scale than the level of whole objects appropriate for the individual and social levels. For example, photons of light stimulate retinal cells and initiate visual processing in the brain, and sound waves affect the structure of the ear and initiate auditory processing. Thus the environment of the neural system consists of those physiological inputs that influence neural firing. Finally, the changes in the neural subsystem include alterations in firing patterns resulting from excitatory and inhibitory inputs from other neurons, as well as alterations in the synaptic connections (see e.g. Dayan and Abbott, 2001; Eliasmith and Anderson, 2003; O'Reilly and Munakata, 2000).

Folk and philosophical psychology totally ignore the neural level, but in current cognitive science the neural and representational levels are increasingly becoming integrated (e.g. Anderson 2007, Smith and Kosslyn, 2007). I have used the term "representational" to refer to familiar structures such as concepts and beliefs, but the activities of neural populations can be representational too, by encoding features of the external and internal world. As an inert object, a single neuron does not represent anything, although there are special cases where the firing activity of individual neurons can stand for things in the world, for example specific actors such as Jennifer Aniston (Quiroga et al., 2005). More commonly, neural representations are accomplished by the

joint firing activity of populations of neurons. Particular self-representations can be performed by populations of neurons that fire in ways that causally correlate with aspects of the self and world.

3.4 The Molecular Self

Just as cognitive psychology has drawn increasingly on neuroscience in the past two decades, neuroscience has drawn increasingly on molecular biology. Neurons are cells consisting of organelles such as nuclei and mitochondria, and the firing activity of neurons is determined by their chemical inputs and internal chemical reactions. Aspects of the self such as personality are influenced by biochemical factors including genes, neurotransmitters, and epigenetic factors that modify the expression of genes.

Genetic effects on behavior are displayed by studies that find higher correlations between some features in identical twins than in non-identical ones, for example in tendencies toward mental illnesses such as schizophrenia. Humans have variation in genes that determine the receptors for more than fifty different neurotransmitters that affect neuronal firing. For instance, there are variations in the gene DRD4 that controls the formation of the D_4 receptor for the neurotransmitter dopamine. These variations are associated with behavioral effects such as the personality trait of novelty seeking (Benjamin et al., 1996) It would be naïve, however, to suppose that there are "genes for" particular behaviors, because of increasing evidence for the importance of multiple genes and for epigenetic effects on the operation of genes (e.g. Richards, 2006). Whether a gene expresses a particular protein depends not only on the gene, but also on the attachment of various chemicals such as methyl groups, which are affected by the overall environment of the cells that contain the genes.

In sum, a self is a system consisting of subsystems at four levels – social, individual, neural, and molecular – each of which includes environment, parts, interconnections, and changes. In writing of the social, individual, neural, and molecular selves, I am **not** taking a person to consist of four separate selves. Rather, the self is the integration of all four levels, as can be shown by considering the relations among them.

4. RELATIONS AMONG LEVELS

From the EPIC perspective on systems, we need to look in detail at the relations between environment, parts, interconnections, and changes at different levels. The relations between parts are the most straightforward. As a first approximation, we can say that the parts at one level are composed of the parts at the next level down. This relation is most obvious at the intersection of the neural and the molecular levels, as biology makes it clear that the parts of neurons include molecular parts such as proteins and genes. But composition is more complicated in other cases. Does it really make sense to say that mental representations are parts of persons, and that neurons are parts of mental representations?

I already argued that the information-processing idea that representations are parts of people should not be rejected because of the commonsense idea that beliefs are properties of people. Concepts can be parts of people in the same way that data structures are parts of computers loaded with software programs. It also takes some conceptual revision to see neurons as parts of mental representations, which in the early days of cognitive science were largely viewed as functional computational entities not tied to any particular kind of physical instantiation. The rapid development of cognitive neuroscience, however, has made it more natural to think of concepts and mental

representations as patterns of neural activity. But are neurons as things – nerve cells – parts of dynamic entities like neural activity, let alone parts of more abstract entities such as patterns?

