Dietrich, E. (2000). Analogy and conceptual change, or You can't step into the same mind twice. In E. Dietrich and A. Markman (eds.) *Cognitive Dynamics: Conceptual change in humans and machines*. Mahwah, NJ: Lawrence Erlbaum, pp. 265 – 294.

Analogy and Conceptual Change, or You can't step into the same mind twice

Eric Dietrich

Philosophy Department

Binghamton Univ., Binghamton, NY, 13902

1. Introduction.

Sometimes analogy researchers talk as if the freshness of an experience of analogy resides solely in seeing that something is like something else -- seeing that the atom is like a solar system, that heat is like flowing water, that paint brushes work like pumps, or that electricity is like a teeming crowd. But analogy is more than this. Analogy isn't just seeing that the atom is like a solar system; rather, it is seeing something *new* about the atom, an observation enabled by 'looking' at atoms from the perspective of one's understanding of solar systems. The question for analogy researchers then is this: Where does this new knowledge about atoms come from? How can an analogy provide new knowledge and new understanding?

My answer is that having an analogy *changes* the concepts involved in the analogy. More specifically, merely having an analogy changes one's concepts. I call this answer the *analogical conceptual change hypothesis* . In this paper, I argue for this hypothesis and explain some of its implications. I have to argue for this hypothesis more or less from first principles, because, as a psychologist colleague pointed out to me, it isn't clear how to test the hypothesis experimentally, at least not right now. This is unfortunate, not just because it means the hypothesis remains untested, but because psychologists have a tendency to lose

interest in ideas that aren't subject to experimental verification or refutation. So, for better or for worse, this is a paper in what we might call *theoretical psychology* .[1]

The next two sections of this paper present needed background, first on analogy and analogical reminding and then on conceptual change and its dynamics. As I lay out this background, I use it to elaborate the analogical conceptual change hypothesis and specify the kind of concept change which I think must occur in order to have analogies (more specifically, to have analogical remindings). Then in the next three sections, I use the traditional tools available to us theoretical types -- logic, plausible assumptions, and others' data -- to argue 1) that probability assessments indicate that the specified conceptual change must occur *before* one can experience an analogy, 2) that the notion of mapping crucial to the theory analogy (defined below) camouflages a serious unpaid theoretical debt, and 3), that paying the debt from 2) while obeying the probability assessments from 1) requires a view of concepts where the types of constituents which make up concepts are not fixed, but can transform into one another rapidly, especially during analogical reminding.[2]

2. Analogy and Analogical Reminding.

2.1 The General Picture and Definitions of Terms

The cognitive phenomenon I am primarily interested in is reminding: specifically, *analogical reminding*. (Sometimes researchers, e.g., Hummel and Holyoak, include analogical reminding as part of the general definition of analogical thinking. See their 1997.) Analogical reminding is common; it occurs any time some concept or percept in one domain recalls, in the right way for an analogy, another concept in another domain. For example, imagine that while walking down a sidewalk one night, you see a jumble of garbage cans, some standing upright, some lying on their side or against each other, and all of a sudden you are reminded of Stonehenge on the Salisbury Plain in England. Here's another example. I was cross-country skiing with a colleague. We paused to rest and drink some water. Though it was cold, we were quite warm and she, being mindful of hypothermia, took off one glove to cool down. She then quipped: "My hand is like a dog's tongue when he's panting in the summer."

Such occurrences are the sort of the phenomena I shall be concerned with. I am not here interested in the phenomenon of a hearer understanding an analogy spoken to her. I am not primarily interested in linguistic analogies, but rather the analogies that occur spontaneously in one's head during cases of reminding. I shall be concerned also with long-term and working memory.

A fair amount is known about analogical reminding and the broader class of *similarity-based retrieval* to which analogical remindings belong. Analogical reminding comprises at least two processes: *access* and *mapping* (Forbus et al., 1995; Hummel and Holyoak, 1997). Access (also simply called "retrieval") is the process of retrieving some memory item (which I will call the *retrieved item*) from long-term memory based on some other item in working memory (which I will call the *retrieving item*). After retrieval the retrieved item and the retrieving item co-exist in working memory.

Before defining mapping, I need to say a few words about my terminology. Generally, I use the term "item" to refer to anything in either long-term or working memory. This gets around the problem of worrying about when a memory element is a concept and when it is not. I will still use the term "concept" when the item referred to is obviously, or traditionally treated as, a concept. This situation most frequently happens when the item is in working memory. However, unlike some psychologists, notably Barsalou, 1989, I will not adhere to the restriction of using "concept" exclusively to refer to items in working memory.

I also assume that concepts and, in general, items, are *representations* of some sort. So for example, in the garbage-cans/Stonehenge case, the perceptually-based representation of the garbage cans in working memory accessed the Stonehenge representation in long-term memory. So the representation of the garbage cans is the retrieving item, and the representation of Stonehenge was the retrieved item. The item in working memory can be a perceptual one, like the representation of the garbage cans, or it can be an item previously retrieved from long-term memory -- as in the stream-of-consciousness phenomenon (for

example, once the memory item representing Stonehenge was in working memory, it might have then accessed the memory of the stone faces on Easter Island).

Mapping is the process of matching constituents of the two items now in working memory. Mapping is essentially a process of finding *functional counterparts* between concepts (see, e.g., Gentner, 1983, 1989; Hummel and Holyoak, 1997). The mapping process locates which *object nodes* (or more simply, "objects") in one concept are the functional counterparts of object nodes in the other concept. For example, in garbage-henge, the representations of the cans are the object nodes, and they map onto the representations of the stone monoliths. In the ungloved-tongue case, the representation of the ungloved hand maps onto the representation of a dog's panting tongue. What makes these objects functional counterparts of each other is the role they are represented as playing in the concept or representation. These roles are represented by *structural relations* among the objects making up the concept. The mapping process ignores *attributes* of objects. Attributes are representations of *properties*, which occur in the external world. For example, dog tongues are wet. Being wet is the property in the world, and internally it is represented by an attribute designated by something like "being wet."

From this preliminary discussion of mapping, we can see the three main constituents making up memory items: objects (object nodes), attributes, and structural relations. These three represent three different parts or aspects of the world: physical or non-physical things, the properties of these things, and the functional roles these things can partake in.

Mapping is a very important notion; I will return to it shortly when I discuss the nature of analogy in more detail. In section 5, I will discuss an interesting and important problem with the notion.

Analogical remindings are individually generated, occurring in the heads of most humans past a certain young age (there is evidence that very young children can recognize analogies provided that their knowledge is manipulated and changed in the appropriate way, see Kotovsky and Gentner, 1996, and Gentner et al. 1995). Analogical remindings are

frequently quite creative and are therefore implicated in theories of creativity (Finke, et al.; Hofstadter, 1995).

Gentner and her colleagues point out that, broadly speaking, there seem to be three large classes of remindings: 1) analogical remindings (e.g. Rutherford noticing that the alpha particles in his experiments were like comets), 2) superficial remindings (like when a yellow balloon reminds you of the sun -- these are sometimes called "mere appearance" or "attribute-similarity remindings"), and 3) mundane remindings (like when garbage cans remind you of other garbage cans or remind you to put out your garbage for tomorrow morning's pickup -- these are sometimes called "literal similarity remindings") (Forbus et al., 1995; Gentner, 1989). They have a theory that explains, in part, the relative frequencies of these three kinds of reminding. According to the experimental evidence, types 2 and 3 (superficial and mundane remindings) are the most common; type 1 is rarer. I do not think that type 1 remindings are as rare as the Rutherford case might lead you to believe. You don't have to have an insight into particle physics to experience a creative, if quirky, analogical reminding. The garbage-henge and ungloved-tongue cases are just such examples. (Gentner and her colleagues might categorize the garbage-henge case as a superficial reminding, rather than a true case of analogical reminding, but I don't think this is right. I will return to this below when I discuss the properties of true analogies.)

I can fine-tune the analogical conceptual change hypothesis a bit, now. The hypothesis predicts that type 1 remindings alter one or both of the items involved in the episode of reminding. (It would be worth considering the extent to which the other types of reminding alter concepts and other memory items, but that is a task for another paper.)

