Solutions to Fodor's Puzzle of Concept Acquisition: Transcript

Symposium at the Annual Cognitive Science Society (Cogsci 2005) in Stresa, Italy
July 22, 2005

Introduction

Sourabh Niyogi: Welcome to the symposium on "Solutions to Fodor's Puzzle of Concept Acquisition." We have a broad array of participants from a large number of perspectives in cognitive science, spanning philosophical, developmental and computational approaches, including Jerry Fodor himself. A brief introduction for everybody (raise your hand and say hello for a second):

Jerry Fodor	Rutgers University
Jean Mandler	UCSD
Frank Keil	Yale University
Alison Gopnik	UC Berkeley
Sourabh Niyogi	MIT
Timothy Rogers	University of Wisconsin
James McClelland	CMU
Dedre Gentner	Northwestern
Stephen Laurence	University of Sheffield
Jesse Snedeker	Harvard University

The questions we'll be asking today – or the topic of discussion – will be centered on the following 3 questions:

1. can a child ever acquire a new conceptual primitive? And if so, how?
2. how might a child expand the combinatorial expressive power of his or her representational system?
3. how might a child expand his or his hypothesis space of possible word meanings?

Over 30 years ago, there was a famous debate was held between Noam Chomsky and Jean Piaget on many foundational questions of cognitive science. Jerry Fodor presented in that debate his own impossibility arguments concerning many of the above key foundational questions. We'll be revisiting many of these issues again here, 30 years later. Fodor's argument core arguments were captured in his 1975 *The Language of Thought,* which I hope appears on your reading lists. We'll see if we can give an update to that in this symposium.

Here's a quick schedule, which we'll try to stick to as close as we can. [schedule] We have a brief introduction to the puzzle and the evolution of the puzzle in his own words from Jerry Fodor, followed by a solution from each of the participants. Each participant will have 25 minutes to discuss their solution – 15 minutes to present their solution followed by a 10 minute Q&A session, starting with Jerry Fodor, then the rest of the panel; if you have burning questions, save them for the roundtables after the symposiums – of which we actually will have two symposiums. After each batch of participants, we'll have a roundtable discussion, of which Jesse Snedeker will leading, which will integrate many of the different proposals.

I'd like to thank the organizers for allowing us to hold this event here at the Cogsci conference, and the participants here (many of them are not cogsci regulars) for venturing out here to make this a most special event. So everyone, thank our organizers and panelists.

Ok, let's get started. We'll start with Jerry Fodor who will give an introduction to the puzzle in his own words.

Jerry Fodor
Rutgers University

Let me just sort of plunge in. The book that was mentioned - the *Language of Thought* - offers an argument against the possibility of concept learning. It's one that many people disliked. It wasn't in fact a particularly original argument, which I will come back to in a moment. Still, the modal response was in fact shock and outrage. In fact, something like more or less you must be out of your expletive-deleted mind. That was a long time ago, when there were still dinosaurs at MIT. I have come over the years to dislike the argument too. But not for exactly the usual reasons, however; not because I think the argument is wrong. My reason is that the most contentious premise -- as given in LOT and as summarized as précis for this discussion – is, I think, dispensable. In fact, one derive the same conclusions from assumptions that aren't contentious, or anyhow some that I don't think can be so transparently true. So all I am going to do *now*, in the hopes of convincing those who are not convinced before, is give a quick characterization of the revised form of the argument and try to convince you that you should like it even if you didn't like the original version. So, here is the old argument, as in LOT and the précis for this talk. It moves in three steps.

Step 1: Concept learning -- if there is such a thing -- should be hypothesis testing and confirmation.

There are 2 ways of thinking about that. One is simply that there aren't any alternatives. There are various ways of putting that claim. But it is absolutely ubiquitous in the literature on concept learning, except insofar that literature isn't simply behaviorist. So concept learning must be hypothesis formation and confirmation, for want of a better learning.

Step 2: On pain of circularity, primitive concepts can't be learned, simply because they're needed to formulate the hypotheses in question.

By the way, and this is why I say this isn't a particular new or striking argument, this argument was about explicit by all non-behaviorist empiricists from about Locke on. Hume says, for example, that complex concepts are learned; primitive concepts are obviously innate. No one even bothers to discuss it -- it just taken as self-evidently true. So from that point of view, the tradition is to consider this; the tradition in at least the empiricist literature is (to say nothing of the rationalist literature) is not to take this argument to be a paradox, but simply to take it as a self-evident truth.

Step 3: Most concepts are primitive.

Then you get the conclusion that most concepts are unlearned.

Now, much of the discussion that this line of thought has engendered -- over the past 20 or 30 or 50 or 100 or whatever it is years -- has in fact about the third step. That is, the assumption that most concepts are primitive. It's been about the status of definitions or constructivist theories of concepts. Anybody who says that most quotidian concepts are complex whether they are thought of as logical constructions or statistical constructions. So what I want to do today, is very rapidly (its a very short argument) to present a somewhat revised formulation, in much the same spirit, that does *without* the anti-constructivist premise, and so may be easier to live with.

Step 1. Concept learning must be hypothesis formation and confirmation, just as before.

Step 2. [[Thinking about being C is sufficient for having the concept C; audio failure]]

I'll come back to explication of that in moment.

Step 3. But formulating the hypothesis that concept C expresses property of being C itself requires the thinking about the property of being C.

So in conclusion, you can't learn a concept you don't already have. This argument to me seems so obviously sound that one then wonders why it hasn't been widely noted. With the agreeable consequent being that the notion "concept learning" should be expunged from the canonical concept learning literature. The answer I think is much deeper than the argument. The answer is, excepting behaviorists (who of course have no truck with concepts at all), Anglophone psychologists and philosophers for the last 100 years or so -- have uniformly -- practically without exception -- have been pragmatists about concepts -- in particular about concept possession. That is, they have identified concept possession with having certain *skills*. In psychology, typically these are sorting skills. On this view, concept possession is what you are able to *do*, rather than what you are able to think about. So the picture in psychology is to have a concept is to be able to sort things – there are ones fall under it and there are the ones that don't. Paul Bloom -- to give an example of one of zillions – for example, in his book about word learning says, look, the test of concept possession is in fact discrimination learning. That, I think has been ubiquitous -- that is, people have thought about concept possession in terms of what you can do rather than what you can think about -- they have thought about it pragmatistically. Andy Clark's paper -- that is, the one he gave here the other day, is a perfectly good example of that.

However, pragmatism got the relation between sorting and concept possession precisely backward. Being able to sort Cs isn't constitutive of having the concept C – it's just a manifestation of your having that concept. And normally, there are lots of other manifestations too. You sort Cs by thinking about which things *are* C, and which aren't -- but not of course, vice versa -- that is, you don't think about things by sorting. You can think about things that you aren't able to sort, but you can't sort things that you aren't able to think about. The pragmatist account of concept possession violates – in fact, explicitly *denies* this truism -- and therefore simply isn't *true*.

Thinking about being C is sufficient for having the concept C. And if that's so, the argument goes through in its revised form -- that's essentially premise (2). Notice that it goes through whether or not constructivism is assumed. Learning about a concept requires thinking about the concept that it expresses. If the concept of a BACHELOR – I take it a paradigm example of a definable concept -- is the concept of an UNMARRIED MAN, then learning the concept BACHELOR requiring thinking about the property of being an unmarried man, precisely because it requires confirming a hypothesis to the effect the concept of bachelor *is* the concept that expresses the property of being an unmarried man. But if you can think about the property of being an unmarried man, then you already have the concept of a BACHELOR. So whether or not, the concept of a BACHELOR can be defined, it can't be learned. QED, as one said in the old days.

That's all there is to it. Except for a brief and sententious closing remark. That concepts can't be learned doesn't of course show that they are innate. The distinction innate-learned isn't exhaustive. For example, concepts might be acquired, but not by a learning process -- for example, in particular, not by hypothesis formation and confirmation. For example, concepts might be acquired, but not by a learning process – that is, not by hypothesis formation and confirmation. See, for example, the ethological literature and the linguistic literature, where notions of concept acquisition, by processes like imprinting, triggering, parameter setting and the like are commonplace. Die-hard empiricists might, in fact, take some comfort in that -- some, but not much.

That concept acquisition requires mind-world interactions is a point that Cartesians (including Descartes) have always insisted on. Descartes was quite aware that you can't lock an infant in a closet and expect it to come out speaking English, or even with the concept of a noun phrase. If that were enough to vindicate concept empiricism, then the present discussion wouldn't be worth having. The question of interest of course isn't of whether concepts are acquired but whether they are *learned*. The line of argument that I have been suggesting shows that they aren't. That is not a paradox, but a straightforward consequence of the inductive character of the learning process. The way to deal with it is to learn to deal to live with it. Thank you.

Sourabh Niyogi: Ok, we'll start with our solution providers. Up first is Jean Mandler from the University of California San Diego.

Jean Mandler

University of California San Diego

Mandler This is one of my favorite quotes – I'll read it for those of you in the back who can't see this.

> "When you put questions to Nature and Nature keeps saying "no," it is not unreasonable to suppose that somewhere among the things you believe there is something that isn't true."
>
> Jerry Fodor 1981

It's been long one of my favorite quotes. The only place I think that we disagree is what it is that Nature is saying "no" to. If I remember correctly, among other things, one of things Nature was supposed to say "no" to is the view that concepts can be defined over primitive components. But I would like defend the idea that they *can* be so defined, as long as we don't ask our definitions to do too much. A concept tells you what the core of something is. Or, it enables you to think about it, as Jerry just said. It doesn't include everything that you know about it. It places it in a framework of knowledge, and we can hold concepts that contradict each other -- so it doesn't have to do with truth, it can admit of exceptions, and so forth. What I would say Nature is saying "no" to is that concepts are learned by hypothesis and test -- at any rate, they don't start out that way -- they start out as a *redescriptive* process – and that's what I am going to be talking about today. I'm going be talking about what babies do, when they first begin to form concepts after birth:

Overhead 1. Some lessons from infancy research

> Many concepts are not learned by hypothesis and test.
>
> Basic-level concepts are not the first to be formed. They are gradually learned achievements.
>
> The conceptual system is built from the top-downward. Animal comes before dog. Container comes before cup.
>
> Even the earliest concepts are complex and structured.

One of the neat things about infancy research is that a lot of things that seem obvious when you study adults suddenly do not seem so obvious when you start looking at infants. The premise that one learns concepts by hypothesis formation and test is one of these. Now that is surely one way for older folks to learn new material, but it is a highly implausible method for the newborn opening her eyes for the first time and gazing out at the world. Even if you assume a moderate about of innate machinery, as I do babies are unlikely to have the wherewithal to engage in hypothesis and test. For one thing, they don't have any concepts yet in which to formulate a hypothesis, or even to express an expectation. Still, they are able to create concepts, and so the question is, how do they do that?

Another thing that seemed obvious before we studied it in infancy was that basic level concepts are the first to be formed, and are foundational thereby. Basic-level concepts are concepts like DOG and CUP and so forth. But this turns out not to be the case. The first concepts are highly global, very abstract and sketchy, superordinate notions like ANIMAL or CONTAINER. These initial global concepts provide the definitions that are going to ground the basic level concepts that come later, such as DOG and CUP. So, a 7-month old looking at a dog will think ANIMAL, or look at a cup and see a cup, but think CONTAINER. Basic-level concepts are in fact *new* concepts that are differentiations out of less detailed notions and it takes a long time to achieve them - many months, in fact. Now I think this is important because part of the problem has been that decomposing basic-level concepts into primitives has proved almost impossible. Decomposing infant concepts turns out to be a lot easier, and this suggests to me that one of the things that we should do is save our decomposition for the conceptual foundations and then see what else can be accomplished through differentiation, and various power-enhancing devices that are going to be described in other talks today.

I’ve hypothesized that the first object concepts are built out of spatial information -- and by the way that we understand space, they have to be structured. For example, a primitive concept is that of CONTAINER, which can be defined as a space with an inside and an outside -- the notion is structured because you can't have one without the other. You can't have an inside without and outside, or vice versa. The only unstructured concepts I know are sensory ones, such as RED, but these are very late and difficult acquisitions. Children can be as old as 3 years before they make any headway on the color domain. It’s even possible, though I don't know of any data to show this, that unstructured sensory concepts require language to be learned, and without natural language you may not be able to acquire them.

Overhead 2. Innate contributions to concept formation

Attention to motion and spatial relations

A redescription mechanism (Perceptual Meaning Analysis) that works on spatial information

A small vocabulary of spatial primitives

Infants are first responsive to spatial information and to motion. They pay attention to what objects *do* much more than what they look like. They attend to the paths that objects take, whether they move by themselves or not, and whether they interact contingently with other objects. Now of course infants begin to form perceptual schemas of what objects look like at a very early age. But that is implicit learning, and it is not necessarily conceptual. Perceptual schemas would not under my, or I think Jerry’s, definition be conceptual. So that's implicit learning -- it doesn't require attention and typically doesn't even reach conscious awareness. Responding to something as familiar is not the same as conceptualizing it.

In my theory, it is in fact attentive processes that are used to form the first concepts, but not by hypothesis and test. Rather, I think that babies find patterns in perceptual data, redescribe them, and generalize from those redescriptions. Perceptual information is in fact incredibly detailed and concepts are not. So you need a mechanism that is going to select and redescribe perceptual information into a simpler form that can be used for thought. The name I have given to this mechanism -- I used to call it “Perceptual Analysis” but in my recent book on this topic, [The *Foundations of Mind*] I've called it “Perceptual Meaning Analysis” -- to get across that this is where meaning comes from.

When attention is brought to something, this mechanism can translate it into conceptual form. For example, we look at faces countless times, and we don't typically take in many of the details conceptually. I was more than 50 years old before my concept of FACE included the notion that the eyes were level with the ears. I used to think that the eyes were up above and the ears were in fact down below. I told this to someone once, and they laughed at me because they said you wouldn't be able to wear glasses if your eyes were up here and ears were down below. But I didn't wear glasses, and this fact had completely missed me. When I did become aware of it, it was in a dull seminar and I started looking at people's faces and looked and looked and I said "Oh my god, their eyes are at the same place as their ears!" [laughter] This is a new part of that concept: including different information from the old. However, this mechanism is a *descriptive* mechanism and in that sense, it is not just a triggering device of the kind that Jerry has described. It’s a descriptive mechanism and so there has to be a vocabulary to couch the description.

As adults, of course, we have words. Looking at faces, we have words in which to couch descriptions, but babies don't. Instead, they are born with a tendency to attend to certain kinds of spatial relations.

In particular, paths, and a few relations like containment and contact. And these, I believe innate, proclivities, are a beginning vocabulary.

So, infants create objects out of an innate base of spatial primitives, a base that is presumably later supplemented by analysis of bodily feelings of force. But we are innately attuned to -- or I think I can use of Jerry's recent terms here -- we can lock onto some kinds of spatial relations. And those parsings are used in the initial workings of Perceptual Meaning Analysis. It seems to be stretching the notion of hypothesis testing to say that this mechanism is an inferential one. It's not: it simply outputs descriptions of what it is looking at.