It is easier to answer this question if we distinguish between occurrent and dispositional aspects of mental representations, following the traditional philosophical distinction between occurrent and dispositional belief. People have beliefs that they are not currently thinking about: Five minutes ago, you were probably not thinking that Canada is in North America, but you probably believed it, in the sense that you had a disposition to say yes when asked if Canada is in North America. Once you are asked, the belief becomes occurrent when you are actually thinking that Canada is in North America. Analogously, a spoonful of sugar has the disposition to be soluble in water that makes it dissolve. Sugar has this disposition because of intermolecular forces arising from its chemical structure and that of water. Similarly, a pattern of neural activity occurs because of synaptic connections between members of a neural population. Hence, from the perspective of cognitive neuroscience, a dispositional belief is a pattern of neural connections that, given external and internal stimuli, will lead to a pattern of neural firing. Because a pattern of neural connections is a combination of neurons and their synaptic links with other neurons, it is natural to say that neurons are parts of mental representations in the dispositional sense. It is only a small step to acknowledge that neurons are also parts of patterns of firing activity in neural populations, in the same way that the colored threads in a quilt are part of the pattern on the quilt.

It might seem that this discussion of composition implies or presupposes a simple reductionist view of the self, with molecules as parts of neurons, which are parts of

mental representations, which are parts of persons, which are parts of groups. However, this unidirectional, asymmetric ordering does not imply that causality needs to be similarly unidirectional: I argue later that social processes can causally affect molecular processes.

Now we can consider the relations between environments that operate in the multilevel system of the self. At the extreme, the large objects that influence the social system are very different from the minute ones that influence the molecular system. Within adjacent levels, however, there seems to be much overlap between environments. Large scale objects in the world such as buildings and rivers influence persons (operating at the social level) and mental representations (operating at the individual level). Such objects also have effects at the neural level, through psychophysical processes of perception, as when light reflects off a building and photons stimulate the retina to initiate a cascade of neural processing. It seems, then, that the relation between levels of environment is sometimes identity, sometimes part-whole (as when the light reflects of the windows of a building), and sometimes a more complex causal process. The complexity of environmental influences derives from the fact that environments are also multilevel systems ranging from microbes to large-scale terrains and climates, with which humans as multilevel systems interact at levels ranging from the cellular to the social.

The third aspect of the EPIC account of systems concerns interconnections, the set of relations that hold between objects, especially the bonds that hold them together. How can we characterize the abstract connection between bonds that operate at one level and bonds that operate at lower ones? Consider a simple physical case. When two pieces of wood are joined by a nail, their bond is the result of physical forces operating at a

lower level, connecting the molecules of the nail with the molecules of the two pieces of wood, where these molecular bonds are in turn the result of subatomic, quantum-mechanical processes. Similarly, for each bond at a higher level in a multilevel system, we should look for a causal process at the next level down that produces it. Higher bonds do not have lower bonds as parts, but rather emerge from causal processes involving lower bonds.

Similarly, in the multiple levels that comprise the self, the bonds at each level are the causal results of processes operating at lower levels. At the social level, groups are formed by bonds between persons that are partly the result of the operations of mental representations at the lower level. For example, when two people become friends, their friendship results from a complex of mental representations that each has about the other, including concepts such as *nice*, beliefs such as "She likes me", and emotions such as feeling happy when the other person is around.

It is harder to connect the bonds between mental representations with underlying neural processes, because detailed knowledge of the relevant neural mechanisms is still lacking. But for some simple cases such as association between concepts, informed conjectures are possible. There is a bond between the concepts *cat* and *dog,* in that both cats and dogs are kinds of animals that are often pets. Activating the concept *cat* will therefore likely lead to activation of the concept *dog,* in a way that can be understood at the neural level. If the two concepts are both patterns of neural firing, then their association results from synaptic links between the neurons involved in one pattern and the neurons involved in the other pattern, which may include some overlapping neurons

and links. Hence the bond between the two concepts that leads to their association plausibly results from the underlying neural structure and activity.

Similarly, the bonds between two neurons – their synaptic connections – are the results of molecular processes that link the axons of the presynaptic neuron with the dendrites of the postsynaptic neuron. Bond relations, like part-relations, seem to be unidirectional and therefore asymmetric: bonds at a higher level result from causal processes at a lower level, but bonds at a lower level are independent of bonds at the higher level. In contrast, the relations between changes at different levels are not asymmetric in this way, as changes at higher levels can cause changes at lower levels (see examples below).

Identifying relations between changes requires considering the parts at both levels, as well as the properties and relations that alter over time. Changes in systems can be described in many ways, using words, diagrams, and mathematical equations. How do changes in groups relate to changes to persons, mental representations, neural populations, and molecular configurations? The simplest answer would be the reductionist one that property changes at the higher level always result from property changes at the lower level. Such determinations are often the case, when changes in group interactions result from changes in mental representations that result from neural and molecular changes. For example, consider the social change of John approaching Mary, because she smiled at him, because she was mentally representing him as attractive, because of firing of neural populations in her visual cortex and dopamine-rich nucleus accumbens. Often, therefore, the reductionist picture is correct in portraying

molecular changes that cause neural changes that cause individual changes that cause social changes.