2.2. Analogy and Structure-Mapping

What makes a reminding an *analogical* reminding is simply that it is a retrieval of an item from long-term memory that results in an analogy with the item doing the retrieving. Analogy is the cognitive process whereby one thing is seen as resembling another. But what does it mean for two concepts to resemble one another, to be similar? This is a deep question. Answering it requires having a theory of analogy and, at least, the beginnings of a theory of

concepts. My answer to this question is derived from Gentner's Structure-Mapping Theory (1983, 1989). I assume her theory for two related reasons. First, as I said, it is really not possible to define analogy beyond a sort of folk definition without appealing to some theory or other, and, secondly, her theory, the central part of it anyway, has more or less achieved the status of "the received view."

On Gentner's Structure-Mapping Theory of analogy, two memory items (concepts) are analogous when one is *mappable* to the other (e.g., if the working memory item *maps* on to the item retrieved from long-term memory). Mapping, as I said, is a process of finding functional counterparts between the two concepts. This process has three parts. First, the objects of one item must map onto the objects of the other item. Consider the well-known analogy between the atom and a solar system (see figure 1). The analogy maps representations for planets onto representations for electrons, and a representation of the sun onto a representation of the nucleus. Second, the two memory items must have the same *structure* for an analogous mapping to be successful. Having the same structure means that their higher-order relations are *identical* . In an analogy, it is these relational structures of the concepts that matter, not the lower-level properties or attributes of the objects. So, third, low-level properties or attributes must be discarded for purposes of the analogy. For example, that the sun is yellow and hot is irrelevant to the analogy, and so the mental representations of these attributes are not a part of the mapping. The analogous concepts needn't share object attributes, and usually won't share any substantive attributes, i.e., attributes beyond things like "physical object". Together, these three parts mean that objects in analogous concepts are represented as *purely* functional counterparts.

So in analogy, both high-level structures and objects are mapped; attributes or properties are not. There is an infelicity, however, in this use of the term "mapping." When analogy researchers speak of mapping structures, the structures have to be identical (at some level). When they speak of mapping objects, the objects *cannot* be identical -- otherwise what would be the point of the analogy? So, mapped objects are not identical, but mapped structures are. This infelicity is not that important in itself, but it does indicate that mapping tends to be under-specified and treated rather loosely in theories of analogy. I will discuss

mapping and these problems in section 5. Since the term "mapping" is the accepted technical term for both kinds of matching, I will use it for both, but the difference between the two should definitely be borne in mind. When it matters to the discussion, I will flag the difference.

At this point, it will be beneficial to step through some simple examples. Thinking that a yellow balloon is like the sun is not an analogy (but it is a similarity comparison) because the similarity of the two (the balloon and the sun) is based on the property of being yellow. Being yellow is not a relation ("yellower than" is a relation, but it is not operative in this case). Thinking that an ungloved hand is like a dog's tongue is an analogy because the similarity between the two is based on the complex relation: "exposed body part causes heat dissipation." Not only is this a relation, but it is a *higher-order* relation between a property (being an exposed body part -- a tongue or a hand) and an event or process (losing or giving off heat). (The fact that a flat hand also resembles a flat tongue was probably important for the retrieval, too, but not the analogy. See Forbus, et al.) Garbage-henge is arguably an analogy (a case of analogy reminding) and not a case of mere superficial similarity reminding because the similarity between the garbage cans and the monoliths of Stonehenge was based on relations such as "lies next to," "stands next to," "lies athwart," and "is leaning on," and not merely simple properties describing the whole collection of objects such as "lies in a semi-circle."

It is clear from these examples that concepts can match or be similar in three ways which correspond directly to the three classes of remindings (see above). The three kinds of similarity are analogical similarity, superficial or mere-appearance similarity, and literal similarity. Analogical similarity, as we have seen, results from structural and object matches, but not attribute matches. Superficial similarity results from object and attribute matches only. And literal similarity matches result from structural, attribute, and object matches.

Structure Mapping Theory also postulates conceptual change. After the analogy between the two analogues has been made, information can be imported from the retrieved item to the retrieving item. This is usually the order because the retrieved item is the one the individual "knows best;" it is the richer one from which knowledge can be imported to the

retrieving item. This process is called *projection of candidate inferences* (Gentner, 1983, 1989, this volume). The wider theory also postulates three other kinds of conceptual change (for a total of four kinds):

1. Progressive Alignment, whereby children's knowledge becomes more abstract so that more high-order similarities can be recognized (this is also sometimes called "unpacking" and "gentrification of knowledge", Kotovsky and Gentner, 1990 and 1996; Gentner, et al., 1995)[3]
2. Highlighting, whereby less salient conceptual properties are made more salient (this volume),
3. Restructuring, whereby whole systems of knowledge get changed (this volume).

It is important to note that all of these changes happen *because* of analogy. In contrast, the analogical conceptual change hypothesis claims that analogy happens *because of* conceptual change (of a certain sort, to be explained below). The analogical conceptual change hypothesis does not deny that there are the sort of post-analogy conceptual changes hypothesized by Gentner. My hypothesis agrees with Gentner on this point: her four kinds of change happen *after* an analogy has been made. Rather, my hypothesis claims that a specific kind of change occurs before the analogy, and that the analogy happens because of this change.

2.3. Analogical Reminding and MAC/FAC

In addition to their theory of analogy, Gentner and her colleagues have a theory and computer model of similarity-based retrieval called MAC/FAC. MAC/FAC stands for "Many are called but few are chosen." (When there is no chance of confusion, I will use the term "MAC/FAC" to refer both to their computer program and their theory of analogical reminding.) The program MAC/FAC incorporates within it a computer model of Gentner's Structure Mapping Theory called the Structure Mapping Engine (see Falkenhainer et al. 1989). MAC/FAC is not intended as a model of reminding in general. It is strictly a model of similarity-based remindings. (And, it is not the only such model. See, e.g., Thagard, et al. (1990), and Hummel and Holyoak, 1997.)

MAC/FAC explains three interesting facts.

Fact one: It explains the observed ratios of the three types of remindings people experience. As I mentioned above, type 1 is rarer than both type 2 and type 3, with type 3 being the most common.

Fact two: It explains why similarity-based retrieval is strongly sensitive to superficial similarity and only weakly sensitive to structural similarity.

Fact three: It explains why high-level relational similarity is a better predictor than surface similarity of how useful a reminding is in terms of making further inferences. This means that analogical remindings are more useful than mere appearance remindings for making inferences. (Mundane remindings are quite useful too, but, importantly, since they are literally similar, they do not usually generate *new* knowledge. Analogies are best for that.)

Briefly, here is MAC/FAC's explanation of these three facts. As we noted above, type 3, the mundane remindings, are retrievals based on literal similarity between the retrieving item and the retrieved item. That is, these remindings are based on matches of both relational structures and attributes. So type 3 remindings are based on the highest quantity and quality of matches. This is why garbage cans tend to remind you of other garbage cans. Type 2 remindings, the superficial ones, are the second most common. This is because they are the easiest to make. Matching attributes and other low level features requires computationally simple operations, like taking the dot product of feature vectors. Finally, analogical remindings (type 1) do occur from time to time because (in part), though they are expensive, they are the most useful in terms of making analogical inferences and importing new knowledge.

All of this suggests that similarity-based reminding is a two-staged process. Stage 1, the MAC stage is sensitive primarily to surface similarities. Retrieval based on surface similarities is easy and cheap, so the MAC stage is relatively fast. On the other hand, the MAC stage is relatively *insensitive* to high-level, relational similarities. This is not so good because relational similarities, not surface similarities, between memory items are the foundation of good inferences and new knowledge. What is needed is another stage that *is* sensitive to relational similarities. So, the MAC stage functions as a wide filter producing a set of potential remindings as output which get passed on to the FAC stage (which is where SME is located in the MAC/FAC program). The FAC stage is slower and more computationally expensive because it is sensitive to high-level relational structure. But this isn't a problem because, thanks to the MAC stage, it is working over a much smaller set than all of the system's memory.