So the question is -- how many innate spatial primitives need there be? As best as I can tell, not all that many are needed get the conceptual system up and running. And the following gives some of them (15 or 16):

Overhead 3. Innate primitives

Path primitives:
Start-of-path
End-of-path
Into container
Out-of container
Onto Surface
Off-of Surface
Up
Down

Link
Container
Surface
Contact

Motion primitives:
Rhythmic (biological)
Straight (mechanical)

Only a few of these are all that is needed to get the concept ANIMAL, for example, started The next overhead [Overhead 4] gives you the primitives that get you the initial concept of ANIMAL:

Overhead 4. How many spatial primitives are needed?

Only a few for a first concept of animal thing:

Self-starting Path (No Contact)

Contingent interaction without Contact (Link)

Rhythmic (biological) Motion

Only a few for a first concept of inanimate thing:

No Path or starts with Contact

No Link without Contact

Straight (mechanical) Motion

Infants redescribe animals as things that start themselves, which is to say they begin paths with no contact from another thing, they move in certain rhythmic or irregular ways, and most importantly, they move contingently in relationship to other objects, and sometimes do so from a distance. This is known in image-schema terminology as a LINK. It's basically a contingency response.

Now an animal defined as self-moving interactor is not a bad core for the concept of an animal, and it's a core that I believe lasts us for a lifetime. A similar set of primitives gets us to inanimate movers as non-interactive things that do not start themselves, that require contact for motion to begin and do not interact with other things from a distance, and that follow straight or mechanical looking paths.

The next overhead [Overhead 5] shows early global concepts that have been studied:

Overhead 5. Some Early Concepts Created from PMA

Objects

- Animal
    - Land animal
    - Air animal
- Vehicle
    - Land vehicle
    - Air vehicle
- Furniture
- Hand
- Utensil

- Plant
- Indoor things?
- Outdoor things?

Relations

- Containment
    - Inside
    - Outside
    - Tight
    - Loose

- Above (Up)
- Below (Down)

- Goal

I have question marks about indoor things and outdoor things because I don't have any data on that distinction before 16 months of age -- but I suspect it emerges quite early. I don't have time to go through these this morning, but I think all can be given "baby definitions" (let's put it that way)-- and that babies do have these concepts by the end of the first year, and some cases even earlier.

Now, let's go back to the innate primitives again. See Overhead 3. You'll see that there are a number of directional path primitives. Into-container, out-of-container, onto-surface, off-of-surface, up, down, link (which I've already mentioned), container, or containment itself, surface, contact, and some motion primitives, which have to do with the way in which things move. Location has got to be somewhere here too (I haven't included it on this overhead) -- these notions I believe also last us for a lifetime. And the next overhead will give an illustration of this -- this is taken from Moby Dick - this is Ishmael talking about whales as a fish:

> "Be it known -- that waiving all argument, I take the good old fashioned ground that the whale is fish."

Now, he knows (and even cites Linneaus about it) that whales suckle their young. But he finds that not as crucial as the fact that whales entirely live in water and never come on land. And that is what makes them fish. I would be willing to bet that all of us carry notions like that around with us today.

The small number of primitives that I've mentioned -- 16 or so – either singly or in combination –go a long way toward getting the system started. I'm sure there are more, but I don't think there need to be very many: 30, maybe? It's not going to be much more than that, if Landau and Jackendoff are right in their analysis of spatial concepts. There aren't an awful lot of spatial concepts, regardless of how our native tongues express then. So what we have then, is some mechanism that extracts some salient perceptual information from perceptual displays and simplifies it using primitives that combine to form concepts.

What is the format of these? See the next overhead [Overhead 6]:

Overhead 6. Possible formats for conceptual redescription

1. Image-schemas ☺
2. Perceptual symbols (a la Barsalou)
3. Other code

Whichever type of representation, it is not conscious. To retrieve a concept into consciousness requires either imagery or language.

I have said that I thought that a reasonable solution to the format of these descriptions is image-schemas. I don't think that is crucial for our topic today and I'm not going to spend any time on it. It could be perceptual symbols, of the sort that Larry Barsalou has talked about, or some other code -- I think image-schemas have a lot to recommend them, because they form not only a way of expressing thought and imagination and analogical learning but also, of course, they are implicated in forming conscious images. They themselves are *not* conscious images. I think I have been misinterpreted by some people to say that these are conscious images - or that I hold a resemblance theory of concepts -- I do not. Image-schemas are more like topological representations than anything else because they consist of limited kinds of spatial information. Path, for example, does not specify direction or speed. Container has no size or shape. In that sense they are more like topological representations.

A final comment (I think I am coming close to my time limit). Several things follow from the first concept of ANIMAL being at a very general level. [Overhead 7]

Overhead 7. Implications of initially global concepts

1. Early concepts can’t include many physical features, such as legs or wings.
2. New concepts are created by Perceptual Meaning Analysis finding new distinctions.

   At first, dogs and cats are just different looking self-movers.

   Language and culture focus attention on differences between them.

   Protracted process, even after language begins. Hence overextension of nouns.
3. Makes the system inherently hierarchical. Concept formation consists largely of differentiation.

First, most global concepts don't (in fact, *can't*) include specific physical features (such as legs or wings) or any other characteristic about what they look like. Second, their generality means that new concepts will be created by a process of *differentiation*. Changes in attention - often with the help of language that parents are speaking to children and the language that surrounds them -- enable new perceptual meaning analysis. They enable children to make different distinctions. But what happens, importantly, is that you get different associations accruing to different perceptual schemas. So you have the concept of ANIMAL -- the language is telling you that some these are called dogs and some of these are called cats, and that parents gets nervous when you pet one of them and they don't get so nervous when you pet another of them, and things of this sort. This is a protracted process, differentiation, and it largely goes by associative learning. So early on, the infant perceptually categorizes dogs and cats as different-looking, but conceptualizes them as land animals that look different. Gradually, a variety of associations accumulate, including the labels given by adults. But notice that even this isn’t hypothesis and test. It is basically pattern-learning.

The third result of concepts being global in nature is that the conceptual system is necessarily hierarchical. It takes a long time to differentiate dogs and cats. But throughout the process, their animal membership is never in doubt. Concept learning is top down from the very beginning and consists largely of a differentiation of initially global notions. Increased detail is one of the ways to increase the power of the system. I think that comes out clearly in Tim and Jay's talk later today. Another way is analogical transfer, which we'll be hearing about from Dedre Gentner, -- transfer of spatial information into other domains, such as time and other abstract domains. That will take me too far afield today, but may get addressed in other talks.

Discussion on Mandler

Sourabh Niyogi: We'll continue now with questions from the panel. Do you want to provide a rebuttal, Fodor. Hypothesis testing?

Jerry Fodor: Exactly... My grandmother taught me, at a very young age (mine, rather than hers) that I should not argue with people who know what they are talking about. So I am not going to argue about the order in which children acquire concepts. I am also not going to argue with her about what a definition is, though I think she has got a view that is certainly different from mine, and I think different from what you find in the literature and I don't think can be sustained in a discussion of concept identity or concept acquisition or concept possession. .

I think definitions have to give conceptually necessary biconditionals. That's why is there is so few of them. And I think if they don't do that, they won't serve the properties that concepts are supposed to serve -- in particular, the properties underlying modal statements. In particular, nothing that wasn't an animal could be a dog. I would be inclined to dig my heels in about that, but it doesn't really matter, as I tried to make clear in my presentation. This is exactly the kind of discussion that has made me want to bypass issues about definition and constructivism and give the argument in a way that doesn't depend on them.

What does depend on them is the issue of whether concept learning is a matter of hypothesis formation and confirmation. I really do think that's not avoidable. It is roughly the only answer that anyone has ever given to the question. How do you go, as Bruner used to say, beyond the data given. To say it is an inductive process doesn't help, because our only theories of inductive processes are hypothesis formation and confirmation.

So everything turns on that. I don't think it's avoidable. I think its what the Brits call non-negotiable. In particular, I don't think it's avoidable by appeal to a differentiation notion of concept learning. Because the kind of differentiation you need, is itself a form of hypothesis formation and confirmation, in a quite non-trivial sense. Let me just give yon one example and I'll shut up (for a while).

Suppose you say -- I don't care about the details -- suppose you think that a dog is a barking animal. Suppose that's the definition that you eventually come to. Start with the notion of ANIMAL -- you get it from somewhere -- from your genes or you get it from somewhere. And you differentiate it into the ones that bark and the ones that don't, and you say, dogs bark and cats do something else, so cats aren't dogs, dogs aren't cats, blah blah blah. How is this going to work?

In particular, how are you going to achieve that differentiation, unless you already have the notion ANIMAL (well that's granted) but also the notion BARKS. What you have to do, is to learn the following thing: on Jean's own authority -- animals are differentiated into things that bark and things that don't bark. Unless you have the concept DOG. Unless you have the concept BARK. You can't state the concept – you can't present to your mind -- the differentiating hypothesis unless you have the concepts that constitute the hypothesis. That I think is a self evident truth. It doesn't matter which form of induction you have in mind when you talk about inductive learning – they're all patterns of hypothesis formation and confirmation. One of the things that is deadly wrong -- and has been for years -- is that people don't recognize the hypothesis formation and confirmation model even as they endorsing it. I claim (this is a very strong claim) that there is *no* account of induction -- no theory of induction -- no theory about how thought can go beyond the data given that isn't a variant of the hypothesis formation-confirmation framework in the sense that is relevant to this discussion.

Sourabh Niyogi: Prof. Mandler, I'd like to get some rebuttal on how you use the term "pattern finding" compared to Fodor's hypothesis testing.

Jean Mandler: Do you want me to answer that question or Jerry's? A descriptive mechanism is different than a triggering mechanism, but it's not a simple learning mechanism. What I have tried to say today -- (very) very briefly, to be sure -- is that there is a whole set of concepts that come from a process that isn't terribly different, I think, from triggering -- it's a descriptive process. And I believe that makes a difference when you start talking about differentiation. Jerry does not, because that involves hypothesis and test -- and perhaps it [differentiation] does. I would be willing to grant that -- or at least to debate it more seriously, more than we could possibly do in this short time that we have today. But if you have a creature that already understands the world in terms of animals and in terms of various kinds of inanimate objects and things that are indoors and outdoors, and you also have one who is in fact observing the world -- this of course is an unconscious process, the observing of the world -- you already have a conceptual system.

Let me give one simple example, and hopefully that will make it clear. Let's assume for the moment that the baby doesn't know about barking, we are not going to worry about whether or not the baby acquires barking, or the difference between dog and cat. Let's just assume that babies only hear a different label for dogs and cats. And that's the only thing that they get from parents, or from the language around them. Now the babies have a perceptual schema formation mechanism that lets them see the difference between dogs and cats. They can associate the labels -- and you can call that hypothesis testing if you want -- that's fine -- they can associate the labels with those things. That's not very much of a concept -- I would say that the difference is something called dog and something called cat – but all they know conceptually is that these are self-moving interactors. That isn't a very good concept, but I think it would be an example. I believe that something like this is something like what babies do. Beyond that, I'm not sure what to say.

Jerry Fodor: I want to say one sentence -- and it would be a long sentence (with many subordinate clauses). Look, I think, this discussion is interesting, precisely because it deals with so many errors in what I think is in the pragmatist tradition -- in particular the empiricist tradition. Namely, in particular in this case, you can't get a learning mechanism -- the data that you like from observing the world -- you have to observe it under a description. In particular to get the datum that dogs bark, this one does, and that does, you have to observe the description under the description "is barking". You can't observe under a description unless you have the concepts that constitute the description. You can't see things as barking unless you have the concept barking.

This is I think tremendously important, because it shows how deeply notions of intensionality penetrate this whole discussion. Intensionality is a basic question – intensionality requires seeing as, and seeing as requires conceptualization.

Frank Keil
Yale University

I should probably start with a confession. I don't think I really have a solution to Jerry's problem, but I'll muddle my way through, but I'll at least try to point out how my views on this have changed, and one possible solution.

I used to think that it was clear that concepts were embedded in theories, and that they got their structures from theories, and that theories are what made them up. I thought this, because I saw powerful links between conceptual change and theory change whether it be in the history of ideas or a particular child, concepts seemed to travel in groups. When one child has a particular kinship term that shifts in development, many others ones shift at the same time in terms of what they mean to the child. Many concepts seem interdependent – I don't know how you can have the concept of NUT without having the concept of a BOLT, BUY without SELL, MOTHER without CHILD, and so on. They seem to be part of a larger relational complex that gives meaning to them and makes up their meaning. thirdly, notions how and why seem to influence all aspects of concept acquisition and use. Often what's very striking is that how often feature occur with instances – that is what makes up prototypes (or syndromes, whatever you want to call them) – that very young kids are not weighing things on the basic of their typicality. The degree to which they think something is causally important is causes them to [audio failure] something that seems like a theory. And I thought where concepts come from in development is by talking about growing webs of belief. Concepts weren't nodes in this network, they were clusters of nodes and links, like one of these little circles. And then as the network grew, a different little cluster meant a different little concept, and they got bigger yet, you got another concept and so on. And there is the whole story – as the web grows, so do your concepts.

But there were problems. And I tended to brush most of these problems under the rug, or thought I had a solution to them. There is the lost in thought problem: wouldn't theories be too slow? Would we get lost working through all these theoretical implications? Yet we use concepts quickly and effortlessly. Isn't there too much change in the theories surrounding concepts, that when the concepts nonetheless changes, things still stay the same? [audio failure] I might talk about more and more links come in, but how do new nodes emerge? That gets backs to Jerry's problem, and I haven't solved anything.

Well, I thought there were ways through all this. For example, there is actually interesting work of how you precompile information using a theory, and then can use it very rapidly. It may be that intuitive folk theories have not changed nearly as much as formal scientific theories. And so the change issue is not as big as it seems, and so on. But all that is beside the point.

Because, when I started looking at what the theories look like in more detail, they weren't what I thought at all. In the last few years, I've really been quite surprised that our intuitive theories are far weaker than we think. Not only are they weak, but they are even weaker than we think they are. I call this the Illusion of Explanatory Depth. We grossly overestimate how well we can explain things.

You can show this very simply by asking someone how well they think they know something works, and then ask them to explain and they fall apart very quickly. Assessing our knowledge of how well we know the facts, procedures (such as how to make international phone calls – I am about so-so on that), and narratives (how well they know the particular plots of books or movies). But they are very bad for estimating their knowledge of how and why. Now, that creates a problem, because if the theories are getting weaker, how much work can they do?

We can retreat to talk of framework theories, or core theories, and kindred kinds of notions, but then we have problems. If you look at some of the theories that are out there in the literature as ascribed to young children – they are at best sometimes 3 nodes and 3 links. The very very young child's theory of mind consists of "I've got desires that cause me to engage in actions" That's it! A little older – "I've got beliefs, that cause desires, that cause me to engage in actions" Very early folkbiology is: "I believe in a vital force, that vital force helps me move, and if there is some left over, it helps me grow" Now if that is all there is to theories, they are not going to a lot for us in terms of articulating the structure of concepts.

Moreover, as had been pointed out earlier, we have a high tolerance of contradictions. As has been shown repeatedly, people can believe rather large chunks of information that are completely contradictory to each other and not realize it until its explicitly pointed out to them.   Bill B__ has some wonderful examples of this in college students.