Frequently, however, causality runs in the opposite direction. When Mary smiles at John, this social interaction is clearly the cause of a course of changes in him that are individual, neural, and molecular. He perceives her smiling and probably infers that she likes him, which are changes in mental representation that are also neural changes. Then social changes cause individual, neural, and molecular changes evident in many other situations, such as:

- Giving a presentation increases levels of the stress hormone cortisol.
- Seeing a beloved causes increased activity of dopamine neurons.
- Men whose favorites sports team have won a game enjoy increased levels of testosterone.
- Male chimpanzees who become dominated have lowered levels of testosterone.
- Women who room together tend to have their menstrual cycles coordinated, altering patterns of estrogen levels.

Hence social changes cause molecular changes.

More contentiously, I want to claim that individual changes can cause neural changes, for example when John's inferring that Mary likes him (a change in mental representation) cause increased neural activity in various brain areas such as the nucleus accumbens. Hence contrary to the reductionist view that causality is always from lower levels to higher, I prefer the interactive view presented in figure 1.

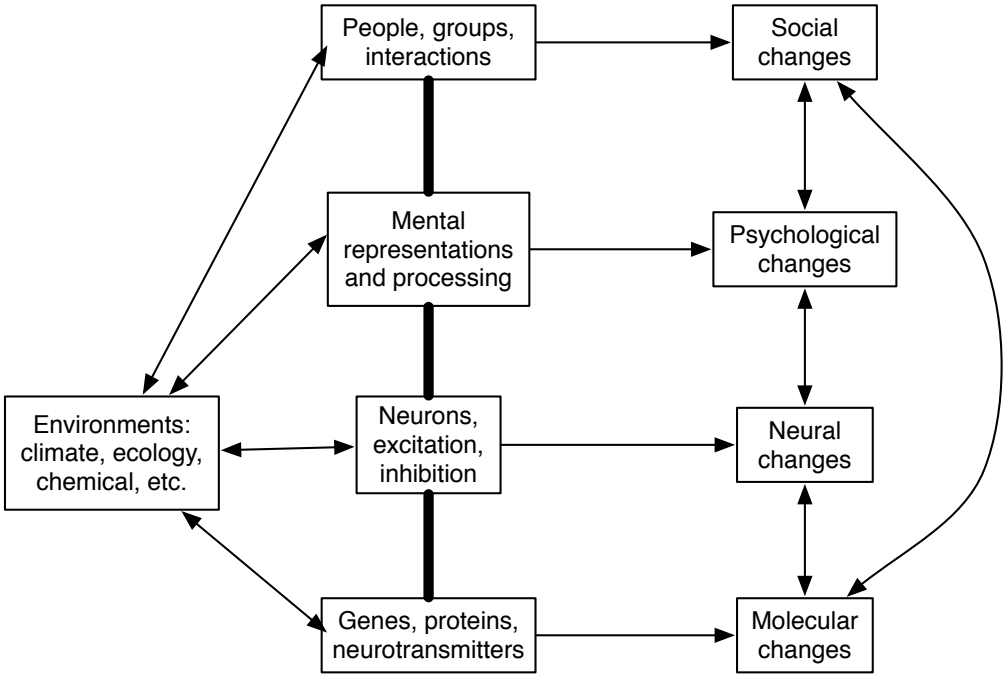

Figure 1. Diagram of the self as a multilevel system. Lines with arrows indicate causality. Thick lines indicate composition.

My account of levels in this paper is largely compatible with discussions by philosophers such as Bechtel (2008), Craver (2007) and Wimsatt (2007). The multilevel mechanisms approach to the self potentially has implications for many other problems in philosophy, psychology, and social science. It suggests an understanding of agents as far more complex than is generally assumed in philosophical discussions of autonomy and personhood, in psychological and sociological discussions of identity, and in economic and political discussions of rational choice and power. Moreover, the MIM view of the self can naturally be generalized to consideration of the interacting mechanisms that operate in all social organizations, from families to nations, which are also multilevel systems.