The interaction of the two stages tends to produce remindings based on both high-level and superficial matches -- type 3 remindings. This accords with the experimental data. Though the FAC stage uses SME, it is not a stage just for analogy matching. If it were, then MAC/FAC would produce mostly analogical remindings, which wouldn't fit the data. Rather, SME is designed so that, while running within MAC/FAC, it tends to produce mostly mundane remindings (based on matches of both relational structure and attributes), yet it still retrieves analogies from time to time (based on relational matches only).[4]

I will *not* be assuming Gentner's MAC/FAC theory of reminding for the reason that my analogical conceptual change hypothesis is incompatible with it (though I will pay attention to the data MAC/FAC are meant to explain).

My hypothesis, to refine it further now, claims that *access* (or retrieval) changes the items involved. And *the change is to the relational structure of the analogous items*. I call this change, *retrieval-based structural change* . It is part of my hypothesis that by the time both items are in working memory, one or both have been changed structurally (if only slightly), and it is this change which funds the analogy -- or better, this change *is* the analogy.

Retrieval-based structural change simply doesn't happen in MAC/FAC. And it does seem that such conceptual change runs counter to some rather deep architectural features in the MAC stage. However, it is not clear to me that retrieval-based structural change is *inconsistent* with the central feature of MAC/FAC: that memory retrieval is a two stage process, one fast and insensitive and the other slow but sensitive. True, MAC/FAC is a model based on collected psychological data, but the data I am discussing here, e.g., the garbage-cans/Stonehenge case, are data over and beyond what MAC/FAC was intended to model. It seems to me that, if my hypothesis is correct and there is such a thing as retrieval-based structural change, MAC/FAC could be altered slightly and augmented without radically altering the general MAC/FAC approach to reminding. Some of the detailed architecture of MAC/FAC would have to be changed if I am right, but not anything central to it.

3. The Mutability of Concepts.

The general idea that concepts change over time is not new. It has been explored by many. Indeed, the idea is venerable. Henri Poincare and William James both hypothesized that concepts were active and nonstatic (Poincare, 1952; James, 1890/1950). In fact, the thread of the idea that concepts are fluid and constructed goes clear back through the Roman materialist philosopher, Lucretius, to the ancient Greek philosopher Heraclitus who said "the moving world can only be known by what is in motion" (Frag. 43).[5]

To a first approximation, the cognitive dynamics of retrieval-based structural change depend on *interacting concepts* . It's the interaction that produces the change. Since concepts are how we conceive things, this squares nicely with the analogical conceptual change hypothesis: the reason the atom - solar system analogy shows us something new about atoms is that it changes the concept representing atoms. But care must be taken here because the interacting concept view suggests that concepts exist ahead of time in long-term memory as static memory items, and it is not obvious that this is true -- to put it mildly (see, e.g., Barsalou, 1989).

There is a fair amount of agreement, at least among psychologists, that long-term memory items are not at all like classical data structures, inertly sitting in one's head waiting to be read or updated. Beyond this, however, there is not much agreement. How stable are the items in long-term memory? Are items in long-term memory retrieved as units? Does an organism retrieve an item from long-term memory, or does it construct the item from something "subconceptual" in long-term memory? When an organism retrieves information (to use a neutral term) from long-term memory at one time, and retrieves the same information (in some sense) at another time, is the resultant working memory item the same both times? If not, what influences the change? (Many cognitive scientists have wrestled with these questions. Barsalou has done an especially interesting job. See his 1983, 1987, 1989. The questions I asked are derived from his 1989).

I have only partial answers to some of these questions. I encapsulate my answers in the following three assumptions, which seem plausible given the data.

Assumption 1. The items in long-term memory are more or less stable (but not static). Think of them as "chunks of knowledge" which might be either conceptual or something subconceptual out of which concepts are built. (We don't have enough information at this time to decide this issue.)

Assumption 2. Items in working memory which come from long-term memory do *not* get there via simple retrieval, like getting a book off a shelf. Instead there is some sort of construction process going on. (This seems to me to be the minimal assumption need to explain data on concept flexibility such as Barsalou's.[6])

It is important to note that assumption 2 does *not* entail that items in long-term memory are subconceptual, though they might be, in fact. The construction process I refer to is one *from* long-term to working memory. So, for example, both working memory and long-term memory items could be something we could reasonably regard as concepts, but the working memory concepts might be assembled from a variety of different long-term memory concepts. In other words, how you conceptualize the world in working memory need not be how you remember the world

in long-term memory. I will still continue to use the term "retrieval" to describe getting information from long-term to working memory. But this process should not be understood as a simple find-and-fetch.

Assumption 3. Items (concepts) in working memory can interact with each other and thereby change each other.[7] However, this is not the kind of conceptual change hypothesized by the analogical conceptual change hypothesis.

Together with the above three assumptions, the analogical conceptual change hypothesis makes the following three claims:

Claim 1. The very process of analogical reminding alters our concepts. Specifically, it alters their high-level structure in some way -- perhaps by constructing new abstractions. (Further research is needed to figure out the details of this structural change.) In any case, how we conceive of the world is thereby altered (however slightly, and perhaps only temporarily). Either item, the retrieving item or the retrieved item or both might be changed. This is what I called retrieval-based structural change. This conceptual change is arguably the central reason why analogical reminding is an important cognitive process and why it is creative.

Claim 2. The kind of conceptual change hypothesized in claim 1 happens at the time of retrieval.

Claim 3. The order of events is this: 1) During an episode of reminding, an item in working memory (a concept, usually) interacts with items in long-term memory (chunks of knowledge) attempting to retrieve at least one of them; 2) during the interaction one (or both) changes in some way; 3) in most cases, mundane remindings occur, but if the change is of the right sort, i.e., if it allows for a mapping of high-level relational structure (which might be new), an analogy results; 4) the person who had the analogy sees something new where he didn't before because his memory items have

changed, if only a little. 5) Events can now proceed as described in Structure Mapping Theory; for example, projection of candidate inferences can now take place.

These claims center around *when* a certain kind of conceptual change occurs. The change I am interested in occurs before the analogy is formed, and in fact leads to the analogy being formed. In the next section I give my argument for this conclusion.

4. The Low Probability Argument

Analogy is mapping objects and relational structures. Analogical reminding is retrieving items with the requisite mappable structures. Two questions need answering: A) Why do the structures match each other? B) What is mapping? In this section, I address the first question; in the next section, I address the second.

Question A) is really a question about timing. It can be re-asked this way: do the two structures match *before* the analogy or do they match *after* (and because of) the reminding? There are, accordingly, two answers to this question: 1) the structures match before the analogy, and 2) the structures do not match ahead of time but are built somehow by the process of analogical reminding itself.[8] MAC/FAC and Structure Mapping Theory (as well as several other theories of analogy and analogical reminding), assume the first answer -- the structures are there ahead of time (e.g., see Falkenhainer, 1988 (e.g., p. 59), Gentner, 1989, p. 213; 1983, p. 158; Gentner and Wolff, this volume; Hummel and Holyoak,1997; and Kotovsky and Gentner, 1990). For example, the reason atoms remind one of solar systems is that the mental representations for each have the same high-level structure, and both representations had that structure before the analogy occurred. I am going to argue that answer 2) is the better answer by arguing that answer 1) is implausible.

Before I get to my argument, though, I need to discuss a couple of matters. The first is that if MAC/FAC assumes that structures are mapped because they match ahead of time, then isn't MAC/FAC committed to the view that long-term memory items are fixed, static things

which are retrieved via a simple find-and-fetch operation? As a matter of fact, MAC/FAC does assume just this. The MAC stage, specifically, assumes that long-term memory items are basically concepts that are simply retrieved and placed in working memory (Forbus, et al., 1995). However, it isn't clear to me that MAC/FAC has to assume this. The essence of MAC/FAC seems to be compatible with the three assumptions about memory I made in section 3.