Well one extreme reaction is to say, well, there is no overall linking structure, knowledge falls apart into little tiny pieces, this would be like Andy diSessa's notion of phenomenal p-prims. But I don't want to go that far.   I think there are ways we can talk about a more relational structure, but it just can't be like a traditional "theory".

So, here is the problem for me at least. There is - and I think this is true for many of us – no theoretical difference between lions and tigers.   I know they are different, I think I know they mean different things, but I cant give you a theoretical reason that distinguishes them. How then, can concepts be created out of or made of theories?

Do we have to move to the notion of *atoms*? Well, I want to resist that, if only because I think we have concepts traveling in groups, with the notion that they can be mutually parasitic off each other for their meaning, and with this critical notion of the centrality of how and why – why it is that features that co-occur equally or correlate in certain ways are ignored or attended to because they fit with some notion of how and why.

What I want to suggest is that we *do* track causal and relational structures in the world in a way that is less theory-like -- perhaps even prepositional -- but critically supports concept acquisition and use.

What are some of the ways that we do track causal structure? It's a long story – I'll just mention a few. One which is very simple is we know what kind of property types are like to do important causal work in a domain. So very young children, infants, and even some other primates seem to know, for example, that when you are talking about tools, shape is going to matter more than color; when you are talking about foodstuff, color matters more than shape. There are these causal relevancy profiles that very young kids, as well as some other species seem to be aware of, that at least in humans I think they know are causal in nature.

Another example, coming out of our own lab, is that infants that intentional agents – these are pre-verbal infants, say, 11 months – are the only things that can create order out of disorder. If you have a pile of disordered blocks, barrier comes up, comes down and they are ordered – they only think that an intentional agent can do it, not something unintentional, like a rolling ball. Reverse the order, and go from order to disorder, and both agents can do it.

Finally, another example. Young preschoolers, in a study we just wrapped up, we've shown that kids as young as 3, if you ask them to ask questions about novel artifacts, and novel animals they have never seen before, they approach them very differently in terms of the kinds of causal regularities they think are at work. So for a novel artifact, they are very likely to ask what the artifact as a whole is for: "What's that for?" The spontaneous questions about novel animals – they are unlikely to ask what the animal as a whole is for, but they will ask about what parts of it are for: "What are their claws for" or "What is this for?" even if they don't know what the name is. They seem to have quite sophisticated expectations about the kinds of relational and causal patterns that go with different domains. And one of the things I want to suggest is, they use such notions to guide their intuitions of the division of cognitive labor. So even if they don't know who knows what, they know there are different kinds of experts out there that they can defer to, and that's critical to how they set up concepts when they have almost none of the details themselves.

So maybe that means we should think of concepts as some kind of chimeras. They are not simply prototypes, they are certainly not definitions, and they are not theories. There may be some rich relational structure that is part of all this story. I think that locking is going to be critical. I think Jerry is right, that somehow, we have to have this idea that we lock onto objects that are often not going to have an underlying supporting propositional structure.

But children are surprisingly sophisticated at linking abstract causal relational patterns to broad domains, such as social interactions, artifacts, intentional beings, mechanical agents and the like; and they use those, to guide categorization, deference and learning.

I'll give you an example – there are all sorts of examples – this is mine. I don't remember now, since I yanked these from the web, which one of these is a weasel and which one is a ferret. One is one, one is the other. But I believe I have both those concepts. Well what does it mean that I have those concepts, since I have absolutely no idea what the difference is between them? It's because I think I know who knows. I think I know who the appropriate experts are and how to access them.

Now part of this may be the notion of "sustaining mechanism" (which I'm sure we'll hear more about later today) – I just want to briefly say that this idea of a mental operation that enable our concepts to lock onto the right sorts of things may be critical, and where much of the cognition is at work.

Now Eric Margolis in one papers talks about 3 kinds of sustaining mechanisms. Those that are theoretical, and allow you to lock onto objects, those that based on deference to experts, and those that are based on a syndrome, or something like a prototype. What I want to suggest is that all 3 are at work in most cases. Our theories are too weak to work on their own. But often when we decide whether it's a ferret or a weasel, what we are doing is having a crude notion of who the right expert is, we then use that to help us defer, and we also then use that to determine which features of the syndrome to attend to.

The critical question, which I am really confused about, and why I don't really have an answer here, is whether the sustaining mechanisms are part of the concept itself, as opposed to just a tool that helps us lock (and I think that's what Jerry is going to say). That basically, they are like a microscope, like what we use to see bacteria.

But maybe not. It may be, that even for microscopes, part of what it means to have the concept BACTERIA is what kind of tool a microscope is. Its not just a -blank- tool – you have to know, for example, that it is a way to get information about invisible microscopic structures, that it has some causal efficacy. And that may be critical to my concept of BACTERIA. So also for experts.

You just can't point any expert at any object and expect to get the right answer. You have to know what kind of expert you are talking about. You have to know that there are different kinds of experts who have different specializations in different causal regularities. And one of the things that has surprised us is how young children are sensitive to this – even by the age of 3 or so, they start to know there are different kinds of experts out in the world. Not everyone knows everything. And that they have to have some mastery of the causal structure of the world to even be able to engage in the notion of deference and the use of expertise.

So: How do we acquire the word *carburetor?* Here is where I'm really not sure, but I'll give it a stab. It's especially interesting because soon *carburetors* will no longer exist – they are vanishing and its all fuel injection (but that probably makes them more interesting). We might hear the word, and we might hypothesis test whether it's an artifact or a natural kind. And I am more than happy to think that the notion of artifact is innate and maybe the *simple* sense of natural kind, not the more complex sense. We then quickly map it onto the artifact domain; there are lots of heuristics to tell whether something is an artifact or not. And then – this is the interesting question – there is whether there is a notion of differentiating sustaining mechanisms. That, initially, our locking is so crude, that we really can't have different kinds of concepts, and that what it means to have differentiating concepts, is to have differentiating sets of sustaining mechanisms. Those are what allow us to be more and more successful.

And they may not proceed by hypothesis testing. They may proceed by getting more and more sensitive to the kind of causal patterns that define different kinds of experts, how to pick out different kinds of regularities in the world. So I am not sure, the concepts might still be roughly atoms, but the sustaining mechanisms will be so linked to them that we may not want to separate them.

So – how do weak theories strongly constrain? It may be that if you violate these abstract causal patterns, you don't have the concept. Someone who thinks that carburetors have microstructural essences, that they have no overall function, or are non-physical, I don't think does have the concept. They could think – they could have the wrong shape, the wrong local function, the wrong material substrate. They'll do it differently if its have a living kind – they'll have different expectations. Weak theories don't tell lions from tigers. But they may provide guidance to deference and ways of access to information. They guide construction and maybe the differentiation of domain-specific sustaining mechanisms.

I don't think I really have a solution to Jerry's problem. I don't have a good way of telling what's in the concept proper vs its distinct enabling cognitive structure. But there seem to be powerful constraints in causal relational patterns that we see as fitting with very high-level domains, and this may be related to what Jean was talking about, and we'll hear from Jay later. Things like living kinds, artifacts, and agents. They are not like traditional theories. And maybe we'll see some sense of a different way of thinking about things from Alison that will work with this. But perhaps they're not always propositional, they are more relational, more like what Dedre refers to (I'm trying to bring everyone in here). But we have to learn how to use, take equally typical features and weigh them differentially, to guide locking. So perhaps its part of the concept after all. And that's the question that I think we have to address.

Discussion on Keil

Jerry Fodor: I actually agree with almost everything that Frank said. It's nice to agree with somebody for a change. But it did remind me a little of one of my favorite jokes. So there is there is this lady standing by the window of a tall office building and this body goes hurtling past the window. And the lady screams! And the person hurtling past looks up and says "Don't worry lady, I'm alright – so far!"

I think Frank is on a slippery slope the bottom of which is atomism. I don't think the atomism is avoidable. There isn't any criterion that will tell you what goes in the concept. And since there isn't, the only conclusion is that nothing does – that's the atomist conclusion. Only the locking matters.

Um, maybe that's … excuse me a second… hold on. Oh, yeah… I think this issue – what is in the concept (if anything is) – as opposed to the relation between the concept and the world. This issue has been kicked about in the philosophy of psychology circles for a very long time. But I think there is now a knock-down argument. And I am not going to give it – I'll just refer you to stuff that Ernie Lepore and I have been doing. In particular, to a book called *The compositionality papers,* which is a collection of papers on compositionality. The basic thought is, look, concepts have to combine to give more complex concepts. You've got to be able to get BROWN COW out of BROWN and COW. That's again, not negotiable.

But it turns out, that test procedures, like consult an expert, use a microscope or something, don't compose. There isn't any way of putting together a bundle of test procedures for brown and a bundle of test procedures for cow thereby deriving – algorithmically, as you would have to -- a bundle of test procedures for brown cow. That, I think is just *clearly* true, when you look at the examples.

Given that it is clearly true, as I am strongly inclined to suppose, it rules out the concept having any content which is concept constitutive. Basically, the story has to be atomistic, and the story has to be atomistic for what is after all not a very surprising reason, which almost everybody would have accepted until Frege; Namely, that the basic semantic relation is reference, and reference is an atomistic relation between a concept and the world.

Frank Keil: I think one question that I am puzzling with is whether the locking mechanisms themselves can differentiate and get more and more refined. And if that is a reasonable notion, especially if the cognitive architecture that underlies locking is not particularly propositional in nature, whether that pattern differentiation is a way to think about how we acquire new concepts.

Jerry Fodor: Well, look, all I can say is, I think its certain that locking mechanisms don't compose. Hence they can't be concept constitutive. Now that's an argument that obviously needs spelling out. But its in print, and I would bet up to a nickel on its being sound. That's a lot by my standards!

Jesse Snedeker: Can I jump in? Because one of the questions that we had when we were reading I think relates directly to this. You also say in *Concepts* that locking mechanisms will turn out to be in the domain of neuroscience rather than cognitive neuroscience. That they are presumably cognitively complex, maybe not even amenable to psychological investigation. And, honestly, that seemed to come out of nowhere. So I am wondering what the justification is, because it's really critical to the kind of solution that Frank is offering.

Jerry Fodor: The question arises, whether or not you can give generalizations over – here's a way to put it: Look, its widely understood now, by just about everybody except connectionists, that intensional states can be -- typically are -- multiply realized. You can't give the conditions for intensional state in neurological terms. Well, the converse argument sort of goes in the case of locking of concepts. The locking mechanisms are themselves are *extremely* diffuse. Very different, heterogenous, exactly as Frank was saying. They go from telescopes, to astronomers, to asking your grandmother, and so forth and so on. There isn't any generalization about the locking mechanism which can be given in terms of locking of an intensional object – it must be just neurological roots.

Jesse Snedeker: But isn't Frank arguing that there are classes of them, that if they aren't rule-governed, that they are least fairly predictable, one for natural kinds might be …

Jerry Fodor: Yes, but the trouble is that they don't compose. The criteria for being in class don't satisfy compositionality. Under any circumstances, you can't combine the mechanisms that lock BROWN to the world with the mechanisms that lock COW to the world, and infer the mechanisms that lock BROWN COW to the world. Your way of finding out whether something is a brown cow can have little or nothing in common with finding out whether it's brown and whether it's a cow. That being so, these locking mechanisms can't be themselves be part of the concept. Well, what *could* they be part of? Well, they could be part of a neurological chain.

Frank Keil: One quick comment. Again, I leave open whether we can put them into the concept itself, because I find your arguments compelling and disconcerting. But I do think we don't want to underestimate the cognitive richness and complexity of the locking mechanisms. I think the notion of essence placeholder, which has been very popular in the literature, is misguided. We have much richer notions of what essences are like, even when we are lacking in the details.

Jerry Fodor: I should say, though, that in a way, perhaps you ought to on my side, because if you take the locking mechanisms as part of the concept, then you have got to tell a story about which part, roughly speaking, is analytic. If they are outside the concept, then you can lock to the world any damn way you please – and I think that's right. The standard way of locking to the world is asking somebody what the answer to "what's that?" is, and paying him if he gets it right.

Stephen Laurence: Suppose the locking mechanisms aren't inside the concept. Then you don't have the problem of compositionality. What you get is that the locking mechanisms set up the concept, and you can acquire them much the same way Frank suggested, which I think is exactly right. And beyond that, compositionality takes over.

Jerry Fodor: But notice what you're saying. The intensional part – that is, the part that has to do with the content of the concept – the part that says, look, BROWN refers to brownness, or that BROWN is the concept that refers to brown things. That part, is no longer in intensional psychology.

Stephen Laurence: I don't agree with that, but –

Alison Gopnik

University of California - Berkeley

Alison Gopnik: I also am not sure that I have a solution to Fodor's problem, but I have a solution to a problem which I think is perhaps more important and I hope will suggest this is the real problem, as opposed to Fodor's problem, by the time we are done – though I think its related to Fodor's problem. Fodor's problem is really an application to general cognitive development and conceptual development of an older problem, I think, in the philosophy of science. That older problem is a problem that's been called "the logic of discovery".

The conventional wisdom in the philosophy of science for many years was that while we could have a logic of confirmation – we could say something about what could confirm hypotheses – we had no logic of discovery. We had no way of actually *getting from* empirical evidence that we had in the world *to* a vocabulary of hypotheses. That was the conventional wisdom. And that's part of the reason why learning seems to be such a hard problem.

What's happened over the past fifteen years or so, perhaps the past 10 years, is that within the philosophy of science and computer science and statistics, we actually *have* discovered a logic of discovery, at least for some particularly important kinds of hypotheses – namely, causal hypotheses.

What I am going to do is very briefly, in 3 minutes, outline a little bit about how this solution to the problem of the logic of causal discovery actually works. And then I am going to say something about how it might apply to problems of concepts, and then I am going to just gesture at the fact that we've recently found that young children are in fact using the same kinds of algorithms and learning mechanisms that are proposed out of this philosophy of science and statistics literature.

Ok. The solution involves representations that are directed graphical causal models, aka Bayes nets and these are representations of causal hypotheses in terms of variables and causal relations, represented by directed edges, in graphical terms – that causally connect those variables. The important point about this representation is that it's systematically related by a couple of important assumptions to patterns of conditional probabilities of events in the world and patterns of outcomes of interventions on effects in the world. And this systematic relation between the structure of the hypotheses – these hypotheses represented as graphs – and patterns of conditional probability and intervention, means that you can actually make systematic predictions about patterns of evidence from these hypotheses. And you also make systematic predictions about what will happen if you intervene in the world in particular ways, based on these structures.

Perhaps the most important thing, for the current discussion is the fact that there are these systematic relationships between the hypotheses – these causal hypotheses represented by variables in graphs – and patterns of evidence – means that you can also do the *inverse* problem: you can say, given this particular pattern of evidence, what *must* the underlying causal structure be? And is *that*, that makes these systems enable you to actually have a logic of causal discovery.

By the way, this is work that has been done by the computer scientist Judea Pearl, my colleague and philosopher of science Clark Glymour and his colleagues at Carnegie Mellon University.