The examples given of effects of the social level on the molecular should make it clear why worries about downward causation are misplaced. Claims such as that a social insult can cause an increase in cortisol levels are unproblematic on all reasonable accounts of causality, even though they cross levels. On probabilistic accounts, the probability of high cortisol levels given an insult is greater than the probability of high levels without an insult. On manipulation accounts, intervening in a social situation by generating an insult clearly results in the higher cortisol levels. On mark-transmission accounts, the social interaction transmits energy in the form of sound waves to the hearer, changing the flow of energy all the way down to the molecular level. The social interaction clearly is a distinct event from the raising of cortisol levels and precedes it, even though people decompose into underlying parts. Changes at time t at one level cause changes at time t+1 at another level. This relation is easier to understand if changes are represented by difference equations or movies rather than by differential equations or static diagrams.

The problem of distinct events is more acute when the causal relations are between adjacent levels, for example between the individual level of mental representations and the neural level. Can we legitimately say that an inference such as an instance of modus ponens operating in the mind of an individual causes neurons to fire? The problem here is that cognitive neuroscience suggests that the propositions used in the modus ponens are just patterns of firing in neural populations, and inference is a process of transformation of firing patterns. Then the relation between the inference and the neural process is identity, not causality. In the abstract, this sounds correct, but in practical circumstances of explanation it does not apply, because we currently do not

have, and may never have, knowledge about the exact instantiation of mental representations. Lacking such knowledge, there is nothing wrong with saying that someone has activation in a particular brain area because he/she made an inference. For example, we could say that Othello (in Shakespeare's play) has activation in his amygdala because he inferred that his wife Desdemona is unfaithful to him. The individual psychological process produced a neural change.

5. EXAMPLE: SELF-CONCEPTS

The scope of the multilevel mechanism account of the self is intended to be very broad, encompassing all the phenomena concerning the self identified by philosophers, psychologists, and sociologists. Figure 2 shows the taxonomy of self-phenomena developed by Thagard and Wood (forthcoming). They provide extensive discussions of seven of the eighty shown in Figure 2: self-concepts, self-presentation, self-esteem, self-enhancement, self-regulation, self-expansion, and self-development. Here I will focus on the multilevel mechanisms relevant to self-concepts, ignoring many experimental studies cited in Thagard and Wood (forthcoming). The goal here is to indicate not only how self-concepts operate at each level, but also how causation crosses the levels at which they operate.

Self-concepts are the many word-like mental representations that people apply to themselves, such as *man, woman, Canadian, American, professor, teacher, happy, sad, introvert, extravert,* and so on. Table 1 provides a concise summary of how EPIC mechanisms for self-concepts occur at all four levels with interactions between them. Self-concepts are most familiar at the individual, psychological level. Psychologists such as (Murphy, 2002) variously consider concepts as prototypes (e.g. the typical

Canadian), sets of exemplars (e.g. William Shatner), and explanatory theories (e.g. he said sorry because he's a polite Canadian). At this level, concepts are the parts that figure into various kinds of interactions, such as inferences, associations, and combinations, as in the generation of *Canadian woman,* or *happy extravert*. Such interactions help to explain many psychological phenomena, such as categorization and inference.