The second matter is a possible source of confusion. One might suppose that any theory that adopts answer 1) is going to be bedeviled by the question: "If the representation for atom already looks like the one for solar system, what is the point of the analogy in first place?" But it would be a mistake to suppose this. For example, within Structure Mapping Theory and MAC/FAC, an analogy allows information, in the form of other predicates, to transfer from one analogue to the other (from the retrieved item to retrieving item). These are called candidate inferences and were discussed in section 2.2. Also, merely knowing that, e.g., atoms are like solar systems, is itself useful. For example, one might construct a new category by describing them both using a unifying notion such as "central force systems" or some such. Moreover, as we also noted in section 2.2, one source for the emergence of structure *is* explained within an extension of Structure Mapping Theory. This extension explains the emergence of relational structure across developmental time from a younger child to an older one, and postulates a process called progressive alignment (Gentner, et al., 1995 and Kotovsky and Gentner, 1996). So even if the structures match ahead of time, there is still something for analogy and analogical reminding to do.

Nevertheless, the first answer should be abandoned. What are the chances the structures of the two memory items resemble each other ahead of time, before the analogy has occurred? It must be quite low -- too low to explain the quantity of analogical remindings that occur in each of us. In short, it seems completely implausible, in the usual case, that the relational structures of the analogues would antecedently match. For example, again suppose you see some overturned, jumbled garbage cans by the curb and are reminded of Stonehenge. Note first that it is highly unlikely that the jumble of cans matches the jumble of monoliths on the Salisbury Plain. But this means that it is highly unlikely that your percept formed by

seeing the cans matches the part of your concept of Stonehenge representing the pattern of the stones seen from a certain perspective. In fact, it is unlikely that your perception of the jumble of the cans antecedently matches even decayed, partial memory of how the stones are arranged at Stonehenge.

If this is right, then since the high-level, relational structures of the two analogues don't antecedently match (except in very rare cases), but they do match at the time the analogy is made, it must be that the high-level structures are constructed at the time of the reminding.

This is what I call the "low-probability argument." It is a plausibility argument; it is not intended to have the force of a theorem in mathematics. I now turn to defending it against some objections. Doing this will also allow me to elaborate it some.

Objection 1: The probability of analogues antecedently matching isn't that low. There are constraints on perception and memory such that the way we perceive and store information guarantees that some items are bound to match other items.

Reply: This objection amounts to a bare assertion that what seems to me to be *im*plausible, is in fact plausible. It is unclear what these constraints appealed to might be, and without a good story about them, we just have dueling assertions based on dueling intuitions. The low probability argument is crucial to my claim that retrieval causes changes in concepts or memory items. Hence, I am willing to give up the low probability argument only if a good explanation is offered as why the analogues have antecedently mappable structures. Since no compelling explanation is currently on the table, and since my intuition still seems the most plausible, I will stick with it.

The objection also clashes with another intuition several cognitive scientists have (me included): concepts are, in many ways, quite plastic and malleable; we can see analogies between all kinds of things. It seems implausible therefore that perception and memory could both support that kind of plasticity while maintaining relational structures that match ahead of time. There are just too many ways two things might be analogous. That we store all those

ways ahead of time seems unlikely. It seems more likely, and even more efficient, that reminding produces the changes in real time.

Objection 2: I've focused on the wrong probability. Consider garbage-henge again. Though the probability is low that your current perceptual image of the garbage cans *identically* matches your imperfectly remembered perceptual image of Stonehenge, this isn't the relevant probability. The relevant probability is the one measuring the likelihood that your garbage can percept was *merely similar* to your Stonehenge concept. This probability might be quite high, high enough to explain the common occurrence of analogical remindings.

Reply: But what does "merely similar" mean? Unless one reduces similarity to identity at some point, one gets an infinite regress of similarities: X is similar to Y because a feature or aspect of X is similar to a feature of Y and these features are similar because their features in turn are similar, etc. This explains nothing. The notion of similarity without identity in some form is vacuous. (Gentner was the first to make this point in this context. See her 1983, fn. 6. In fact, this objection amounts to rejecting Structure Mapping Theory.)

Consider five strings of characters:

a) a s d f g h j k l

b) q w e r t y u i o

c) q w e r x y u i o p

d) z c v b n m

e)

Letterwise, string a) is not similar to any other string. Strings b) and c) are similar because some they have identical substrings. However, one could say that strings a) and b) are similar because they have the same number of letters (a bit of information about both strings that is implicitly represented -- to draw it out requires abstracting). This is perfectly legitimate, but notice, this requires that the two strings have *identical* cardinalities. String d) is not similar to any of the others preceding it (even using cardinality) unless one says that it, like the others is made up of letters from the English alphabet. Again, a perfectly legitimate move, but one that requires saying that strings a), b), c), and d), were all drawn from *identical* alphabets.

Finally, try to imagine what it could mean to say that a string was similar to one of the first four strings without some feature of it being identical to one of their features. Imagine a fifth string, string e). Suppose that it shares no identical features with any of the other four strings, but that it is nevertheless similar to, say, string a). Neither its subparts nor any of its abstractions are identical to string a), nevertheless it is similar to a). What does string e) look like? I cannot think up such a string, and I believe this is because there is no such string.

These observations illustrate a general principle: similarity must reduce to identity of some aspect or other. If this is right, then there is no such thing as analogical retrieval based exclusively on the two items being "merely similar". The retrieving item and the retrieved item might in fact be similar, but that is because some feature of the two is identical. In analogy, this feature is the relational structure of the objects.

Objection 3: But there is empirical evidence that retrieval is governed primarily by surface similarity or commonality. This evidence is in fact one of the reasons for the MAC stage in MAC/FAC. So retrieval *is* based on similarity and not identity.

Reply: The MAC stage of MAC/FAC assumes identity. It assumes that two memory items are candidates for analogical mapping based on an estimate of their structural similarity. This estimate is the dot product of their content vectors. And the dot product multiplies identically matching vector components. So identity is crucial to MAC/FAC.

This objection is actually based on an ambiguity in the word “similarity.” When Gentner and her colleagues use the term, they don’t mean “similarity without identity,” rather they mean “similarity because of identity.”

Objection 4: Conceding then that the relevant high-level structures probably don’t match ahead of time, why does the conceptual change happen at the time of the *reminding*? Isn’t it more plausible that it isn’t the reminding that causes the construction of the matching high-level structures, but rather the analogy itself?

Reply: Analogical remindings happen very quickly. This objection requires that the reminding occur, and then the change associated with the analogy, and then the analogy, i.e., the mapping. It seems implausible that there is enough time for this to occur. It seems more plausible that the very process of retrieving items from long-term memory alters the items retrieved. The alteration or change might be very slight, nevertheless, it does seem likely that such change occurs.

However, since the nature of mapping is up in the air, I concede that the mapping process itself might include a process responsible for changing memory items. This would complicate the analogy process, but it might be correct, and it is certainly worth exploring. If this objection were correct, my central point would remain however: high-level structures don't antecedently match, so they are changed at some point during the process of analogical reminding.

Objection 5: Aren't I ignoring the data which gave rise to MAC/FAC?

Reply: No. The data are that retrieval is most sensitive to surface matches, relatively insensitive to high-level structure, but that analogies are nevertheless based on matching and mapping high-level structure. The analogical conceptual change hypothesis, retrieval-based structural change, and the low-probability argument are all compatible with this data. My hypothesis is *not* the claim that an item in working memory can retrieve just any item it wants from long-term memory merely by transforming the latter's high-level structure. It seems likely that making such changes costs in energy, time, and/or space. Analogical reminding could therefore be relatively expensive -- more expensive than mundane and mere-appearance remindings (even though mundane remindings match at the structural level, too, there isn't much to change here because the structures literally match). So, retrieval could be a process of probing long-term memory, attempting to construct several different items in parallel based on the retrieving item, and then retrieving the one that is most easily changed to match the retrieving item. It could quite often be the case that changing a long-term memory item is too expensive relative to retrieving some item that amounts to surface reminding or mundane

reminding. This seems even more likely if goals for accessing long-term memory are factored into the retrieving process.

Here's where we are. I have used the low probability argument to argue that it is unlikely that the retrieving item and the retrieved item have matching high-level, relational structures prior to the reminding. Assuming this is correct, then since an analogical reminding includes some sort of identity mapping between the structures of the retrieving item and the retrieved item, it must be that the process of analogical reminding itself changes this structure of one or both of the two items involved. I have argued the it is the retrieval process itself that is responsible for the changes, but it could be the mapping process provided that the mapping process was made more complicated. The changes involved might be slight, not permanent, and in the usual case too expensive to complete before a more ordinary reminding is completed, but once in a while the changes are completed and we experience an analogy, which could be either quirky or sublime, but is always interesting.