Ok, let me briefly just say a little bit more about how these systems work. So one way of thinking about all of these systems that is relevant to the discussions we are having today is that causal Bayes nets give you a way of mapping hierarchical componential systems – these causal graphs – onto the kind of data that for example, connectionists have traditional paid attention to: statistical data about patterns of conditional probability. What these systems do is to take a couple – in fact, 3 – simple assumptions and use those assumptions to map a hierarchical componential structure onto patterns of statistical evidence. And I'm not going to detail about these assumptions.

The first assumption is called the causal Markov assumption. And, given the causal Markov assumption, you can take this kind of representation where these variables can be anything – they can be linear, they can be non-linear, they can be any kind of thing – any way of categorizing an event in the world. And the relations can be any kind of relation, provided it's a causal relation. What the causal Markov assumption tells you is: *if* you have a particular causal representation of the world, then certain kinds of conditional probabilities are going to hold among these various kinds of variables; And other kinds of conditional probabilities are not going to hold among those variables. In particular, for example, given this structure, you can tell that S is going to be independent of W conditional on X. And you can generate a whole variety of these kinds of predictions. These are *normatively* accurate, given the assumptions. That just logically has to be the case.

You can also make assumptions about interventions. What makes these representations *causal* is they not only tell you about patterns of conditional probability, but they also tell you a rather different thing, which is: what will happen if you actually intervene on the world, if actually *do* something, and wiggle one of these variables – what will happen to the other variables. And they do this by making another set of assumptions (that I won't go into) about interventions. And essentially the way this works is you think of an intervention as being a cause, its coming in from outside of the system and perturbing it, and that intervention fixes a variable to a particular value. And the result of that [computer error] is that arrow from S to X will disappear as a result of that intervention, and you'll actually have a new graph. And now you can say ah, if I had intervened, then here's what would have happened. Not only does this say something about what will happen if you intervene in the world, it – exactly this machinery – allows you to generate counterfactual predictions. Judea Pearl has outlined this quite beautifully – so you can say, if I had done X – or if the world had done X – then what would the result have been? So not only can you make predictions about what will happen, you can also make counterfactual predictions about what would happen if I did X, or what would happen if I had done Y.

Between those two basic fundamental assumptions, means you can also go backward. You can solve as it were the *inverse* problem with these representations. And what you can do is take the patterns of independence and conditional independence that you see in the data, and normatively, logically say, *only* these kinds of causal hypotheses are compatible with that kind of data. And there are computationally tractable search algorithms, which are provably accurate in the limit, which enable you to take patterns of data, and say, given this pattern of data, *only* this hypothesis can be correct. That's why it's a logic of discovery. It is not any more that you simply have to start out with a set of hypotheses and test each one of them.

I want to emphasize that these models are *normative* models. They didn't start out with anything to do with psychology at all. And I think a very interesting and fruitful analogy is the way that we've used information in logic of vision as it were in vision science. So in vision science – probably the most successful aspect of cognitive science – what's happened is that people have discovered abstract normative principles about what a visual system *must* be like if its going to accurately recover information about the world. By making some very general assumptions about what the spatial structure of the world is like, and how it gives rise to patterns of visual evidence, we can have normative theories about what a visual system has to do to recover spatial information from patterns of evidence. In vision what we do is we have geometry, which tells us about 3-d object representations, we have optics, which tells us about how those 3-d object representations must be related to visual data, and then we have ideal observer theory, which tells to assume that in fact caused by those patterns out there in the world. And that enables us to in fact recover the structure from the evidence.

In exactly the same way, the theory of causal Bayes nets gives you representations of causal structure, in the form of acyclic graphs. It gives you principled assumptions about how those patterns are related to patterns of data and evidence in the world – particularly, statistical evidence. And then, one assumes that the evidence was caused by the data, and that enables you to normatively recover the causal structure from the data.

Let me mention that I think that these representations really are representations that -- the reason why I got interested in this in the first place – and this speaks back to Frank's point – is that these capture most of what's important about what we mean by intuitive theories in cognitive development.

So although I think these representations – and I think Frank's point is relevant here – seem much sparser than what we've traditionally thought of in intuitive theories, I think these do in fact do all the work we would like intuitive theories to do. In particular, they allow us to go out into the world and make accurate predictions and correct interventions and correct counterfactual causal claims about what the world is like. And I think that's all you really want a theory to do.

So far what I have shown is that we have a way of representing the causal structure of the world, it seems to me that what theories are is representations of the causal structure of the world. And that that system can be normatively accurately learned, provably learned, from patterns of evidence that we know we actually get in the data – namely, patterns of conditional probabilities and independence.

Everything I've talked about so far is just mathematics, nothing to do with psychology.

Now, you might ask, what does any of this have to do with *concepts*? I think it's got two things to do with concepts. First of all, what Bayes nets do is specify the causal role of those variables (those nodes). And, at least on *some* views of concepts, what you really want to *know* about a concept is *precisely* what its causal role is. So once you've specified a variable – here is a variable and here is its causal role. That is thing you want to know – to know what concept that variable belongs in.

But there is also a stronger claim, which also relates back to Frank's point. Which is that Bayes net learning algorithm can also normatively imply unobserved variables. Not just unobserved variables, but unobserved causal structure, and they can lead you to split or combine existing variables.

So given certain kinds of patterns of evidence, what you can conclude is, none of the variables that I can see in this net are correct, there must be something else that I *haven't* seen, that has a *different* causal structure than any of the variables that I have encountered so far, and that thing is responsible for this pattern of causal structure. And again, you can do that, normatively.

I won't have enough time to say exactly how that is done, but trust me that you can.

So there are 2 senses that you can think about Bayes nets as specifying concepts. And I want emphasize – all the stuff that you'd want, everything about concepts that you could want. There's probably a whole bunch of other stuff about concepts, which has to do with deference and sustaining mechanisms and locking mechanisms. But I'm not sure that's stuff that you actually need to get concepts to do the work that you want. Namely, work about prediction, intervention and counterfactual reasoning.

This is all very well and good theoretically – it is nice to know that there *could* be a solution to the problem. Is this a solution that we use, or that children use? What we've shown empirically so far is that in fact, preschool children – 3- and 4-year olds – can indeed infer causal structure from conditional dependencies in ways that are consistent with this. In more recent work, including work that I talked about this morning, with Laura Schulz at MIT, we've shown that they can infer unobserved causal variables appropriately and normatively, from this kind of data – including unobserved causal structure.

What does this set of ideas and solutions, how does this relate to the other kinds of approaches? What's important, and what makes these learning solutions different from associationism, connectionism, analogy, generalization – all the kind of traditional candidates, the usual suspects– is that this kind of inference involves rational, normative inference. So one of the points is that – the most crucial point that Jerry mentioned before is – is this actually learning or is it just something like triggering? And this system provides you with a rational means of induction – a kind of inductive logic – about when you should draw the right kinds of causal conclusions, given certain kinds of data. That's not true about any of the other solutions that people have proposed.

It’s also true that this kind of learning is much more constrained than classical associationism and connectionism and so forth. Its much more general than triggering or parameter setting, but it *only* applies to causal structure, and it only applies if these particular assumptions about relations between causal structure and evidence are true. Just like ideal observer theory or signal detection theory only applies if certain kinds of assumptions about the relation between the relation between the world and perceptual data and evidence are true.

So it seems to me it sort of just right – it gives us something that is much more general than a kind of triggering or parameter-setting solutions that people have proposed. On the other hand, its much more constrained than association, generalization, analogy, differentiation, all the rest of the guys. In particular the way that its constrained – the fact that its constrained in terms of these assumptions about the relationship between structure and statistics, means that its constrained in just the right way to ensure that you actually get the right solution to the problems.

So what I think I’ve argued for (at much more length in other places) – is that we have a way of getting the kind of structure -- causal representations of the world which enable the right kinds of predictions, that we would like our concepts to have. And what’s more is we have normative, rational ways of learning about that structure from the kinds of pattern of evidence that we know that we have – patterns of conditional probability and dependence. And, we have evidence, that, as a matter of fact, that even very young children are indeed, using those learning mechanisms.

Discussion on Gopnik

Jerry Fodor: You should never believe anything a Bayesian tells you. (laughs) – I take it back. That’s an answer, maybe it’s *the* answer, to the wrong question. The problem about concept formation is different from the problem about belief fixation. What we have here is a model – either a good or a bad one, I am not in a position to judge – of when a belief is rational, to put it normatively, relative to a certain body of data. But to characterize the body of data, or the content of the belief, you have to know what the data are data about, you have to have a way of describing them, and you have to describe the hypothesis in the same terms that you use for the data. If the data are about noses, then the hypotheses have to be something like “how many noses does a person have”, or something like that.

The problem of concept formation is *prior* to the problem of belief fixation. The model that is being suggested is in fact *presupposing* the very worry I have in fact been trying to raise. You can see that it in fact *has* to be true. The model is purely formal. It doesn’t *care* what the values of the variables are. So it doesn’t distinguish between any two causal concepts. Presumably we have more than one causal concept. Presumably. But this model doesn’t distinguish between them. It tells you what it is for the concept to count as causal.

But we want a model that least distinguishes between “breaking the bottle” and “destroying the city”, or breaking the bottle and building a house, or something of that kind. All the causal concepts -- simply because the schema is characterized by the variables -- satisfy it equally well. If you want to look at it that way, it’s infinitely too coarse-grained to give you an inventory of concepts.

Alison Gopnik: Ok, so there’s two things to say about that.

The first thing to say is that, one thing that this does give you – and again, I went over this much too quickly – it gives you a potential for having a actually having a constructivist solution for having new concepts. So if you identify the variables with concepts – which seems like one way you could solve the problem – that is, those nodes are what we are going to call concepts. We have every reason to believe that we are going to start out with some innate set of variables, in just the way that Jerry would propose. But we actually have *procedures*, given this kind of formalism, that actually lets us go from that innate set of variables, lets us get more information about those variables, and lets us propose *new* variables, lets us propose *new* nodes that have different kinds of causal relations. That's the first thing to say.

The second thing to say, which I think is more important, and this is in some sense an empirical question, is it seems to me at least plausible that much of what we actually mean when we talk about conceptual structure – what's important and counts to us about concepts – is exactly their formal structure of causal relations to other things.

Jerry Fodor: That can't be right, because all causal concepts, ipso facto, have the same formal structure.

Alison Gopnik: No, they don't.

Jerry Fodor: But they all satisfy the same axioms!

Alison Gopnik: But they all have different graphs. I mean that's exactly the point. The point is that what's going to count as being the concept is … Its not the labels on the nodes which distinguish between the graphs, it's the *structure* of the graphs that distinguish between the graphs.

Jesse Snedeker: So, if that's the case then, then how is that these concepts get tied to the world in any way. How do they have reference solely by their links to the other ones, if they are individuated solely by their links to each other?

Alison Gopnik: That's a good question. Again, there's 2 kinds of answers to it that you could give. One answer is that these variables (and this relates to Jean's point) – are in fact going to be related to other information like spatial information that you have in the first instance. So in the first instance, I think exactly what's going to happen is you are going to be using for example spatial relations, among other things when things are part of the variable – at least to begin with.

But then there is this potential kind of bootstrapping mechanism that you can have, where those things are going to become, as you actually see more relationships among the events that you see in the world, you're actually going to be proposing new kinds of variables.

Jerry Fodor: This is very important, because the same issue comes up in discussing connectionism. If you have two graphs that have the same shape, you can't tell what concepts the graphs are grasping. So you can't distinguish, for example, the causal relation between rhinoceroses and mice – if rhinoceros are afraid of mice, and the causal relation between trains and people – if trains run over people. That you need to know not just the *form* of the concept is – what it is with variables, you have to know what the values of the variables can *be*. In particular you have to know that to be able to describe the data, because data about rhinoceroses are not ipso facto data about subway trains. That's what it costs you to have a formal theory. What you are saying is very like saying: look, first order logic is a theory of concepts, because it tells you the structure of rational arguments. They happen to be deductive, not inductive. But first-order logic does tell you the structure of certain rational arguments, but it doesn't distinguish one application of modus ponens from another. You extract from the conte nt – you extract from the concepts -- exactly the same thing holds here.

Alison Gopnik: So the question really is – one of things that's powerful about these systems is the idea that in fact the structure of the graph —the way that concepts are related to other concepts – and things about the parameterization of the graphs too exactly what you think the relations are --

Jerry Fodor: You have to say what the other concepts are. Its no good to say that the difference between rhinoceros and platypus is that rhinoceros is related to girls and platypus is related to boys. That's won't do because you are going to have to have a characterization of the relation. That's what its not giving you.

Alison Gopnik: Lets think of an example actually like scientific theories, rather than intuitive theories. One of the things that is in fact is characteristic of scientific theories which one would assume are conceptual is precisely that if you want to characterize what the concepts mean you characterize them in terms of the sets of relations among the concepts.

Jerry Fodor: Empirical theories are empirical, not conceptual. That's what the problem is. Empirical theories, use (among other things)conceptualizations. But the generalizations they are after are typically empirical. That's to say, they apply to world in virtue of their *content*, not of their *form.* Now there is no way out of it – the way to think of this I think is: why don't you make the same claim about deductive logic. Why not say, look, modus ponens is constitutive of concept, so we know what the concept is – it's the one that undergoes modus ponens – the trouble is, *all* concepts undergo modus ponens. If you want to know what's constitutive of this concept, you have to know not just that it undergoes modus ponens, but that it undergoes modus ponens in such premises, as if something is a dog, then its an animal. But you have no way of saying that – all you have got is variables. What you need is constants.

Jesse Snedeker: Frank, would you like to jump in at all? It seems to relate to some of the issues that came up in your talk.

Frank Keil: One way -- I'll try to be a peacemaker -- could be to try to argue that the different causal patterns that you pick out are what you associate with large content domains. So you know that this domain of artifacts has a particular causal pattern, and this other one has a different causal pattern – that they are distinctive – and they aren't really constituting the concepts per se but they are enabling these different kinds of locking mechanisms.

Jerry Fodor: That doesn't work because of the compositionality argument – both of these balls have to be kept in the air in order to make any progress. The causal role of a brown cow is not predictable from the causal role of cows in general and the causal role of brown things in general. So you can't take causal roles – even if you wanted to – as concept constitutive.

Alison Gopnik: I think part of the problem is that there is a sort of confusion that we have been led into by language, which is thinking that the kinds of things that are going to be picked out as same or different things by language are actually the things that are going to be doing the kinds of cognitive work that we typically want concepts to do. So it seems to me the kind of cognitive work that we want concepts to do is not to go out in the world and be able to identify this word applies here and that word applies there …

[audio failure]

Sourabh Niyogi: One of the persistent intuitions that I think many of us have is that concepts are relational – for example, between NUT and BOLT – that you can't have the concept NUT unless you also have the concept BOLT, or that you can't have the concept BLICKET without having the concept BLICKET DETECTOR. Concepts appear to be interdefined in terms of each other, and so we expect our theories to contain a characterization of these relations.

Jerry Fodor: It doesn't follow from that you can't have the concept unless you know the relation. That's a very strong and very specific characterization of what the possession conditions are, and why I want to explicitly deny – what I want to say is, look the possession conditions – about being able to think about pears – what you're thinking about is a relational concept – if you can think about nuts – you have got the concept NUT – but of course it's a relational concept – that's as it were, a piece of metaphysics – not a piece of semantics – its part of the story about what a NUT is, not part a story of what the concept NUT is, or what it is to have the concept of a NUT. Now if you think that's wrong, you owe an argument.