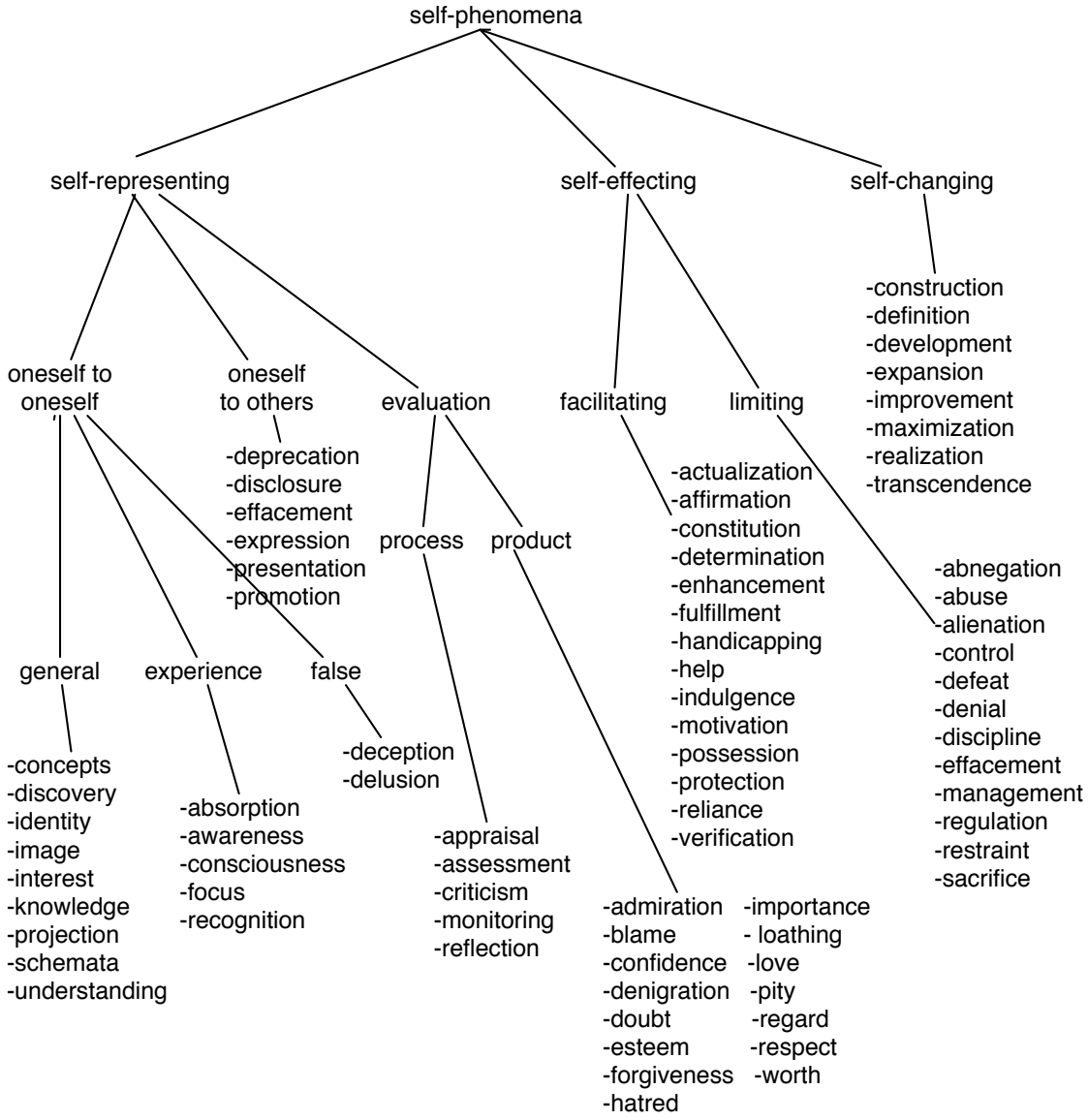

Figure 2. Grouping of many self-phenomena into six main classes, including three kinds of self-representing, two kinds of self-effecting, and self-changing.

	Environment	Parts	Interactions	Changes
Social	Groups	People	Communication	Behaviors
Individual	Physical and social	Concepts	Inferences, combination	Categorization, reasoning
Neural	Physical and social	Neurons Semantic pointers	Excitation and inhibition, binding	Firing patterns, synaptic connections
Molecular	Bodies	Neurotransmitters, hormones	Biochemical reactions	Chemical concentrations

Table 1. Summary of how four aspects of EPIC mechanisms operate at four different levels relevant to self-concepts.

At the social level, self-concepts play an important role in the communications and behaviors that take place in groups of people. When people apply concepts to themselves, they identify themselves as members of groups in ways that often affect how they talk to other people and behave in their presence. For example, locating and identifying yourself as a member of an academic department can lead you to communicate and behave in ways different from what happens when you are involved with a sports team. Such interactions show how social changes can produce individual-level changes in self-concepts, when associating with different groups leads you to think

of yourself in different ways. (Psychologists call this the "malleability of the self-concept" Markus and Kunda, 1986). Of course, causation can work in the other direction as well, when the concepts that people apply to themselves lead them to seek out and spend time with various groups. Thus the social and the individual level mechanisms interpenetrate: the interactions of people change the application of concepts, and the application of concepts changes the interaction of people.

At the individual, psychological level, concepts including self-concepts seem hard to pin down, as psychological evidence fails to univocally support prototype, exemplar, or explanatory-theory accounts of concepts. Thagard (2010a) suggested that moving down to the neural level could provide a unified account of the nature of concepts, and new simulations show that such unity is indeed achievable (Blouw, Solodkin, Eliasmith, and Thagard (forthcoming). At this level, the parts are neurons that interact through processes of excitation and inhibition that enable them to be organized into populations of neurons whose patterns of firing allow them to functions as concepts, including self-concepts. Eliasmith (in press) discusses concepts in terms of semantic pointers, a powerful, complex kind of neural representation that is capable of both symbol-like functioning and expansion into associate sensory information. If concepts are semantic pointers, then it becomes possible to see how they can have diverse functions such as categorization and inference while retaining some contact with sensory experience. Thagard (2012) argues that scientific concepts can fruitfully be understood as semantic pointers.