5. Mapping and The Paradox of Analogy

The analogical conceptual change hypothesis amounts to the claim that retrieval-based structural change must occur if reminding is to produce analogies from time to time, which it clearly does. The argument for this claim is based on plausibility assessments that the probability of memory items matching antecedently is too low -- at least in the general case. Conceptual changes, therefore, must occur as part of the reminding process, and therefore, reminding is constructive and concepts are quite mutable.

So, we know *why* the changes occur, but we don't yet know in detail *what* changes (beyond saying it is structure). To work toward an understanding of what changes, in this section, I will argue that a natural interpretation of the notion of mapping leads to the conclusion that analogy is *impossible* . In the section 6, I offer some changes to our notion of concepts and analogical reminding that solve this problem.

(The important question of *how* the changes are produced will be left for another time, mainly because I don't know how they are produced, but see Oshima (1996) for some interesting speculation on this question. Also, though we do know why the changes occur -- the memory items don't match ahead of time -- we only now the answer to this question in proximal, shallow way. A deeper question is: Why do retrieving items try to transform retrieved items in the first place? That is, why do analogical remindings occur at all? I suspect the answer to this question lies in the realm of the evolution of cognition, and will probably have to appeal to the notions of exaptation, situated action, and the fact our ancestors couldn't draw as many distinctions as we do. Briefly, the explanation might go something like this. Assuming situated action is the best explanation for low-level perceptual and motor abilities (a big assumption), it is reasonable to infer that the question of how to explain higher cognition would also benefit from a situated action approach. This requires postulating that concepts interact with each other (since that is all that is available for any organism capable of higher cognition). The move here amounts to modeling conceptual interaction as a sort of *perspective shift* -- an *inner* perspective shift. Since, in general, it's in our survival interest to see such relations in the world as there are, the most advantageous way for concepts to interact is to attempt to change each other to highlight their similarities. Voila, analogy. Of course, this is pure speculation at this point, but it does make a certain amount of sense. For more details see Dietrich and Fields (1996) and Dietrich et al. (1996).)

Now to the paradox. Imagine once again that you are walking down the street at night and see some garbage cans strewn about. Suddenly you *see* Stonehenge right there on the curb and spilling out into the street. Such things rarely occur, which is good, since Stonehenge is on a plain in Salisbury, England, and not on your curb. But why don't such weird things occur? On a plausible interpretation of the mapping operation (explained below), you ought to see Stonehenge on the curb. But you don't. So we have a paradox. I call it the *paradox of analogy*. And it needs to be explained away.

I phrased the paradox in terms of *seeing* Stonehenge on your curb for dramatic effect. Of course, retrieving your Stonehenge memory fully is not enough to get you to see Stonehenge on the curb. Retrieving memories does not produce hallucinations, usually. The

technical point is this: on a natural interpretation of the notion of mapping (that it is activation), your entire Stonehenge memory -- objects, relations, *and* attributes -- ought to be retrieved and activated in a case of "analogical" reminding, but it isn't. . . .Why? Any memory retrieval involving mapping always ought to result in *complete* retrieval of a concept. And so there ought to be no such thing as analogy. But there is. This is our paradox and this is the matter to which I now turn.

The usual response to the paradox is to re-invoke the notion of mapping and say that in an analogy only objects and structures get mapped. Mapping, as we know, is defined as a structure-sensitive comparison. So, mapping finds the invariant relational structure between the object nodes of two memory items (refer once again to figure 1.) Given this, the paradox is dissolved: since the lower-level properties of the memory item don't get mapped, it follows that the whole memory isn't part of the analogy. For example, in garbage-henge, "is-made-of-stone" is an attribute predicate that doesn't get mapped, and so is not a part of the analogy. So of course you wouldn't see Stonehenge on your curb. (Or put correctly: so of course all of your Stonehenge memory wouldn't get retrieved and activated.)

But what is a structure-sensitive comparison, really? What does mapping two memory items onto each other amount to in the brain? The simplest and most natural neural interpretation of the term "mapping" is to say that mapping is *activation* . But if mapping is activation, then the above answer to the paradox won't do. Here's why.

Think in terms of a network of nodes and activation arcs between them. Activation usually *spreads*. So, given that the object nodes and the structure nodes of the retrieved item get activated, why doesn't the activation spread from these two areas to activate the attribute nodes of the retrieved item, too? In garbage-henge, the high-level structure gets activated and the object nodes get activated (cans map onto monoliths, "lies-next-to" in the garbage-can percept maps onto "lies-next-to" in the Stonehenge memory, etc.), and presumably the activation spreads both from the objects and from components making up the high-level structure. So why don't the attribute nodes get activated? Why doesn't "is-made-of-stone" get activated? Doesn't activation spread to it? If it and other attribute nodes were to get

activated, then your entire memory of Stonehenge would have been active, and hence retrieved. Which isn't what happened. So something is still wrong. On the assumption that mapping is activation, we predict a phenomenon that simply doesn't occur, namely the retrieval and activation of entire memory items which instead ought to be analogous. The paradox makes the phenomenon of analogy disappear. Yet analogy clearly exists.

One answer that would work, but seems a tad desperate on the face of it, is that attributes don't get activated because then you wouldn't get an analogy! This answer requires that the system (or organism) know ahead of time that it was constructing an analogy and not an ordinary, veridical reminding. Consider Rutherford. The behavior of alpha particles in his experiments reminded him of comets in their orbits around our sun. The proposed solution here is that his brain analogically retrieved comets because it was searching for an analogy in the first place. And since it was searching for an analogy in the first place, the retrieval process didn't activate the attribute or property nodes of his comet concept. Hence only the structure and object nodes were available for mapping (activation), and hence the behavior of alpha particles in his experiments analogically reminded Rutherford of comets. And that is how he had his analogy.

Note that there is no logical problem with this answer. The answer is *not* the claim that the system knows ahead of time which analogy that it wants. That *would* be impossible. The claim here is that the system knows ahead of time that it wants an analogy *of some sort* . A system could know this ahead of time. Still, I don't think this solution to the paradox, in its current form, can be right for three reasons. One, it breaks up memory retrieval into at least two disjoint processes, one for analogical retrieval and one for ordinary (both mundane and superficial?) remindings. While this might be correct, it seems ad hoc. One should postulate multiple process only if one has to. It is frequently better science to try to unify processes under one framework if possible. This is in fact what MAC/FAC does, and one of its principle pluses: MAC/FAC is a *single* process with two stages. Two, this solution makes it difficult to explain the data which supports MAC/FAC. Indeed the alleged two separate processes (one for analogical retrieval and the one for mundane retrieval, say) would *have* to interact to have a chance of explaining the MAC/FAC data, and if they interact, this solution reduces to

MAC/FAC. Finally, three, this solution seems to make what is a property or an attribute in a concept fixed and unchangeable; indeed, it seems to fix all three components of memory items: attributes, relational structures, and object nodes. In order for the analogical retrieval process to look for an analogy, it has to ignore attributes, and map everything else. It can only do that if attributes are there to ignore and everything else is there to map. For example, if "is-made-of-stone" or "is-made-mostly-of-ice" are attribute nodes in one's concept of the Stonehenge monoliths and comets, respectively, then this solution fixes them as attributes permanently, because only that way could the analogical retrieval process know to ignore them. They can never be rendered as relational structures, for example. But memory items seem more plastic than this.

Another solution to the paradox of analogy, where mapping is assumed to be activation, advocates a sort of general demotion of properties. On this solution, properties in the world aren't that important, so in turn, the attributes in a concept simply don't matter much for purposes of retrieval and inference. This solution seems incorrect because sometimes representing properties is important and so attributes are important in reminding.

Still another solution is to say that mapping isn't activation. But activation has to occur anyway, that is arguably what retrieval amounts to. Certainly activation is a necessary part of retrieval. So the paradox remains (it depends on activation). And now mapping is something over and above activation. But what could that be? Hence, not only doesn't this solution work, it leaves mapping undefined.