Frank Keil. I don't understand this – this is a critical problem for me, because that's why I am trying to the sustaining mechanism back into the concept because I think it might give us a way to put nuts and bolts together. And so, I think we agree, the metaphysics and epistemology part is tricky for me, but I think we agree, that if you have the concept NUT you must have the concept BOLT.

Jerry Fodor: No, I don't agree with that. What I think is correct is: if you have the concept NUT, then you have the concept OF something, which is essentially related, in a certain way, to bolts – that's true. It doesn't follow that to learn the concept is to learn a relation. Again – I think this is why this stuff is so fascinating. All the skeletons come out of the closet. *One* of the skeletons – that's been haunting this (if skeletons can haunt) discussion for 100 years is the assumption that metaphysical truths are ipso facto conceptual or semantic. That nuts are used to screw on bolts (or in whatever in hell they are used for) is a fact about *nuts*, not a fact about a *concept*.

Alison Gopnik: But -- just about metaphysics and epistemology. The whole point about having a cognitive system is to capture things about metaphysics. Having a conceptual system that actually isn't designed to capture the metaphysical structure of the world would be really stupid …

Jerry Fodor: Having a discussion about what cognition was evolved to do – or what it is "designed" for – any of those kinds of things -- what they are doing is whistling in the dark.

Alison Gopnik: But all of vision science does that.

Jerry Fodor: Wait. What we have is a system that allows us to arrive at true beliefs. Whether that's "designed for", whether that is "evolved" – God only knows – certainly, no finite living mind does – what we know is that if you have the concept NUT, it allows you to formulate the belief that nuts have a relation to bolts. The question now is, is that belief a conceptual truth? Or is it a fact about what it is for something – is it a metaphysical fact, a fact about essence, a fact about what it is to be a bolt -- One simply *cannot* beg that question. You are not simply given as a premise, that essential properties are ipso facto semantic.

Dedre Gentner: I'll try to be brief. Just to change this a bit – going back to "brother" – to me it seems relevant that if we are thinking about a learning mechanism for how to get stuff in there, starting at this very high relational level, doesn't really fit the facts really very well. So what kids typically think brother means, is "someone who looks a lot like my brother." They don't think uncles or guys with pipes. They have a very hard time figuring out that a grownup in a chair could be the brother of his mother and so on – and likewise for uncle.

Jerry Fodor: That's on my side, not yours.

Dedre Gentner: No, its actually on my side – you just don't understand my side thoroughly. It could be on your side too and that would be delightful.

Jerry Fodor: Look, the point is this – Anybody who has cognitive commerce with "brothers" ends up believing that brothers are siblings. That is common ground. What is at issue is whether that belief is constitutive of the concept BROTHER that occurs in such propositions as for example "Brothers are siblings". You have (roughly speaking, with lots of wrinkles) two choices: you can say: yes, its concept constituitive, so the proposition is conceptually true, or you can say, No, what you learned is a fact about brothers – it's a metaphysical necessity, not a conceptual or semantic necessity.

Now you simply mustn't beg the question, whether all metaphysical necessities are semantic. In fact, we know (unless Kripke and Putnam are wildly wrong, which they may be – but modulo that assumption), we know there are lots of necessities that *aren't* semantic. Well, now you have to tell me why brothers go with siblings and nuts go with bolts aren't that kind of necessity.

Sourabh Niyogi
MIT

Sourabh Niyogi: Ok, let me see of some of these issues might show up in this toy world that I am exploring here. I've been looking a lot of Fodor's commentary on lexical semantics—he's a very harsh critic of people who try to come up with meaning primitives – the standard picture for lexical semantics is something like the following: it's a very old architecture, dating back to the generative semantics of Lakoff and company in the late 60s, Schank in the 70s, Jackendoff Pinker and many others in the 80s. And its essentially the following; the project in lexical semantics has been: what are the concept primitives that span some conceptual space? How can a vocabulary item actually map onto one of these concepts.

I've been working on richer conceptual model that I think is strictly more powerful than I think the one that the lexical semantic folks have been working with – I'll call it the Universal Theory Model of concepts. A theory acquisition device works with some set of theory primitives to output some theory T*, there is some concept generator G that maps T* to a set of lexicalizable concepts. And what is interesting about this is: whereas the Standard Picture has a fixed set of lexicalizable concepts -- that is, once you know P (a fixed set of conceptual primitives), your conceptual space cannot expand. In the new model, you can have a variable input to the vocabulary acquisition device – that is, if your theory T* changes, then the space of possible concepts that you can reach also changes.

This is the kind of architecture that I think exposes at least some of issues that are at stake when Fodor says "you have a concept", because there's actually 2 viewpoints you can take. One is the viewpoint of the developmentalist – the developmentalist might say of an individual – well, what input does the VAD have? Well it only has those concepts accessible from the current theory T* that you have right now. The nativist might take a different viewpoint – the set of concepts that a VAD has is the union of all the concepts that are accessible from all possible theories.

Which viewpoint you take is largely a matter of perspective. So the first viewpoint says, well concept acquisition is possible, because you can get a new concept (or a new hypothesis) for your VAD just by changing your theory T*. The second nativist viewpoint says, well concept acquisition is impossible, because in a certain sense, if you take the union of all possible outputs T* from this TAD, then the input is basically fixed.

You can expose some of these key state variables – the theory T* and the space of possible concepts G(T*) – in a toy world. The model that I've been toying with is – a theory is a set of kinds, attributes, relations and causal laws. Theories are generative models of possible worlds, and by direct analogy to Universal Grammar, this theory of the TAD – or the theory primitives – I'll call Universal Theory. You can try to say that essentially a lot of discussion that people have had on "theory theory" all sort of fall into this general architecture and what different approaches are arguing about are what the format is of these theory primitives, what the possible constraints on this state space are. Some of these proposals I think are exactly on the money here, and try to give content of what those theory primitives might look like.

Here is the toy world – it's a toy world I started working with Josh Tenenbaum on – it was itself inspired by work by Alison Gopnik [and colleagues]'s "blicket detector" studies. In this world, subjects actually have to discover a theory, and they have to discover that there are these 3 verbs that underlies the causal laws that govern these blocks. So in this world there happens to be 4 kinds of blocks – BLOCKs, Bs, Ds, and Qs – and there are 3 causal laws that govern how these blocks interact with each other. So it turns out, that every B activates every D. Each of these BLOCKs over here co-light each other up. Every Q can activate every other Q – the Q with the higher internal property beta will activate the one which has a lower beta.

If you throw subjects into an application like this – let me show this to you in 10 seconds here [demo] -- subjects basically drag and drop blocks onto each other, and they are given a simple linguistic cue – like, "F is gorping D", and they have to discover there are in fact, these 4 kinds, but they aren't given any direct perceptual features – they have to discover that there are these 4 kinds, and they have to guess what these 3 verbs mean, just from playing within this application.

What they are tested on, in the middle of the application, is their knowledge of naming conditions – they are asked to touch one of the blocks to another one of the blocks, and they asked, hey, what do you think happened? Did Z gorp L? Did Z seb L? Did Z pilk L? And they give a forced choice response – one of 6 responses – and they are asked in the beginning and at the end of the experiment, just roughly, what they think gorp, pilk and seb mean.

And it turns out that not all subjects "get" the theory. Some subjects do – and you see this most clearly in how some subjects organize the blocks spatially -- and some subjects don't. There are T1 subjects, T2 subjects and T3 subjects. T1 subjects don't organize the blocks spatially at all – they just figure out that there is one kind, that blocks sort of activate each other at random. On the other hand, there are T2 and T3 subjects – and I'm not sure whether this "belief fixation" -- but T2 and T3 subjects figure out that there are Bs, Ds, Qs, and BLOCKS – the difference between a T3 subject and a T2 subject is whether they figure out that there is this internal property beta (underlying Qs).

What our challenge might be is: how can we figure out how their theories and their lexical items might map onto each other. Well if you ask T1 subjects, they are at chance on this naming task. If you ask them what is going on when Z is activating L, they pretty much can't tell gorp pilk and seb from each other. On the other hand, T2 and T3 subjects can figure it out. If you ask subjects for direct definitions, T1 subjects say well, *gorp pilk* and *seb* mean "cause to light up" while T2 and T3 subjects give something that is consistent with their knowledge of the (4) kinds.

So you can try to come up your set of theory primitives. These aren't Bayes nets, but they might be something that you might call weak theories. But you can make a distinct proposal about what Universal Theory might look like. In my system – I won't go the details – you can postulate sets of kinds, attributes, relations, and laws – and each of these sets are relational in structure. In order to define an attribute as a map from a kind to a space, you've got to refer back to these other sets (of kinds and spaces). Critically, these theories form a generative model for possible worlds – they say what can and cannot happen.

So you can situate each of the key state variables – the theories and the lexicon – within this architecture. You can ask: what do T1 subjects have, when they have knowledge of how Causal Blocksworld works? And you can describe them with these theory primitives. A T1 subject knows that there is just this one kind BLOCK; they know that BLOCKs are either lit, or they are not. They know that there is a contact or activates relation which holds between them. And, they only know about one kind of causal mechanism. And because they are stuck with that one kind of causal mechanism, they can only have 1 (verb) concept. They can only map these 3 lexical items onto that one causal mechanism.

On the other hand, if you've got to T3, you have knowledge of 4 kinds, and not 1 but 3 causal mechanisms. And because you are at T3, and have knowledge of these 3 causal mechanisms, you can actually map these 3 lexical items gorp pilk and seb onto 3 distinct concepts. And we can argue about these individual things are atoms, where belief fixation is happening exactly – well we can define what those terms mean exactly. These properties I think matches a lot of what the folks in theory theory have been asking for, of their conceptual structure.

Theories help you parse the universe. If you have a theory, say if you're a T3 subject, and you have some know of say a particular causal mechanism, and you observe some activation – say of Z activating L, formatted in some perceptual vocabulary. Well if you have some knowledge of the kind of object that Z is (that it is of kind B), the fact that you have this causal mechanism (law1), allows you infer that, well, L is of kind D.

Theories are generative models of possible worlds. A lot of the stuff that we've learned from grammar induction are inherited by theory induction. A lot of the techniques that Gopnik referred to are very relevant here. You can actually say, here is how we can acquire T1 or T2 or T3, given the data that is available to the learner. And it is not that we have unanalyzable, ineffable primitives like the lexical semanticists have had. We instead have generative model for possible worlds.

A lot of the stuff we've learned from lexical semantics has to be reinvented in this concept generator which maps these theories onto lexicalizable items. The simplest that works here is that you have a simple one-to-one mapping. For each causal mechanism, there is single atomic concept that corresponds to something that might be a possible word meaning – that you can map gorp pilk and seb onto.

You can have 2 perspectives on this, again – you can see how the theories, and possible lexicalizable items may map onto it – you can say, well the developmentalist can look at it one way – well the T1 subjects just have access to that *one* concept, whereas the T3 subjects have access to just those *three* concepts—that's the viewpoint that the developmentalist might take. The nativist, well he can say you have access to the entire space. That is the T1 subjects have access to just that space as well as the union of the entire thing!

And which viewpoint you take is really a matter of perspective. I think, in the ambiguity of "having a concept", you can dissolve the puzzle of concept acquisition, and ask instead a different set of questions: What are the theory primitives that span the space of possible theories? What concept generator allows you to map theories to the lexicon or lexicalizable concepts? What are typical trajectories T*(t) or initial state of T* (at t=0)? What are actual lexical databases that we can talk about building, that are theory-based in nature? And then there is the questions of standard cognitive psychology: what are the mechanisms that allow us to acquire a theory or acquire a word meaning?

Discussion on Niyogi

Jerry Fodor: I'm actually going to be very brief, which I think you'll glad to hear. As far I understand it, this model makes two assumption. One, I think, is that concepts are definable. And in that sense, the model is in the old tradition of Bruner, Goodman and folks like that. And second, that they are definable interior to theories. Well quotidian concepts are in general not definable, which is how we got into this trouble in the first place. And concepts are *prior* to theories, not *posterior* to theories. Concepts are what you use to state the theories *in.* They are related to theories in the way that words are to sentences. Now its been part of the pragmatist tragedy to deny that last claim. And one can continue to deny it if one likes. The trouble is, its true!

Sourabh Niyogi: So when you say I'm making a mistake in defining these concepts, I don't think I am making that error. I have a separate vocabulary for theories, where a concept over here is defined in terms of whatever this output is for this concept generator. I'm not making the lexical semanticists' mistake --

Jerry Fodor: No, but I mentioned Bruner Goodman and Austin with malice and forethought. There is a difference between a theory of concept learning and a theory of word learning. This is exactly the mistake that Bruner et al made. The theory of word learning assumes that you have essentially got the concept, and answers the question you are trying to answer: how do you figure out which word in the expressible vocabulary expresses it? But concept learning isn't word learning. Concept learning is prior to word learning, in exactly the way that concepts are prior to theories. And words are prior to sentence, by the way.

Sourabh Niyogi: Why can't we just say that what you call concept learning is just theory acquisition?

Jerry Fodor: Ah, because what I call concepts are what the child converges on. I don't really much care what path he uses to converge on it. What he ends up with are concepts in my sense. In what sense are they concepts in? Well, for example, in particular, they are compositional. Right? If you have definitions – if concepts are definitions, then you have compositionality – because definitions in fact do compose. If you have a definition of *cow* and a definition of *brown*, then you would be able to derive a definition of *brown cow*. However, the antecedent is false. You don't have a definition of *brown*, and we don't have a definition of *cow,* and that's, I take it, because the corresponding concepts are atomic.

James McClelland: It seems to me that there is something about this piece of work that I thought was appealing that I thought were at the issues that you were getting at. And that is there is a process at work here, that creates clusters. We certainly have to admit that each of the little scrabble pieces is individuated, to start with, but there is a process at work that clusters *them*, and those clusters are the things that become the constitutive concepts that then get labeled. Now whether you call it theory acquisition or not, I don't think that's particularly relevant. What I do think is interesting is that that process started with, you know, mere individuation of tokens, and came up with a clustering of them, which then could be treated as concepts in just the way you described. Namely, they are the inputs to the problem of figuring out what we label.

Jerry Fodor: No, actually they can't be described that way. That's the other half of the dilemma. Look, the clusters are in fact definitions, then they'll be compositional all right, so I have no arguments. The trouble is, as you also believe, they can't be definitions. So what could they be? They could *statistical* clusters. Which is I take it, all you guys think they are. That can't be right. Because statistical clusters don't compose. So that would be a lovely story, if only the predictions were true. But they are false.

Alison Gopnik: But Jerry, it seems to me that there's two different things that are getting confused. One is what kinds of things do you want to have in your representation to do the kind of work that you want concepts to do. And other one is, what are the kinds of things that map onto natural language and words. And its no particular reason to believe that those things are going to turn out to be the same.