If concepts, including self-concepts, are identified with semantic pointers, it becomes complicated to see how there can be causal relations between the individual

level and the neural level, which might seem to be collapsed. Despite the identification, however, it still makes sense to talk of inter-level causation because the processes (changes) are so different. For example, consider what happens when a person undergoing surgery for epilepsy has neurons electrically stimulated, generating a memory of being a child. Then it makes sense to say that the neural manipulation causes activation of the self-concept *child,* even if that concept is construed as a pattern of activation in a population of neurons. Going in the other direction, when people make inferences about themselves such as "I'm happy being with other people, so I'm an extravert", this psychological event has a neural effect, namely increased firing in the neural population corresponding to the self-concept *extravert*.

The molecular mechanisms associated with self-concepts are rarely discussed, but can nevertheless be recognized. Like other concepts, self-concepts have associated emotional valences, positive or negative. For most people, the concepts of *successful* and *failure* respectively have positive and negative valence. Such valences are associated with neural activity in identifiable brain areas such as the amygdala and nucleus accumbens, but are also closely related to neurotransmitters such as dopamine and serotonin. Other neurochemicals such as oxytocin, cortisol, testosterone, and estrogen can also influence emotional processing. Hence a full understanding of the emotional component of self-concepts requires taking into account mechanisms at the molecular level.

Causal relations between the individual and molecular levels with respect to self-concepts operate in both directions. Telling people that they are good-looking, nice, and successful will usually produce in them feelings of pleasure associated with increased

activity in the dopamine system, whereas insults increase cortisol levels. Going in the other direction, ingestion of drugs like opiates, stimulants, and hallucinogens produces molecular changes in the brain that can lead to self-attribution of different concepts. Hence a full understanding of self-concepts requires attention to interacting mechanisms at all four levels: molecular, neural, individual, and social. A similar case can be made for other self-phenomena (Thagard and Wood, forthcoming).

6. OBJECTIONS

The multilevel view of the self is open to objections from many directions. Some philosophers will think that I have slighted phenomenological aspects of the self - what it feels like to be you (e.g. Gallagher and Zahavi, 2008; Strawson, 2009; Zahavi 2005). Perhaps selfhood is more a matter of ongoing lived experience than the result of multiple mechanisms. My response is that qualitative experiences such as emotional consciousness are in fact amenable to mechanistic explanation, particularly at the neural level (Thagard and Aubie, 2008; Thagard, 2010a). Emphasis on raw phenomenology over mechanism encourages hanging on to transcendental views of the self as soul, or swinging in the opposite direction toward deflationary views of the self as just a series of experiences.

A second objection is that my account has neglected insights into the powerful role of embodiment and action in constituting thinking and personhood, for example in self-control. Many philosophers, psychologists, and linguists have made interesting observations about the often neglected role of the body in human thinking (see e.g. Gallagher, 2006; Gibbs, 2005), but the multilevel approach is compatible with these insights (Thagard, 2010c, 2012). Brains operate with a continuous flow of information

from inside and outside the body, and the inclusion of environment in the characterization of systems at each level shows the compatibility of my account of the self with views that understand cognition as intimately coupled with bodies and the physical and social worlds with which they interact (Legrand and Ruby, 2009). I reject, however, extreme positions that claim that dynamic embodiment shows that minds do not require mental representations, which are a crucial part of all the self-phenomena. Cockroaches have dynamic embodiment, but they lack selves. Understanding the self as a dynamic system situated in physical and social worlds requires attention to internal representational models (Ismael, 2007)

A final objection is that the account of selves as multilevel systems is terminally obscure, bereft of explanatory power. I grant that this account is very broad, but maintain that much of the details are being worked out through characterizing in detail the mechanisms at each level. At the individual level, there have been decades of work on mental representations and the computational processes that operate on them. At the neural level, investigation of the kinds of high-level cognition relevant to understanding the self is much more recent, but the past decade has brought major advances concerning how brains represent and process information. There is even a start on explaining such neuropathologies of the self as anosognosia, asomatognosia, delusional misidentification, depersonalization, and Capgras and Fregoli syndromes (Feinberg, 2009). Psychologists and philosophers have tended to ignore the molecular level, but increased focus on neural mechanisms is inevitably leading also to increased attention to molecular mechanisms.