Nevertheless, having objected to all three solutions, I still think there is something worth exploring in them, especially the first two: analogical reminding is a separate, independent process and properties are, in general, less important than relational structure. The first two solutions become more palatable if one assumes that memory items are malleable in a certain way and that working memory items are constructed during reminding. In the next section I will consider detailed variants of all three solutions.

So here is where we are. On a natural interpretation of mapping -- it is activation -- we are stuck with the vexed question: Why doesn't retrieval always retrieve a whole item from long-term memory? Why doesn't retrieval activate the object nodes, the structural relation nodes, *and* the attribute nodes of a given item? In short, why is there any such thing as analogy? ...Why don't you *see* Stonehenge on your curb?[9]

6. What is a concept that a human may make analogies with it?

The low probability argument is a constraint on reminding. It entails that during analogical reminding, the relational structure of one or both of the analogous items are changed by the retrieval. The paradox is a constraint on analogy. It entails that the crucial notion of mapping can't be simple spreading activation. We can satisfy these two constraints by assuming that *reminding includes a process of constructing mappable structures*. This assumption ramifies, giving us a novel picture of concepts, conceptual change, and analogical reminding.

There are actually two pictures: a simpler one, and one that is complex and more speculative. Each picture corresponds to a way of dissolving the paradox of analogy, and both pictures assume that the low-probability argument is correct. I discuss both pictures in this section.

6.1. Dissolving the Paradox: the Simple solution.

The low-probability argument requires that structures be built during analogical reminding. The simplest way to do that is to assume that constructed structures are made from other structures, and that the construction process is really one of *altering existing structures* . So, the relational structures of (at least one of) the retriever and the retrieved item are altered slightly during reminding (with Gentner, we can assume that it is usually the structure of the retriever that is altered the most).

Now, the simplest way to dissolve the paradox is to say that the mapping process occurs during the retrieval process -- and not afterwards -- and that the mapping process *is* the structure-altering process, altering the existing relational structure of one (or both) of the analogous items. To see that this dissolves the paradox, we need only note that on this solution the mapping process is *not* a process of spreading activation, but rather one of structure altering. So, the reason attributes don't get activated is that activation is not being passed to them. Relational structures are being altered, but activation is a separate process. In fact, memory item activation, on this solution, is left unspecified.

Here's an example. My representation of Stonehenge has a structural component specifying how I remember the monoliths being arranged. On the Salisbury Plain, the monoliths are in an open circle, and the circle is incomplete now because several of the monoliths have fallen over or tilted. Suppose there are some garbage cans arranged (by accident) in a kind of semi-circle, and they too were a jumble -- some upright, some tilted, some having fallen over. The configuration of the monoliths was not exactly the same as the configuration of the garbage cans. This much is certain. But consider my representations of the cans and of Stonehenge, especially the structural components representing the configurations of the two groups of things. If we assume that 1) these structural components were not identical before the reminding, and 2) after the analogical reminding they were identical (at least at some level -- a quite abstract level, perhaps), then we are led to conclude that the reminding process aligned the structural relation components of the representations (i.e., the representations of the configurations of the stones and the cans), and that this aligning process had to *alter* the structural relations in (at least) my perceptual representation of the cans. But assumption 1) is just the low-probability argument, and assumption 2) is just Gentner's prevailing theory of analogy. We can safely conclude that viewing mapping as a purely structure-altering process (i.e., as a *non-activating*, structure-altering process) which occurs during reminding dissolves the paradox.

On this solution, the reason whole memory items are not retrieved is that analogical reminding is first and foremost a structure-altering process. Since structures are defined over object nodes, object nodes come along for the ride. Activating the entire memory item, i.e., the

attributes, too, is never a problem. Analogy occurs, therefore, because of a split between activation and representation construction.

This solution to the paradox is like the third one discussed above in section 5. And accordingly it has the main problem that one had: its explanation for why attributes don't get activated during analogical retrieval is too ad hoc, because activation is not incorporated into this solution and is left as a problem for another day. It would be nice to have a solution that incorporated activation. The next one does that. But it also requires a much more complicated view of concepts.

6.2. Dissolving the Paradox: the Speculative solution.

This solution does not leave the problem of activation dangling. The general framework for this solution is this: replace the traditional, tripartite view of concepts, which sees them as comprising objects, attributes (properties), and structural relations, with a more process-oriented view in which none of the three components are fixed, but rather change (or can change) during analogical remindings. So, nothing is essentially a property or an object, but rather takes on that role in certain contexts. (If this is right, then it is possible that other (all?) cognitive processes result in this sort of conceptual change, but here I am only concerned with analogical reminding.)

The key to this second solution is this: *analogical reminding can change what counts as relational structures in memory items*.

The second solution uses the following four premises:

1. Representing the relations things in the world can partake in is crucial to an organism's survival -- much more important than merely representing the things themselves together with (or as the nexus of) their properties or attributes.
2. Mapping is activation, and activation spreads. (In the first solution, mapping had to be purely structure-altering.)

3. There is no such thing as reminding, in general. Rather, there are different kinds of reminding. Analogical reminding is just one species of reminding. Analogical reminding exists in order to highlight and categorize the relations between things the cognitive organism finds in the world.
4. People can transform information represented by attributes into logically similar information represented as relational structures. That is, they can represent the color of an apple as Red(apple) at one time, and then at another represent the color of the apple as Color-of(apple, red).

Premise 1 does *not* imply that we don't store information about the attributes of things in the world. We clearly do. But premise 1 does suggest that attributes are important in mental representing only in certain contexts; they can be ignored in other contexts. Premise 1 seems plausible on inductive grounds, once it is noted that relations represent functional roles (in general, to say R(A, B) is to say that A functions in a certain R way relative to B). It is very rare in life for a thing-in-itself to be important to us. Usually what matters is the role the thing plays, and several things can usually play that role. Think about the two biggies in life: food and sex -- all that matters is whether something is edible or impregnable, and both are relations. In fact, it seems plausible that most of our categories are functionally defined. If so, then relations tell us what types of things there are in the world. Even knowledge of particulars (*this* coffee cup; *that* green mechanical pencil) is arguably functional, at least in part: successfully referring to this particular coffee cup in the world requires using the nexus of a collection of (internally represented) properties (white, thick, heavy) together with a collection of functional relations (holds coffee, reminds me of the University of West Florida).[10]

On a deeper, more philosophical level, premise 1 is plausible because we live in a universe that is a vast collection of processes; nothing is just a static object. Heraclitus was right: all is change, and you can't step into the same river twice. But relations are just a way of representing processes (on the *situated action* way of viewing things, *all* relations (even "greener-than") really amount to representing a process of some sort; see Bickhard and Terveen, 1995).

Premise 2 is simply the best way to fold activation into a solution to the paradox.

Premise 3 is quite controversial. But it makes sense, especially if premise 1 is true. Here is an argument for it. Recall that relational structure is the functional organization of objects (object nodes in mental representations). To represent a dog's panting tongue as releasing heat to the surrounding air is to represent that tongue as playing a role in a certain process. That role is the tongue's function at that time. To represent an ungloved hand as releasing heat to the surrounding cold air is to represent that hand as playing the same role. That is the foundation of the analogy between the two. But it is unlikely that that role, the representation of a dog's tongue and the representation of the ungloved hand, identically matched between the two items before the analogy. For starters, it is unlikely that that role was salient or highlighted in each item the same way and to the same extent before the analogy occurred. Indeed the *point* of the analogy, it seems, was to highlight or make salient the relevant heat-dissipation role (represented by some structure) that became common between the two items. Since this structure wasn't there ahead of time, constructing and matching this structure must have occurred with the construction of the analogy. But this makes analogical reminding unlike other forms of reminding -- analogical reminding is the type of reminding used for the special purpose of constructing and aligning structure. Hence, analogical reminding is just one species of reminding.[11]

(I want to stress again that I am not assuming that the matching structure was created out of nothing. Nor am I assuming that any two memory items can be made analogous. The relevant matching structure was no doubt created from already existing structures and attributes specific to each of the two items which might have been *similar* to each other before the analogy occurred. But remember, merely being similar isn't good enough. The structures, or least some aspect of them, must be able to be made identical.)