Jerry Fodor: Sure --

Alison Gopnik: So it seems to me the kinds of things you are going to want in your representations that are going to do the kind of work you want them to do. Now what I would claim is, if you are thinking – that needn't necessarily be specifically Bayes nets – the range of things that are like those kinds of representations. What they are going to do is give you a representational system which will do the kind of work that you want. Those aren't statistical clusters. And its important that in those cases that those aren't statistical clusters. Those are abstract representations of variables and relations that are then *related to* statistical information. But there might be quite an indirect route from things that do say counterfactual support and the theories to the things that actually specify when it is that you actually give the same word, or when you don't give something the same word. And it seems to me you are assuming - - those things compose! Those variables compose beautifully. That is exactly the point – with those variables you're going to get a very clear compositional story about what happens with variable, in variables, and causal relations within the theory. It might very well be that the reason that why the browns and the cows don't compose is that they are the result of these other kind of mechanisms and processes which are only parasitical from, derivative from, the basic thing that's doing the cognitive work.

Jerry Fodor: I don't care about language at all, for the present purposes. The question is about the compositionality of concepts. But concepts have to compose. Can they compose according to that schema? No.

Niyogi. So that's where I can disagree. I can show you how "blickets gorp gazzers" compose exactly.

Jerry Fodor: That's because they are definable!

Jesse Snedeker: The claim here is that this doesn't characterize real concept acquisition --

Jerry Fodor: If you give me two definitions, I can show you how they compose. The trouble is, there aren't any definitions. If you give me something that there are, say, like, stereotypes or feature bundles or something, that's fine. People have – but they don't compose. That's what is called a dilemma. You can have one horn or the other, but you can't have both!

Jesse Snedeker: Jerry, how do atoms compose?

Jerry Fodor: Atoms? They don't: you don't get bigger atoms out of smaller atoms.

Jesse Snedeker: Well, no, but how do atoms subserve compositionality?

Jerry Fodor: Well, there are laws of molecule formation. I mean, why do you think I should deny that? Look, its straightforward, I mean, there is no hidden agenda or something. If you take the concept of COW (not the word, the concept), and you take the concept of BROWN – let's assume the former is a stereotype and the latter is an exemplar, if you like that kind of talk – there isn't, in the general case, any way of putting them together and getting the concept "brown cow". The only way of putting them together – and that includes set-theoretic ways of putting them together – but what we are worried about is specific to these concepts.

Ok, you have essentially two options. One option is to say, well, they do compose, but they are atoms. And then everything is fine. Brown is an atom, cow is an atom, and we know how to compose *that,* its just a set theoretic intersection.

Or, you can say, well, they don't compose, but that's because "brown cow" isn't compositional. The trouble with that is that it's false.

Stephen Laurence: I think an advocate of conceptual role semantics—and I think this would apply to Alison Gopnik's talk as well—I think that an advocate of conceptual role semantics should say that the conceptual role fixes the content of the atoms. And then once the content of the atom is fixed, then it composes …

Jerry Fodor: That can't be true, because conceptual roles don't compose.

Stephen Laurence: No, but the composition doesn't work by the composition of the conceptual roles. It works by composition of the contents, which are determined by the conceptual role.

Jerry Fodor: Then what is the functon that takes you from a conceptual role to a content? That is what conceptual role theory is lacking.

Stephen Laurence: Well, conceptual role theory just says, if you have got this conceptual role, then you've got this certain content.

Jerry Fodor: Well I'd certainly like to see an example.

Stephen Laurence: Suppose that have a certain conceptual role gives you the content "brown", and having another conceptual role gives you the content "cow", now you've got the content "brown" and the content "cow", those can combine via a compositional semantics.

Jerry Fodor: Not unless it's Christmas and giving is a mystery. I mean, I suppose there is some function boringly that maps conceptual role onto meanings. And if there isn't, it's because the notion of conceptual role is so undefined. The question is, if the meanings are themselves the conceptual roles – we know the meanings have to compose (that's not up for grabs) – if the meanings aren't the conceptual roles, what are they? They're not definitions, and they're not conceptual roles (by assumption), and they are not statistical bundles, because ..

Stephen Laurence: They are referential relations.

Jerry Fodor: But how do you get referential relations from conceptual roles?

Stephen Laurence: Well that is how I would read the conceptual role theory….

Jesse Snedeker: I think you'll get a chance to get back to that. What I would like to do now is: there are a couple of questions that I have gathered from people that they wanted to ask, to folks in this debate. And, after we do just a couple of those, I'll cut over and I'll promptly -- 5 minutes or so (say, 7 minutes?), then there will be chance for all of you guys [in audience] to ask questions. So I don't want to lose that opportunity.

Round table I

Moderator: Jesse Snedeker, Harvard University

Jesse Snedeker: Maybe one of the first questions I'll start with is: One of the things the folks that I was reading these articles with was to characterize the kind of approaches you all had to Fodor's argument. We really thought that there were 3 kinds of approaches that people said that they were doing, and in some cases we felt they were maybe better classified as one of the other approaches. What I want to do is just present this, and see how each of you respond as to whether you were correctly categorized, or how you would argue against that. In some cases where there are folks who haven't gotten a chance to speak, they can either say "I'll handle that in my talk and we can talk about it later" or they can choose to jump in if they like.

Basically, what struck us is that the first possible response you could have to Fodor's argument is that is fallacious in some way. Typically when people did that, they denied that learning had to be hypothesis testing, either by saying that hypothesis testing was explicit, and that the kind of learning they were proposing wasn't explicit, or by saying that somehow new things were generated, that weren't compositionally generated on the basis of a primitive. In most cases, we weren't convinced that they were actually doing that – Mandler's abstract, Dedre's abstract, Alison's abstract – to say that they were denying the argument itself.

But in all cases, when we read the articles, we saw what appeared to be primitives in the theories that they were proposing. Primitives, which did, in some ways, compose the concepts themselves. Now frequently, these compositions was not definitional, or follow say a prototype theory – but there were, for example, in Alison's theory, these variables, and there were the links between the variables, and when you looked at those, it completely defined the set of concepts you could possibly have, given that model. We actually reclassified those mentally, as the second response, which struck us as a very logical response. You could argue that just because decomposition has failed us in the past, it doesn't mean its going to be wrong. Maybe we've just been going about it wrong. Definitions? Sure, maybe definitions won't work, maybe prototypes won't work. But, we felt, that most of the participants in this symposium had something "new" – either "I have a new kind of primitive" (one that might be provided by perceptual meaning analysis), "I have a new kind of combinatorial apparatus" (I think best characterized Rogers and McClelland kind of contribution) or "I have a new procedure for recovering combination" (maybe like Bayes Nets or via analogy) – something that allows us to avoid exhaustive hypothesis search and focus in on just the right ones. I guess what I'd like to hear in response to this is – maybe we haven't been using the right tools.

Finally, another possibility is to just accept the argument, and try to account for developmental change and escape radical nativism. And we'll be talking about a little bit more in the second half.

Jerry Fodor: This why I wanted to get rid of the decomposition thing. I think it's really a red herring. You can really run the argument without it, for reasons I started out saying. Once you understand the possession conditions, that is, they are given in terms of what you can think about, the argument runs through equally well for complex concepts as basic ones. However, look, I've got primitive concepts *too.* It's just that my primitive concepts include brown and cow, and that's why I can compose with them to get the concept brown cow. What is the content of the concept brown cow – its exactly what you'd think – it's the set intersection of the brown things with the cow things – everything comes out exactly right. Look, God knew what he was doing what he made the lexicon – it could end up with exactly the same content, if only he had written in a longer book. God has very longs books that are available, so do neurons – we have billions of the damn things. The point is, you can't get the decomposition to anything much finer grained—intodecomposition into morphemes, right?

Jesse to That's not part of the logical structure of the argument. I mean there what you are really saying is that it has failed in the past, and with atoms it doesn't seem to be any better.

Jerry Fodor: Yeah, that's what I say about anti-gravity too. Nobody has an idea of how to make it work, least of all, its proponents. The proponents don't have an idea about how to make it work, because they haven't even understood the question.

Jesse Snedeker: So I guess I'd like to ask Jean and Alison, do you object to being categorized as (2), instead of (1)?

Alison Gopnik: The only caveat I would have about it – and I think that that is actually right. One of the things that happens is that you can take some of these primitives – I mean, the striking fact is that if you – I mean certainly I would think that some of the general graph structure – the relationship, the Causal Markov, intervention and faithfulness assumptions – I do think that those are primitives. And I think we start out with an initial set of variables that we are designed to identify, using some the same sort of procedures that Jean is talking about. I do think that, one of the things that happens – I mean, it's a *fact* that one of the things that can happen – as a result of new data, you can actually develop new variables. I think by the time we are at all sophisticated, certainly by the time we are in science, most of the content of our variables is no longer given by the kinds of things that Jean is talking about, but is in fact given by their conceptual and causal role.

Now in terms of Sourabh's point – you still have this argument with – Clark is my collaborator about this – who is someone in machine learning. And I would say, I can't stand these damn nativists, and he would say – if you are talking about anything other than the space of all logical concepts, and you think that you don't just learn just everything that you possibly logically could, then you have to be a nativist, right? Then the only choice is, do you have only a narrow set, when you are talking about triggering, or do you have a bigger set, and it's still less that the set of all the possible logical conclusions that you could ever draw. He argued it was like George Bernard Shaw arguing to the lady about all they were doing is haggling over the price, right? You would agree that you were a nativist, and you were just haggling about how much of a nativist you are. So in that sense, I think Sourabh is right – there is some sense in which you could say, if you can put any constraint on the kind of conclusions you are going to draw, then there is some sense in which you could conclude that all of those conclusions were dictated by what was there in the first place. But that seems to be a much weaker sense than what Jerry would want – about what it would mean to say that was innate. There is some set of restrictions and constraints – you can't get to any of the logically possible places, doesn't seem to me to count as being a nativist.

Sourabh Niyogi: The question is what the format of the primitives are – the task of the nativist is to answer: what are the constraints on possible theories? how do those generate lexicalizable concepts? You've identified what you regard some of the key constraints – in at least in Bayes Nets – Keil has a different set of proposals – say, 3 sustaining mechanisms that he thinks are important – and its our task (for cognitive science) to figure out what those constraints are. I don't see why *that* couldn't be a program that is analogous to what the generative linguists claim to do for syntax.

Jesse Snedeker: Jean, do you have anything you want to add?

Jean Mandler: Yes, I really fall between (1) and (2) I guess. I have an easier task, or a simpler task, than most of the other participants in this debate. Because all I need to do is to characterize the conceptual world of a 1 year old, or a 1 ½ year old. And that's fairly crude. And I think I know the way to get there, and I gave you a description of it. That's the first point.

The second point is that I don't think we have enough information yet but I think we will get more information about whether (2) is correct – that is, just because decomposition has failed, doesn't mean that it is wrong – coupled with new primitives. I suspect, although I realize that Jerry would disagree strongly with this, that it is possible to develop new primitives, from a mechanism of the kind that I described. I'm not sure of the answer to that – I may be wrong, but I think this is something that is modelable, and would be very interesting to do, and to try and do. Because I think it probably can be done. I think there is a lot more power in the spatial primitives than I think we give credit for. And I also agree with Frank, that our theories – or the way we put concepts together – is much sketchier and more primitive, and less developed than we tend to think, even for adults.

So, as long as you want your baby to up about 2 or 1 ½, I think you can get away with the nativist approach. And what I've called not a triggering approach, but a triggering plus descriptive approach. The question of whether or not we can go beyond that, I think we don't have enough data. But I think we have some idea of how to go about finding out.

Dedre Gentner: I'll be brief since its not really my turn yet, since I'm up there. I actually find decomposition really puzzling. So I used to spend my theoretical life thinking about for example how verbs might decompose into primitive, and tried to come up with evidence that they did and so on. But, I would now say that decomposition is an ongoing process. Its not at all what precedes concepts in general – and that in fact, the notion of "brown" is really hard to get. Dimensional notions, I think may be learned frequently by learning language, rather than the other way around.

And that brings me to another point I wanted to mention. I think in some cases, some concepts lead language, and in other cases, language invites why people might be using the word in that way, and for example, comparisons between the entities named by the word, and so on. So who is leading who, its just not a one way street.

Jerry Fodor: I am always puzzled by this. Why aren't you worried about why, if somebody doesn't have the concept of a tomato – sorry, *how* somebody doesn't have the concept of a tomato, he can learn the word "tomato". That seems to me like a straightforward paradox.

Dedre Gentner: Let me do "brother", which is easier – which seems easier to me now. Now I am going to say that the fact of being in a sibling relation is part of the concept – its not just metaphysics – its part of what you are supposed to know. But it isn't known at the start, by most kids. They freely use the word brother without understanding that what made him a brother is the sibling relation. But this is a messy little stand in, for what is going to be a real concept. But the fact that they are calling this kid "brother" helps them remember that that simple sort of link can be refined – and of course I'll suggest – the way you do it is that someone else gets called a brother, and the kid wonders why that's true. In fact, they sometimes object: you can't be a mommy, *that's* a mommy. But they have to deal with it – it's out there, its not going to go away. So they begin asking themselves, why, and then so on. That's the dumb way that can pave the way for the smart version. And I don't know that can work for tomato, but it works for relational concepts, which are almost never understood in their full relational structure.

Jerry Fodor: Nothing that has the form -- that he sees that that's a brother and that that's brother and proceeds from there in some fashion or another – nothing has the form can proceed unless he brings that concept to the data.

Dedre Gentner: That's a good point, Jerry, and I don't have a full answer. What I might have to say is two things – there's got to be some stock of primitives, and I don't know what they are. I think Jean's are a very interesting set. I think the set of things kids – we have to make up relational terms for kids – like "allgone", which seems to be made up in language after language. Those might be another good set. [Fodor interjection] Well, wait, there is a lot of implicit stuff – and the difference is whether it becomes explicit or not.

Jerry Fodor: Let me put one of my least favorite authors: "If you keep putting questions to Nature and Nature keeps saying so, you should take seriously the view that you believe something that's not true." People have been putting the question about that theory to Nature for 300 years. Nature has said no. Only Jackendoff and people like that think that it's said yes, and they only think it about a couple of verbs!

Look the view that you learn quotidian concepts by assembling them from more primitive concepts.

Jesse Snedeker: And now we're going to allow the audience to ask some questions – you'll get another chance to go at it in the second round table.

Audience: There is an issue that I see in the debate. That I would characterize the lead on a part of some of the members of the panel that rigor is formalizable on some of the level. Namely, there are deep assumptions that we can formalize at the concept level, a theory of concepts. On the other hand, there is some gist of the idea that its not possible given the nature of the world. I am wondering whether this rigor – those who say, given the nature of the world it is not possible – I suppose they are concerned with relevance as opposed to rigor. Now those who go after rigor, they go after it from a mathematical [perspective]. But those who are concerned with relevance or the reality of the world, go after weak theories or simply saying that concepts are not possible to formalize at the semantic level. And I'm wondering if you have any thoughts on that.