What is most obviously lacking in current discussions is a rich, general understanding of the relations among levels. I have maintained that there are interlevel

feedback loops that account for much of the richness and unpredictability of human behavior, and have sketched how this works for self-concepts. But much more research needs to be done to better comprehend the relations among the social, individual, neural, and molecular levels. Insights from the growing field of systems biology should be useful here. Any organism is also a multilevel system, and increased appreciation of the relations among bodies, organs, tissues, cells, genes, and proteins should be help to illuminate the analogous relations among the multiple subsystems that constitute the self.

7. CONCLUSION

So who are you? My answer is that a self – a person – is a complex system operating at four levels, each of which consists of an EPIC subsystem composed of environment, parts, interconnections, and changes. Each level includes mechanisms consisting of networks of parts whose interactions produce regular changes, as summarized in figure 1. Because the interactions in these subsystems typically involve nonlinear dynamics resulting from feedback loops that magnify effects of small differences in initial conditions, the behaviors of such mechanisms are often hard to predict. In particular, the behavior of the parts at each level is typically difficult to predict from the behavior of parts at lower levels. Forecasting is made even more difficult by the existence of causal relations among levels, for example social influences on molecular changes and vice versa. Moreover, at all levels the subsystem interacts with environments that include other complex systems such as climate and ecology, each of which can have changes that are difficult to predict. My multilevel account is not yet a theory of the self, but rather a framework for developing specific theories that describe mechanisms that operate within and between levels.

The justification for adopting the multilevel system view of the self is that it is superior to alternative accounts in potentially explaining a wide range of phenomena concerning human behavior. Unlike transcendental views of the self as a supernatural soul, the multilevel view understands the self as a natural but highly complex kind of entity, like a state, university, living body, organ, or molecule. Unlike deflationary views of the self as a fiction, multilevelism maintains that a scientific concept of the self has sufficiently broad explanatory power to justify belief in selves akin to belief in atoms, viruses, fields, genes, ecologies, organizations, and other important theoretical entities posited by successful sciences. The multilevel systems approach, like that of James (1890), aims to account for both the unity of the self emphasized by Kant and the diversity of the self emphasized by Hume.

The multilevel approach to the self is both methodological and ontological. Methodologically, it recommends that understanding of the self is best achieved by developing much richer scientific accounts than currently exist of the relevant social, individual, neural, and molecular mechanisms, as well as of the interactions among these mechanisms. Ontologically, it contends that the best available theory of the self consists of the hypothesis that selves *are* complex systems consisting of multilevel interacting mechanisms. This theory will become much more rich, precise, and satisfying through discovery and integration of the relevant mechanisms.

My account of the self exemplifies an approach to the social sciences that might be called the method of multilevel interacting mechanisms (MIM). This method is implicit in various creative investigations of human behavior going back at least to the work of Herbert Simon (1962), but it has rarely been aggressively pursued. Simpler

approaches, concentrating on one level or at most two, are cognitively simpler and less professionally risky. The cost of simplicity, unfortunately, is inability to explain many of the most important aspects of human behavior, such as ongoing political conflicts, economic crises, and the nature of the self.

The above discussion should make it clear that the MIM method is neither reductionist nor holistic. It is not holistic, because I do not consider higher levels such as the social as independent from or exclusively determining what happens at lower levels. It is not reductionist, because I reject the common picture that causality moves only from lower levels up to higher. Not only do causal mechanisms operate at each level, but higher-level mechanisms can have causal influences on lower-level mechanisms. The parts at higher levels have emergent properties in the non-mystical sense that the properties belong only to parts at that level, not to parts at lower levels or to simple aggregates of those parts (Bunge, 2003; Wimsatt, 2007).

Hence multilevelism is interactive rather than mystically holistic or simplistically reductionist. The justification for this approach should not depend only on its success in making sense of the self, but also in applications to many other important human phenomena, including emotion (Thagard, 2006), creativity (Thagard and Stewart, 2011), economics (Thagard, 2010b), and culture (Thagard, forthcoming).

My concern in this paper has been narrower: to make sense of the self by considering it as a multilevel system consisting of interacting social, individual, neural, and molecular mechanisms. Thagard and Wood (forthcoming) show the relevance of all of these levels to additional important phenomena: self-presentation, self-esteem, self-enhancement, self-regulation, self-expansion, and self-development. These are

representative of three general classes (self-representing, self-efficacy, and self-changing) that cover more than eighty self-phenomena important in psychological, philosophical, and sociological discussions of the self. The self is neither simple nor fictional, but can be understood, from a sufficiently rich, multidisciplinary perspective, as a complex system.