For premise 4, it is well-known that people can represent the same information in different ways -- the color of an apple, for example. But can people transform, e.g., Red(apple) into Color-of(apple, red)? Gentner and her colleagues have evidence that such

changes occur *during development* (Gentner, et al., 1995; Kotovsky and Gentner, 1996; Kotovsky and Gentner, 1990). The leap we have make here is that such changes can occur very quickly -- during reminding, in fact. It is this that makes the second solution appear radical.

Now, the second solution is this. Creatures with robust cognitive abilities need to be able to represent and compare relations (premise 1). Analogical reminding is a process that focuses on structure (premise 3). It can alter and rearrange existing structure to meet this need, but also analogical reminding can change what counts as relational structures in memory items. In particular, it can change attributes into relational structures (premise 4). So, the reason attributes don't get activated during the mapping process on the second solution (premise 2) is that with respect to analogical reminding attributes qua attributes *don't exist*. The second solution dissolves the paradox because attributes are not activated *qua attributes* (i.e., qua single-place predicates). This, however, doesn't sideline *the information* attributes contain because attributes and structures can slide back and forth, each changing into the other. Attribute-hood is a relative thing, and not fixed.

This completes the second solution to the paradox.

7. Conclusion.

The second solution is my personal favorite because it hypothesizes mental representations that are quite malleable and which change and alter due to cognitive pressures, and it seems to me that only such malleable mental representations have a chance of explaining the robustness and creativity of human cognition. Furthermore, if the analogical change hypothesis is right, then, since analogical reminding is a very common psychological process, we get the conclusion that the constituents of the mind, the constituents of thought, change rather frequently.

The picture of the human mind (and indeed other animal minds) which emerges from the second solution is quite exciting to contemplate, for it is a picture of a dynamic and fluid mind. This it seems to me is a welcome result for many reasons, not least of which is that everything in the analogical change hypothesis is compatible with computationalism. So a robust cognitive dynamics can be had within the computational paradigm, which is good, because it is the best paradigm we have. We can now rigorously start exploring the plasticity and malleability of concepts and memories, for this is the key to how we create new knowledge out of old, and see what we hadn't seen before.[12]

Endnotes

1. The phrase "theoretical psychology" makes some cognitive scientists shudder. I think this is because researchers are worried about lapsing into the kind of speculation that eventually lead to behaviorism in the mid-twentieth century. But I think theoretical psychology is important and has a place in modern cognitive science; we should not avoid it.

2. To dispel any misconceptions, it does not follow from 1)-3) that the computational hypothesis fundamental to cognitive science is false, nor that there are no such things as mental representations. I am arguing that one of the computational processes in our heads, analogy, is a representation construction process.

3. Perhaps another name could be "Gentnerfication" because the process produces increasingly mappable structures of the kind postulated by Structure Mapping Theory.

4. This is an interesting feature of SME: though it is primarily thought of as an analogy engine (in fact, as an implementation of Structure Mapping Theory), in MAC/FAC, SME nevertheless produces mostly mundane remindings. Briefly, this is primarily due to the MAC stage: SME produces mundane remindings because that is mostly what the MAC stage gives it. SME can run in one of three different modes: analogy mode, literal similarity mode, and mere appearance mode. In MAC/FAC, SME is run in literal similarity mode. Yet, within MAC/FAC, it still produces analogies and mere appearance matches from time to time. Indeed SME is almost always run in literal similarity mode, and still it produces analogies and mere appearance matches. SME works by coalescing initial local matches into global matches. Apparently, the ratios of the three types of remindings depend on which local matches are produced and how successfully they combine into global matches. For the technical details of SME's architecture, see Falkenhainer, et al. (1989). The details for the way SME works within MAC/FAC can be found in Forbus, et al. (1995).

5. Many psychologists have discussed conceptual change and the constructive processes involved in analogy and related processes, and there are several hypotheses with psychology

about analogy and conceptual change. For example, see Barsalou (1983, 1987, 1989); Camac and Glucksberg (1984); Gentner (1983, 1989); Gentner and Markman (1995); Gentner and Wolff (this volume); Glenberg, et al. (1994), Glucksberg and Keysar (1990); Kelly and Keil (1987); and Markman and Medin (in press). See also Black (1979). For a good synopsis of the field of analogy and some interesting speculation on its future direction, see Hoffman, (1995).

Intuition-based artificial intelligence has also contributed a fair number of intriguing hypotheses about the nature of analogical conceptual change. For example, see French (1995); Hofstadter, (1995); Mitchell (1993); Indurkhya (1997); and Schank (1982).

Some psychologists are now beginning to explore *how* some of the changes are produced in humans, that is, what the detailed mental processes are that result in conceptual change. See Gentner and Wolff (this volume), Gentner et al. (1995), and Kotovsky and Gentner (1996).

In AI, the "how" question takes on complicated empirical and methodological baggage. For an AI program that implements a model or hypothesis about analogical conceptual change, we definitely know how the change occurs: you can't write code without specifying the details of a process. But does the process in the machine of changing knowledge representations have anything to do with the process in humans of conceptual change? This is the empirical/methodological question. Unfortunately, for a variety of reasons, including the fact that we still don't know all that much about how concepts change in humans, many of AI modelers to date have had to invent their own processes by relying on introspection and intuitions of what is plausible. I think, by and large, such speculation is a good thing simply because it widens the pool of what we consider possible, but it should be remembered that it is speculation and speculation based on data derived from introspection.

Not all AI programs, however, are based on introspective data. Some robust and highly suggestive computer models of analogy and conceptual change are based on psychological data collected in experiments. Moreover, the performance of these models has also been experimentally compared with human performance. In many ways, these programs and their ties to psychological experimental represent one of cognitive science's real success stories (Forbus et al., 1998). See ACME and its associated model of memory access, ARCS (Holyoak and Thagard, 1989; Thagard et al. 1990), IAM (Keane and Brayshaw, 1988; Keane Ledgeway

and Duff, 1994), LISA (Hummel and Holyoak, 1997), Phineas (Falkenhainer 1990a&b), and the Structure Mapping Engine (SME) and it's associated model of memory access and retrieval MAC/FAC (Falkenhainer, Forbus, and Gentner, 1989; Forbus et al., 1995).

6. For example, Barsalou has shown that what counts as the typical member of a category, e.g., the category of birds, can change depending on context and that such changes are reflected by changes in representation. Also, he has shown that different individuals in the same population produce different examples of what counts as a typical member of a category. Finally, some of his data contravenes the more or less traditional view of how we represent categories (like birds). This view assumes category representations have a stable, definitional core. In some of Barsalou's experiments, subjects explicitly relying on definitions of categories did not exhibit the expected conceptual stability. See, Barsalou, 1985, 1987, and 1989.

7. One can discern in the literature two views: 1) concepts exist ahead of time in long-term memory and in working-memory they interact with and change each other, 2) concepts do not exist ahead of time in long-term memory but are constructed from more or less stable chunks of knowledge and that once in working memory the items there interact and change. Since we currently don't have enough information to decide how long-term memory items are stored, the point of these two views is the same for our purposes: concepts in working memory interact and change each other. So I am not going to pick between these two views. Besides, it is not clear one *can* pick between these two views. It is possible to set up these two views so that it is *logically* impossible to distinguish between them, sort like the two claims that the world was created five minutes ago complete with memories and the world was not created 5 minutes ago, and past events really did happen.

8. The reader might think there is a third answer: the structures *partially* match at the beginning, before the analogy, and then match more completely in the way required for analogy after the reminding and because of the reminding. I have no doubt that this sometimes occurs, maybe it is even the usual case, but it is important to notice that this is a variant of answer 2). Answer 2) is the answer I favor. It says that the relevant conceptual structures do

not match *in the way required for an analogy* before the analogy (except rarely, perhaps). According to Structure Mapping Theory, analogies are *isomorphisms* between high-level structure. In an important sense, the two concepts simply share the one structure that funds their being analogous. (There is a fair amount of agreement on this point, though it is not universally accepted. But as I said, I do assume it is true because I am assuming Structure Mapping Theory.) Now, a partial match, by definition, is not an isomorphism. So, granted, a partial match might get the analogy process started, but the question remains, where do the isomorphic structures come from? Or: Where does the unifying structure funding the analogy come from? Answer 2) just says that the structures weren't there ahead of time -- they are not part of the concept nor are they stored in some generalization or ISA hierarchy (as in Falkenhainer, 1988). From this fact, I infer that the relevant structures were constructed in real time. This is simply a restatement of the analogical conceptual change hypothesis.