Jerry Fodor: Well, look if you can't have a theory of the mind that is given in terms of intensional like concepts, then you can't. And you might as well do something boring, like neurology. But we're all assuming, and I think, not unreasonably, since it's the only candidate that anyone's ever had, is that the appropriate vocabulary for couching a cognitive psychology, say, is a vocabulary of beliefs and concepts and desires and thoughts and so forth and so on. That could be wrong. I mean in which case you can't do the kind of psychology that …

Rochel Gelman: I don't expect an answer at this point. I do want very much to hear people talk about the problem of attention. There is a problem of selective attention: No matter what your theory as to why, a novice as a learner, attends to one variable or another, red when its food, but shape when its toy, etc. I've heard no discussion of this, and I think it does go back to the question of, what are you going to do, about "brown cow". Why do you attend to those two things, in the name of having a combination. And, I'm missing that altogether, and I think others in the audience are too. I don't want you to try to answer this now, I just want to put it on the floor.

Jerry Fodor: I give you a very short answer, but I know you won't like it – you attend to those things because you are born with those concepts. If you didn't have those concepts, you wouldn't attend to them – that's a truism.

Rochel Gelman: I know what your answer is, but I haven't heard the answer from other people!

Jesse Snedeker: My understanding is that the technicians need to eat, so we need to stop.

Timothy Rogers
University of Wisconsin – Madison

Jay McClelland
Carnegie Mellon University

This is one great thing about speaking after lunch. I don't think that Professor Fodor has made it back yet. So I get off easy! All the important stuff is going to be in the first slide.

This is work that I've done with Jay McClelland. And, although I'm going to be speaking here at the microphone, Jay is right here, so if I can't take the heat, he is available to answer the hard questions.

When I got Sourabh's invitation to speak in this symposium, I remember it as being fairly judiciously worded, to say something like: Some people think that some of your research might have some relevance to some of the issues that are going to be discussed in this symposium. And at the time I thought I might agree with that proposition. After this morning's talk, I think I feel the same way. (Please shout at me if I turn my head, and you can't hear what I'm saying.)

So, just to refresh your memory from this morning's talk, as articulated in Professor Fodor's précis, here's what he thinks we are trying to do. We're trying to understand the process of non-demonstrative inference, with respect to a body of data, specifically, specifying positive and negative instances of the extension of a concept. The explanation ought to refer to hypotheses about candidate identifications of the concept, and some confirmation metric about adjudicating the hypotheses according to the data.

Now we are coming to what I think is a similar aspect of human behavior, from a somewhat different set of initial questions. Specifically, we focus on what I think of as generalization, rather than a process of inference. That is to say, reasoning from premises to some conclusion. And, although our approach does have data, its not so much specifications of positive and negative instances of a concept, but rather experience tabulated over some initial ability to detect different elements of similarity and difference across different events – which I'll call properties, and I'll unpack this all, of course.

There are no explicit hypotheses about concept under our approach. Rather, we posit that there are knowledge structures, which lead people to implicitly predict unobserved properties from observed properties. So they derive expectations, which can either be met or not met. And rather than some confirmation metric, we posit a learning process, which will adjust these knowledge structures in response to incorrect predictions. (Now I'm going to unpack all that in 15 minutes, and I won't go a second over, I promise.)

So the cartoon of the problem we are trying to solve is the following. Here are 4 objects. I have a 3-month old infant at home. And if I presented these 4 objects to him, I would expect that he would be able to detect different elements of similarity and differences amongst them. For instance, these 3 items are all white, and I'd expect he'd to be able to see that. These 3 items share a similar overall shape, and I'd expect he'd be able to detect that. If he saw them moving, he might notice that these 3 things are all moving through the air, as well as this one moving on the water, etc. etc.

And, the mystery that we are trying to understand, is how it is through a series of experiences of different objects in different contexts in different situations in the world, we can come to know that a name like "bird" comes to apply to these 3 things, even through they don't have any of those sort of primitives in common, and exclude this one, despite its similarity in shape and color to other items.

The cartoon of our answer is something like this. My little son Elliot will, over the next couple of years, experience different kinds of objects, in particular situations. For instance, he might see a dog, and he might see a man calling to it, and on the basis of this experience, he will derive certain expectations about what will happen next. So for instance, most animals, when you shout at them, will run away. But dogs are different in this respect. This expectation will not be met, the dog will run toward the man instead of away, and from this discrepancy between what is expected and what is observed, we will expect he will learn to improve his predictions, specific to this particular item in the future.

So the simplest possible instantiation of this kind of idea, in a parallel distributed processing model, will look like this: So this is a model that consists of simple neuron-like units (represented by ovals here), and connected together by weighted connections. So, here, wherever there is an arrow, every unit on the left is connected to every unit on the right. Input comes in on the left, and prompts a flow of activation through the network, toward sets of units on the right hand side. And here, these sets of units represent individual properties, or different elements of similarity and difference that the model can, from its starting state, detect among different items. So for instance, it can detect elements of similarity amongst things can move, things that can fly, things that grow, etc. So the idea here is that there are 8 particular items that can appear in the world, and there are 4 particular situations in which can be encountered. And to represent that, we simply turn on whenever the thing labeled a canary appears, we turn this guy on. And, whenever its in a situation where we are paying attention to its behavior, we turn this guy on. And turning these units on, prompts a flow of activation forward through the network, and if the weights are configured correctly, it should at the outside be able to complete the proposition represented by these two inputs. So, in the network's world, the canary can fly and sing. (move should be up there as well but I think it's been covered up)

So the question is, how do we get the weights into the right state – I should add that 8 different sorts of things in 4 different contexts, and each conjunction of item and context leads to a different set of attributes being appropriate in the output. So, if we had "canary…has", the network should answer that it has wings, has eyes, feathers, and things like that. And, if we looked at the pine tree, and what it can do, the only thing the pine tree can do is grow. So after many sweeps through the corpus, with the learning procedure I'm about to tell you about, the network should be able to answer correctly any question that you'd pose to it, about any of the 8 items in its environment.

How do we get the weights into that state? Well, it's sort of a direct manipulation of this violation of the expectancy rule that I talked to you about just a moment ago. The idea is, initially, we set the weights to small random values, so that the network is effectively making the same null prediction about what is going to happen next, for all different items. We set these to small random values; you get a mush of activity that propagates forward. But then what we get to do is tell the network what actually does happen next. So it looks at the canary, says "can", it doesn't know anything. But then it sees that the canary can fly, or it sees the canary singing. And we use the discrepancy between what it predicts and what it observes to make small adjustments to the weights, all the way through the network, so that the next time it sees "canary…can", it's a little bit closer towards coming up with the correct conclusion.

Now why is any of this interesting? Well there are 3 points that I want to try to communicate to you in the time that I have left. The third one is the one that is most interesting and relevant for the purposes of today's topic. But to get there, I need to lead you through the other two.

So the first one is that the learned similarity relationships in this model provide a mechanism for inductive generalization -- I won't say inductive inference -- but for generalization of what you've learned in the past, an application to new things that you might encounter. This mechanism that I am going to tell you about comes to weight some features, some ways of thinking of things of being similar or different, more heavily than others, for purposes of inductive generalization.

Finally, the most interesting part: in certain cases, it can learn to generalize across sets of items that share no properties, in the training pattern, either in either the input side or the output side. And I will argue that, as a consequence our model exemplifies a certain kind of process that *can* acquire a concept – that is to say, a knowledge structure that permits generalization across some set of differentiable items -- which *doesn't* actually reduce to its initial feature set represented in the input or the output.

So the first point: how does it act as a mechanism of generalization? I've put this slide up just to remind me to tell you that the particular properties that we teach the network about were all derived from the influential Collins and Quillian model, illustrated here. So any predicate, or proposition that you recover from this tree, enters into the training corpus that we expose the model to.

We input different item context pairs, propagate activation forward, give it the right answer, and just adjust weights a little bit. And, what we pay attention to, and what I want to direct your attention to, is the patterns of activity that arise across this set of units. So whenever I turn on "canary" here, I get a certain pattern across here. When I turn on "pine tree", I get a pattern across the same set of units. And we can expect those visually, after the network has learned to correctly answer all of these questions. And I think you can see that the patterns are not especially random. There is a degree of similarity structure. There's the pine, rose, oak and daisy. If you look across the histograms, they have similar patterns of activity across those units, though not identical. And if you were to inspect this very closely, you see that the rose and daisy are somewhat more similar to one another, than they are to the pine and oak.

Rather than take my word for it, you can actually do a multidimensional scaling of the actual distances between these vectors, which is represented in the current plot. So this is a bit of a busy plot, but direct your attention toward the endpoints. The black labeled endpoints here. These represent the similarities in that representation layer, after the network has finished learning all the propositions. And what you can see is that the proximity on this plot actually of reflects the degree of semantic relatedness amongst the individual concepts.

So the robin and the canary are more similar to one another than the either is to the various fish. But the two fish are somewhat more similar to the birds, than either is to the various plants. Now its this recapitulation of the various similarity relationships in that hidden layer that provides a mechanism for generalization, which is what the shading on this plot is intended to illustrate.

Of what the model knows about these different objects, some things are specific to the canary and not true of the robin, like the property to sing. So the model knows if its going to active "to sing" in the output, it needs to be exactly on the canary representation. And if it gets a little far away, it should turn that guy off.

But other properties tend to be shared between by "robin" and the "canary", like having "wings" and the ability to "fly". As a consequence, the trained model knows that whenever its anywhere in the neighborhood of the canary and the robin, its licensed to activate "can…fly" in the output. Similarly, there are properties that are true of all the animals, and as long as the model finds its representations somewhere in this space, it ought to attest to those properties being true of the item.

In order for knowledge to generalize under this sort of framework, we need to specify a representation for a new bird as being somewhere in this cloud. And as long as it is not identical to the canary or identical to the robin, the network will attest to properties that are true of birds generally, but will not attest to properties that are true only of the canary or the robin. That was point 1.

Point 2 is that its not obvious from this plot that this generalization mechanism leads to a mechanism for understanding how different kinds of properties end up being differentially important for generalization and induction. The point can be made by paying attention to the lines in this plot, which illustrate the trajectories that these representations undergo over the course of learning. So initially, weights are small and random, so all the representations, no matter what you are looking at, are effectively similar – these red dots in the middle of the plot. But after just a little bit of learning, what we see is that the 4 different animal representations move away from the 4 different plant representations. And the interesting thing about this particular trajectory is that at this point in learning, the final similarities that are represented in the output, and that the model ultimately comes to master, are not perfectly reflected in these early representations. That is to say there is no differentiation of the canary and the robin here from the two fish, and there is no differentiation in the flowers from the trees over here. It is as though the model is preferentially attending to whatever differentiates the animals from the plants.

A little while later there is a non-linear change in the similarities. What you can see is that the birds and the fish split apart, the flowers split apart from the trees. But within these little groups, there is still no individuation of the two individual fish, or two individual birds. It is as through the model is now paying attention to the individual subcategories but not the individual items. And, finally, the individual items split apart.

What I want to give you insight into (in the next slide) is why this process happens, and its relevance for understanding how different features come to be differentially important for generalization.

So what happens in the model now – in its starting state, everything is represented as similar. If I learn that the canary can grow in a given episode, that information is going to tend to generalize to everything that the model knows about. Which is in the case of growing is appropriate, because everything *can* grow in this little environment, but for other properties it's not appropriate. If I learn that the canary can move, the next time I see a pine tree (which is represented similarly), I have to unlearn that – I have to learn that it can't move. And as a consequence, I'm going to have great difficulty learning anything that differs amongst this clump.

A little while later, when the animals are similar to one another – more similar to one another than to the plants – things change dramatically. Now I learn that a canary can move, it's going to generalize more to the other animals than to the various plants. And when I learn that the pine tree can't move, its going to generalize more to the various plants than to the animals. So given this slight differentiation, I now have some foothold for learning the properties that vary systematically with this difference.

The first point is that in the forward part of the network, the structure of the representations makes some properties easier to learn than others, and these are the ones that the network is going to be able to recover.

The second and slightly more subtle point is that these representations themselves are changing as a consequence of the learning process. So I can learn that the canary can move a little better, if I move the canary closer to the center of all the animals. So if all I've learned in the forward weights is that some things can move and some things can't, then I can capitalize on that forward knowledge by adjusting the representations so the animals form a tighter cluster.

In contrast, (here we go) moving pushes the representations apart. When it comes to these other properties, well I haven't learned to differentiate any of them in the forward part of the network, so they are exerting very little influence on how the representations change down here. In other words, I don't have the sort of notion that some things fly and that other ones that swim. The information that is coming back from these representations does little propel these representations apart. After the more superordinate information is mastered, tiny tiny changes begin to non-linearly build up until finally there is a cascade when the fish and the birds burst apart, at which point I can now master their properties. We view this little model as exemplifying a learning that results in a learning process that results in a progressive differentiation of concepts very similar to what Jean was talking about.

I need to explain how these ideas will lead to discovery of similarity that is not represented in the input or the output.

So here is a version of the same model that I was just showing you. What is highlighted in green is identical to the model that I just showed you – there are 8 concepts, 4 relationships and these activate overlapping sets of properties in this green output layer. What I have added into this model are 3 other sets of 8 items, with their own relationships, which activate their own sets of properties in the output. The important insight is that across these different colors, there is no overlap on either the input side or the output side. But, all of these things project through the same intermediating network structure. I've trained them on a set of patterns across these output units that exemplify the following similarity relationships: Each of the shapes down here at the bottom denotes one of the 4 sets. We can see that the 4 sets have nothing in common – that's why they are falling in different branches of the tree. But within each set, the 8 items enter into the same set of similarity relationships with one another. So in some sense this little guy here is to its family – enters into the same role – as this guy little here to his family.   Its not just a consequence of the clustering algorithm. You can look at the raw similarity matrices, and you can see that there is clustering structure and there is nothing between sets.

Now if we look at the representations that emerge in that representation layer, we see something very different. We see that the model is tending to group these things by their role within their respective families. Here items that share no properties in common but enter the same relationship to their families are represented as similar. It's not perfectly represented here with the more fine-grained structure, but its getting into the right ball park at least. And again, you can see that its not a consequence of the clustering algorithm, the raw similarity matrices show that the network has learned these guys are in some sense similar to these guys. And these similarities can be used to foster inductive generalization, for instance, if we were to learn that these things all have the same name.

So the take home points from this simulation.   Despite having no overlap in the training patterns, items that play similar roles within their respective families come to be represented as similar. I haven't shown you this, but the network can capitalize similarity for purposes of inductive generalization. And so we would contend, at least within the limits of what we understand concepts to be -- knowledge structures that foster inductive generalization -- the network has learned a structure that doesn't reduce to its constituent properties.

I don't have time to talk about this, but this stability is a consequence of innate starting parameters. It depends on the particular similarities that are detected in the input and output features, which you can think of as proto-concepts or initial ways of seeing things as similar and different. It depends on the particular network architecture. It depends on there being similar initial starting representations. And it depends on the learning mechanism, which adapts knowledge structures in response to discrepancies between expected and observed events.

Thanks for your attention.

Discussion on Rogers and McClelland

Sourabh Niyogi: Questions from the panel.

Jesse Snedeker: So when we were reading your paper, we saw a lot of parallels between the model that you are proposing and a prototype theory.   So like a prototype theory, you have a set of input features that are going to receive a weighting (right?), that are going to in some ways delimit the set of concepts that you can have. We also saw some ways in which it went beyond that. And maybe you could tell us how you think its different from a prototype theory.