Acknowledgements: Thanks to Joanne Wood and anonymous referees for comments on earlier drafts. This research has been supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Alicke, M. D., Dunning, D. A., & Krueger, J. I. (Eds.). (2005). *The self in social judgment*. New York: Psychology Press.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (2007). *How can the mind occur in the physical universe?* Oxford: Oxford University Press.

Aronson, E. (2007). *The social animal*. New York: Worth Publishers.

Baumeister, R. (Ed.). (1999). *The self in social psychology*. Philadelphia: Psychology Press.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Benjamin, J., Li, L., Patterson, C., Greenberg, B. D., Murphy, D. L., & Hamer, D. H. (1996). Population and familial association between the D4 dopamine receptor gene and measures of Novelty Seeking. *Nature Genetics, 12*(1), 81-84.

Blouw, P., Solodkin, E., Eliasmith, C., & Thagard, P. (in progress). Concepts.

Bunge, M. (2003). *Emergence and convergence: Qualitative novelty and the unity of knowledge*. Toronto: University of Toronto Press.

Callero, P. L. (2003). The sociology of the self. *Annual Review of Sociology, 29*, 115-133.

Churchland, P. M. (2007). *Neurophilosophy at work*. Cambridge: Cambridge University Press.

Churchland, P. S., & Sejnowski, T. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*, 547-663.

Darden, L. (2006). *Reasoning in biological discoveries*. Cambridge: Cambridge University Press.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown.

Eliasmith, C. (in press). *How to build a brain*. Oxford: Oxford University Press.

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

Feinberg, T. E. (2009). *From axons to identity: Neurological explorations of the nature of the self*. New York: Norton.

Gallagher, S. (2006). *How the body shapes the mind*. Oxford: Oxford University Press.

Gallagher, S. (Ed.). (2011). *The Oxford Handbook of the Self*. Oxford: Oxford University Press.

Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. London: Routledge.

Gibbs, R. W. (2006). *Embodiment and cognitive science*. Cambridge: Cambridge University Press.

Hume, D. (1888). *A treatise of human nature*. Oxford: Clarendon Press.

Ismael, J. T. (2007). *The situated self*. Oxford: Oxford University Press.

James, W. (1890). *The principles of psychology*. New York: H. Holt.

Legrand, D., & Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review, 116*, 252-282.

Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology, 51*, 858-866.

McCauley, R. N., & Bechtel, W. (2001). Explanatory pluralism and the heuristic identity theory. *Theory & Psychology, 11*, 736-760.

Mead, G. H. (1967). *Mind, self & society from the standpoint of a social behaviorist*. Chicago: University of Chicago Press.

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York: Basic Books.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

Organ, T. W. (1987). *Philosophy and the self: East and west*. Selinsgrove, PA: Susquehanna University Press.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*(7045), 1102-1107.

Richards, E. J. (2006). Inherited epigenetic variation--revisiting soft inheritance. *Nature Reviews Genetics, 7*(5), 395-401.

Siderits, M., Thompson, E., & Zahavi, D. (2011). *Self, no self?* Oxford: Oxford University Press.

Simon, H. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society, 106*, 467-482.

Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive psychology: Mind and brain*. Upper Saddle River, NJ: Pearson Prentice Hall.

Strawson, G. (2009). *Selves: An essay in revisionary metaphysics*. Oxford: Clarendon Press.

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

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Thagard, P. (2008). How cognition meets emotion: Beliefs, desires, and feelings as neural activity. In G. Brun, U. Doguoglu & D. Kuenzle (Eds.), *Epistemology and emotions* (pp. 167-184). Aldershot: Ashgate.

Thagard, P. (2010a). *The brain and the meaning of life*. Princeton, NJ: Princeton University Press.

Thagard, P. (2010b). Explaining economic crises: Are there collective representations? *Episteme, 7*, 266-283.

Thagard, P. (2010c). How brains make mental models. In L. Magnani, W. Carnielli & C. Pizzi (Eds.), *Model-based reasoning in science and technology. Abduction, logic, and computational discovery* (pp. 447-461). Berlin: Springer.

Thagard, P. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Thagard, P. (forthcoming). Mapping minds across cultures. In R. Sun (Ed.), *Grounding social sciences in cognitive sciences*. Cambridge, MA: MIT Press.

Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition, 17*, 811-834.

Thagard, P., & Stewart, T. C. (2011). The Aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35*, 1-33.

Thagard, P., & Wood, J. V. (forthcoming). Eighty phenomena to be explained by a theory of the self. *Unpublished, University of Waterloo*.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge, MA: Harvard University Press.

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge, MA: MIT Press.