9. Actually, this is a version of a general problem. The question of where to stop the activation is a problem in *every* case of reminding, and hence in every case of thinking. Think of cats. Not everything you know about cats is activated when you do this. Of course, it would be bad to design an intelligent system that always retrieved everything it knew about any subject when it was reminded of that subject. But how do we design a system so that it retrieves what it needs without retrieving everything, given that, in the general case, it doesn't know ahead of time what it needs? I think this problem is quite interesting and requires for its solution a marriage between heuristic-driven retrieval (most of time, an intelligent system need only retrieve the "standard" information for a concept), a theory of conceptual boundaries, i.e., a theory of where one concept ends and another begins, and a theory of how concepts coalesce to form the larger structures we call knowledge.

10. This claim has a strong Berkeleyan flavor to it, but only a flavor. I am not saying that things *in the world* are only collections of perceived properties. They might be, but I am not committed to that view here.

11. The form of highlighting I mentioned *seems* very similar to Gentner and Wolff's notion in their paper in this volume. But I am not sure of this because on her theory of analogy, she also

seems committed to the view that analogies happen because the relevant structures antecedently match. It is quite clear that Gentner thinks analogy is responsible for interesting conceptual change. But, throughout her many papers, Gentner seems to try to have it both ways: analogous structures are there ahead of time, *and* analogous structures (or parts of structures) are constructed at some time around the analogy. I have struggled with this antinomy in her theory, and my considered opinion now is that it is really an unresolved issue with Structure Mapping Theory. If this right, then Structure Mapping Theory could be changed to accommodate my retrieval-based structural change and indeed the whole of my analogical conceptual change hypothesis without doing it serious damage, and, I suggest, improving it slightly by making it better able to handle concept's robust capacity for change.

12. I thank Robert Davidson, and Art Markman for good comments on an earlier draft. I thank Chris Fields, Ken Forbus, Bob French, Dedre Gentner, Celia Klin, and Art Markman for discussing these matters with me. And I thank my graduate research group: Jon Beskin, Doug Beyer, Phil Gross, Lewis Loren, Clay Morrison, and Michiharu Oshima for helping me formulate these ideas.

References

Barsalou, L. (1983). Ad hoc categories. *Memory and Cognition 11* , 211-227.

Barsalou, L. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *J. of Experimental Psychology: Learning, Memory, and Cognition* 11, pp. 629-654.

Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (ed.) *Concepts and conceptual development: ecological and intellectual factors in categorization* . Cambridge: Cambridge Univ. Press, pp. 101-140.

Barsalou, L. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (eds.) *Similarity and analogical reasoning* . pp. 76-121.

Black, M. (1979). More about metaphor. In A. Ortony (ed.) *Metaphor and thought* . Cambridge,UK: Cambridge Univ. Press pp. 19-41.

Camac, M. and Glucksberg, S. (1984). Metaphors do not use assoications between concepts, they are used to create them. *Journal of Psycholinguistic Research* 13: 443-455.

Dietrich, E. and Fields, C. (1996). The role of the frame problem in Fodor's modularity thesis: a case study of rationalist cognitive science. In K. Ford, and Z. Pylyshyn (eds.) *The robot's dilemma revisited: The frame problem in artificial intelligence*. Norwood, NJ.: Ablex. pp. 9-24.

Dietrich, E., Morrison, C., Oshima, M. (1996). Conceptual change as change of inner perspective. *Embodied cogntion and action: Proceedings from the 1996 AAAI Fall Symposium* , Menlo Park, CA.: AAAI Press, pp. 37-41.

Falkenhainer, B. (1988). *Learning from physical analogies: A study in analogy and the explanation of process.* Ph.D. Dissertation, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign. Report no. UIUCDCS-R-88-1479.

Falkenhainer, B. (1990a). A unified approach to explanation and theory formation. In Shrager and Langley, (eds.) *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, CA: Morgan Kaufmann.

Falkenhainer, B. (1990b). Analogical interpretation in context. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society. Cambridge, MA: Lawrence Erlbaum.

Falkenhainer, B., Forbus, K., & Gentner, D., (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence 41* (1), 1-63.

Finke, R., Ward, T., and Smith, S. (1992). *Creative Cognition* . Cambridge, MA: MIT Press.

Forbus, K., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19, pp. 141-205.

Forbus, K., Gentner, D., Markman, A., and Ferguson, R. (1998). Analogy just looks like high-level perception: Why a domain-general approach to analogical mapping is right. *J. of Experimental and Theoretical AI.*

French, R. (1995). *The subtlety of sameness* . Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science 7* , 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (eds.) *Similarity and analogical reasoning* . Cambrigde, UK: Cambridge Univ. Press, pp.199-241.

Gentner, D. and Markman, A. (1995). Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (ed.) *Proceedings of the workshop on Similarity at the Univeristy of San Marino* . Milan, Italy: Bompiani.

Gentner, D., Rattermann, M. J., Markman, A., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (eds.) *Developing cognitive competence: New approaches to process modeling.* Hillsdale, NJ: Lawrence Erlbaum. (pp. 263-313).

Glenberg, A., Kruley, P., and Langston, W. (1994). Analogical processes in comprehension. In *Handbook of psycholinguistics*. Academic Press, pp. 609 - 640.

Glucksberg, S. adn Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* , 97, 3-18.

Heraclitus. Trans. with commentary, Philip Wheelwright, *Heraclitus*. (1964). New York: Antheum. (Originally published by Princeton Univ. Press, 1959). p. 58.

Hoffman, R. (1995). Monster analogies. *AI Magazine* 16 (3), pp.11-35.

Hofstadter, D.R. (1995). *Fluid concepts and creative analogies*. New York: Basic Books.

Holyoak, K. and Thagard, P. (1989). Analogical mapping by constrain satisfaction. *Cognitive Science* 13 (3) pp. 295-355.

Hummel, J. and Holyoak, K. (1997) Distributed Representations of Structure: A theory of analogical mapping. *Psychological Review* .

Indurkhya, B. (1997). Metaphor as change of representation. *J. of Experimental and Theoretical AI* .

James, W. (1950). *The prinicples of psychology* (vol. 1). New York: Dover. (Originally published in 1890).

Keane. M. and Brayshaw. M., (1988). The incremental analogy machine: A computational model of analogy. In D. Sleeman (ed.) *Third European Working Session on Machine Learning* . London: Pitman.

Keane, M., Ledgeway, T., and Duff, S. (1994). Constaints on analogical mapping: A comparison of three models. *Cognitive Science* 18, pp. 387-438.

Kelly, M. and Keil, F. (1987). Metaphor comprehension and knowledge of semantic domains. *Metahpor and Symbolic Activity* 2 pp. 35-51.

Kotovsky, L. and Gentner, D., (1990). Pack light: You will go farther. *Proceedings of the Second Midwest Artif. Intell. and Cog. Sci. Society Conf.* pp. 60-72.

Kotovsky, L. and Gentner, D., (1996). Comparison and categorization in the development of relational similarity. *Child Development* 67, pp. 2797-2822.

Lucretius, Titus. *On the nature of the universe.* Trans. by R. E. Latham, New York: Penguin Books.

Markman, A. and Medin, D. (in press). Similarity and alignment in choice. *Organizational Behavior and Human Decision Processes* .

Mitchell, M. (1993). *Analogy-making as perception* . Cambridge, MA: MIT Press.

Poincare, H. (1952). *Science and method* . New York: Dover. (Originally published in 1908, Paris; English version in 1914, London).

Oshima, M. (1996). Analogy as a creative process. Master's Thesis. Program in Philosophy, Computers and Cognitive Science, SUNY Binghamton, Binghamton, NY.

Schank, R. (1982). *Dynamic memory* . Cambridge, UK: Cambridge Univ. Press.

Thagard, P., Holyoak, K., Nelson, G., & Gochfield, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence* 46, 259-310.