Timothy Rogers: (Can I go back up, I have some additional slides? I really want to be able to point to the model.) One thing that makes our model different from prototype models, which I didn't talk about in the talk, but what's important for the models ability to attest to any property being true of any other item is the particular situation or context in which it is encountered. In this case, seeing a canary in the context of wanting to know what its behavior is, give you a different set of output properties than otherwise.

Jesse Snedeker: So its compositionality to the primitives basically.

Timothy Rogers: Right – so in prototype theory, a prototype is whatever it is independent of what you are doing with it. What is interesting in this sort of framework is that different situations and contexts entail knowledge about properties that engender different similarity relationships. So for instance, things that all behave similarly may in other sorts of contexts be very different from one another, and vice versa.

Think about functional features like telephones. Telephones can look very different, but in the context of using them, you use them to subserve the same end. So in that context, you need to represent them as similar. Whereas if you are paying attention of how they are going to fit into the décor of your living room, you need to pay attention to what they look like, and that's going to lead you to represent different sets of similarity relationships among the same items. In this case, you have a representation of the item that is suited – it doesn't get any inputs from the context – it is suited as it can be to all the contexts that you know about. It finds a structure that is as suitable as it can be across the different situations and contexts. But in this layer here, you can shade that representation so in different situations and contexts you can exaggerate or contract different kinds of similarity representations suited to the particular context that you are interacting with the thing in.

I would say that there are reasons why that's useful and interesting – that differentiate our theory from a prototype theory.

Sourabh Niyogi: I'm curious as to why you didn't take the step to add new items for bird, animal, and so on in the network. It seems to be one way to say, well, we are creating new nodes in our network that weren't there before.

Timothy Rogers: The answer to that question is that although in this particular network, we have gone a fair way using localist representations of individual items and attributes precisely because – its not the case that one grows new brain tissue every single time one learns a new concept. We conceive of the inputs and the outputs both being distributed patterns of activity across perceptual, motor and language areas, such that encountering a new item is not so much adding a new node as presenting a new distributed pattern of activity across a set of initial feature detectors. So in the very extended version of this argument presented in the book Jay and I published last summer, we have a version of this model where the attributes on the output are simply copied into the input. And instead of having a thing that stands for canary, you have the thing that stands for the yellow thing that shaped like a canary and has these wings sticking out. And now when you have a new bird, you say, the feathery thing that's brown that has these wings sticking out. So it's really a distribution over properties on the input. What's different between our model and a feature-based prototype model, in addition to the context sensitivity, is that it can learn conjunctively, how to respond to different combinations of particular patterns in the input.

Jean Mandler: I have one question about the context. You have just put those in – whatever is watching the scene – "can" and "has" and "is" – but those are also primitives?

Timothy Rogers: So this is a great question. There are two things that we think of as being contextual input. One could be the state of the rest of your input systems. For instance, when I am observing a dog, I am observing it through my eyes, and you're telling me "Hey look at the dog" I am hearing that through my ears and I've got some sort of conjunction of visual input and auditory linguistic input that I can then use to construct some Gestalt that I can unpack further information from. So context could be other elements of the scene, that you're representing in the input.

The second thing, that we go through in some detail in the last chapter of the book, is that there is a temporal context. For instance, in observing a squirrel running towards some sort of barrier, what I really have are successive snapshots of the relative position of the squirrel and the barrier. And I can use those, then, to anticipate what is going to happen next. So long as you allow for some mechanism for retaining information over time, that temporal context can then be used to constrain what information is coming to mind on the output side in a given situation.

Frank Keil: Independent of context shifts, one of the things I've always found intriguing about this work if not also frustrating, is that contrary to my naïve notion, similarity doesn't always build from the bottom up. You can have similarity relations that are very high level overriding lower level similarity relations as the structure differentiates downwards. So, you can have what looks to people like me as a thing highly abstract and causal saying no, look at this correlation, don't look at this correlation. That seems quite novel and different, and possesses an interesting challenge to those of us who think that was a hallmark of high-level causal schemata that was influencing. Am I right in that interpretation?

Timothy Rogers: Right – I didn't have the time to exactly unpack this point, but the notion is that according to that differentiation process that I described, the link that I didn't get to articulate is that the thing that drives presentations to push apart is the high order covariation amongst the sets of attributes that it can detect. That is to say, systems of properties that come and go together – on our view, they come and go together because there's causal structure in the worlds that leads them into these homogenous bundles. But the learning mechanism is sensitive to the fact that they come and go together, its that covariation that propels these representations apart. And as I think we did see, once the representations are propelled apart, it becomes very easy for the model to learn about those particular properties. In fact, the model will show this tendency to *mis*attribute some of those properties to items that don't in fact have them, so long as those items otherwise participate in the system of covariation. So for instance, you can see the model thinking that bats should have feathers instead of fur because in fact bat share other properties with birds that consistently come and go together.

Jesse Snedeker: Stephen, would you be willing to comment on what of the weaknesses of the prototype theory, for example, this kind of model might inherit.

Stephen Laurence: What I was interested in actually, is something along the lines of what Frank just asked about, higher level properties that might be associated with a concept– because many of the properties that are listed here are broadly perceptual properties. I'm wondering whether you want to build in (as innate) all sort of other properties, like functional properties, properties of being a mammal, …..– I don't know what you want to build in.

Timothy Rogers: This is a place where focus of my current thinking is, right on that question. We are beginning what I think of as the easiest abstractions to understand, from the point of view that we are trying to put forward. So for instance, in the last simulation that I showed, I was showing the discernment of similarity across sets of items that don't share any properties at all. So that's like a pure abstraction of some form – do you disagree with that?

Stephen Laurence: No, that's fine, there is a similarity – the question is whether there is a similarity that corresponds to function, or --

Timothy Rogers: So linking that to one of those similarities hasn't been done. But the kinds of abstractions where I think we can make a link are in things like, for instance, goal-directedness. This is a long story, and I don't want to bore people – but there's a story out there that goal directedness comes out of image schemas of the kind that Jean was talking about, that have to do with co-termination and self-initiated movement, contingent movement, that give you this concept of animacy – and goal-directedness is a part of that. Now I believe that observing episodes of different kinds of motion, some of which only occur with contact and others of which occur without contact, and the fact that little piece of information allows you to make strong predictions about the trajectories of the object. A billiard ball is perfectly predictable, but the frog that is hopping toward the barrier is much less predictable. Holding onto that piece of information becomes very important for how you represent the events subsequently. The accumulation of those influences over time can lead you to represent goal-directed agents as similar in certain respects in much the way that this model comes to represent items from the different families as similar in certain respects, despite not sharing any perceptual properties. The point is that it can extract these similarities, even when there is nothing shared, so long as there is other systematic structure that it can capitalize on.

Dedre Gentner
Northwestern University

Ok, taking as my theme one of Jerry's earlier challenges. I want to begin with the intuition about our great stock and trade as a species is that we are really good learners. I think the key to what makes us so smart is innate processes, not innate content. One of our special abilities as a species is analogical ability, and I mean that in a very broad sense which I'll describe in a moment. Another is of course, language, about which I will say very little, because there isn't very much time.

So here's a couple of very specific points where I am going to diverge from, in the same way as many of the other speakers have. I'm not going to buy hypothesis testing as the major way of learning, certainly not by kids. I think analogical learning is another rational and non-deterministic learning process. There's also non-trivial learning that isn't constructing a concept out of primitive concepts, although I think concepts do derive parts of their meaning by being related to other concepts. Furthermore, new concepts – or maybe I should use the word "notions" – can arise from comparing experientially acquired concepts (or notions), and once you get this new notion, it can then act as part of a definition, or at least part of connective structure that helps maintain the stability of other concepts.

I am going to say a few words about analogy and give a couple of examples of analogical learning, and come back to the issues of the symposium.

What's an analogy? It's a likeness of relational structure. And very importantly, the corresponding objects in the aligned relational structure don't really have to match -- an abstract analogy of the form that we all think is very clever – the relations will match but the objects won't.

But the same processes (that I'll describe in a moment) do apply when its an easy match and when the objects match – so a hen and a chick can match a mare and her colt – but a hen and a chick also match another hen and her chick. That's a really easy similarity, and importantly, that one can be understood by young kids, even by young kids who aren't too clear on the relational structure. There's some implicit constraints people use we care about 1-to-1 correspondence, we care about structural consistency, which also involves parallel connectivity – so the two relations correspond, then their arguments should also correspond in the same order. And systematicity (my apologies Jerry, its not used in the same way as you do, but I started using the word in 1981 or so and it was too late to change it when Jerry's systematicity came along) in the analogical means a preference for connected structure. You don't want to map a set of numerous but unconnected predicates to a domain – that would be just finding a set of coincidences. You want systems of relations that have some causal relation to them, or some other connective tissue.

I said earlier that I believe this process doesn't have to involve hypothesis testing. An important part of the claim here has to do with the way I believe it's done. You don't have to start by having an idea of what the analogy is going to show. Analogies—that is, let put it more broadly – structure mappings can start completely blind. For some strange reason, you start making a comparison – perhaps someone told you, perhaps two things have the same name, perhaps there was some stray surface juxtaposition that got you started. The first stage is – you have these two representations and you do a blind local match of the content – so you find any identities you can across the two situations. This is a highly structurally inconsistent – typically, you get 1-to-N mappings and so on. This is all done in parallel, its easy, its fast, its really stupid. Next stage you do a structural consistency enforcement— so at this point, all that mess breaks into little consistency clusters which we call kernels. And finally, in the next stage you form those kernels into the largest structurally consistent mapping that you can.

That last stage will involve a couple of interesting points. One is that we care about not just how many matches there but whether they are connected, so systematicity means that we are going to be looking for large connected relational structures. And two, once you get this maximal set, there's frequently there is something still present in, lets call it the base domain, that isn't present in the target domain –

but that is connected to the common structure, that we've mapped across as what we'd call a candidate inference. Two things come out that didn't have to be anticipated. One is the alignment, which actually forms a possible abstraction, and doesn't have to be known in advance – that is, you don't have to know how the analogy is going to turn out to be. And you also don't know what the inference is going to turn out to be.

So you can think of analogy as generating hypotheses, which then have to be tested. And certainly, many of the candidate inferences will turn out to be wrong. Not all analogies are right, clearly. But it's a very selective process. It's the sort of thing where by no means can any old thing be postulated and based on its presence in one of the items: you have to find common structure first.

One other important thing is – because literal similarity or ordinary similarity behaves the same way. Very young children can engage in this process without knowing the whole story, and that means that because relational structure is favored over isolated matches, that means relations may emerge when comparing two things that the infant or young child was previously at least not explicitly aware of.

Now I see some of this offers a solution to some of Fodor's conundra. Probably, I would have to say, more accurately, this process offers some way of approaching some of Fodor's conundra.

Let's go a little further, and I'll try to calibrate it better how this fits in – let's take an example or two. Here's an example of analogical inference. If I say, I give you analogy, and say "Walcorp divested itself of Best Tires bought a more profitable tire company" and likewise, "Martha divorced George, and …" lead you to make the inference, you'll probably infer that she acquired a more advantageous husband in some way or another (richer, more powerful, better-looking, …) – you won't infer that she bought a tire company, even though I told you that's what Walcorp did. In other words, you're going to look for something that has the same structural relation that Walcorp has to its tire company as the husband does to the new husband. That's point 1.

Analogy can also invite re-representation. Now what I mean by that is reconstruing predicates that don't quite match, to make a better match. For some reason, people like to find alignments. So what you see happening is, if you've got a match that's going pretty well but there's a pair of predicates that needs to be there but don't quite match, people will make efforts to reconstrue them so that they match better. So for example, if I give people that same analogy – and I say, tell me what's in common if I give people "Walcorp divested itself of Best Tires" and "Martha divorced George" – they are going to come up with something like "they each got rid of something they didn't want", or "they terminated a relation they no longer cared about". So in some sense that termination relation is now a little mini-abstraction brought out of the more specific statements, *divorce* and *divest*. So you could think of it, as on the diagram on the left, as a new abstraction.

You can also think of it in terms of a late decomposition, that you never really kind of noticed before – that divorce and divest have this notion of termination in common, that could be used in thinking about defining both of them.

Stepping back at this point, this is not at all evidence that the person didn't, in some *implicit* way, already understand the notion of termination and getting rid of. Of course an adult understands it an explicit way as well. So what I am claiming is that the idea becomes a little more conscious, a little portable, a little more explicit, and able to be more thought about.

But lets consider a couple of cases of kids, because that's makes perhaps a better arena to see stuff that's more like real learning. In this little example, I want to get across the idea that because comparison favors relational structure, it can lead to insight that really wasn't there before. In this kind of study, Laura Namy and I use a word extension task, you say to a kid, "See this? This is a dax" (in Martian language) while pointing to that tricycle. "Which one of these other two is also a dax?" And a four year old will typically point to the spectacles. As Linda Smith, Barbara Landau and others have demonstrated, common shape, common perceptual structures are important in this particular task, in this word extension task at least. Similarly, you get a new group of five year olds, and they too also point to the spectacles. And by the way, you ask them later, and they know that the glasses are glasses, and its not that we fooled them by giving them bad pictures or something. So overall you're getting mostly perceptual matches, 59% perceptual only 41% category match at this stage.

Now in comes the next group of 4 year olds, this time we give them both the bike and the trike. And say, "these are both daxes" – if you thought comparison was just a matter of concatenating perceptual features, then there should be now double the information. It tells you that the other dax has to be the glasses. But that's not what happens. The child at this point will choose the skateboard. And, my explanation for this is that if you get the bike and trike together, and they compare them, some information that was not exactly missing, but was certainly not explicitly available becomes more available – because a connected structure that has to do with riding things, and turning left on the wheels means the object turns left, and who knows what else – what other fragments of knowledge the kid is drawing on. But at that point that becomes more explicit and more compelling and the kid things thinks of these things more as vehicles and less as a pair of circles.

Let me give you one more example. A natural response here is, for someone who wants to be skeptical (as of course, we all should), is: look, there's no right answer, all you did was kind of hint to him which way he ought to go, or something like that. So I'm going to take a case where there's really a right answer. Kids weren't really getting it, and comparison helped him get it.

This is another very simple little task. This is a perceptual similarity triad. One of the reasons I like perceptual similarity for these purposes is that in a way, all the information, is, so to speak, on the table – there is nothing I know that the kid can't see with his own eyes. You give the kid – by the way this is done with Laura Kotovsky – you give the kid the top picture, and say, "tell me which one of these bottom ones is more similar" – and the rule is that there is always something that matches it relationally -- in this case its monotonic increase – and there is some other thing is nothing but a scrambled version of the relational match. So the non-relational match really has nothing going for it: its just a terrible match.

Now what happens is, the within dimension triads (like that top one), 4 year olds are really very good at – they're about nearly 70% correct. They can see the right one, they can say this is more similar – by the way no matter what the say we just __ them and just go on to the next one – no correction. But when we get to the bottom kind of triad, the cross dimension ones (you have to go from monotonic change in size, to monotonic change in shading) they are completely at chance. They are very frustrated – 4 year olds say things like "hey, they can't be the big one, because this one is red" – things like that – they reveal frustration with the task.

Now, here's what allows them to get the task. If instead of giving them mixed triads, as in the previous experiment I told you about, and put them in a new group, and given all the within-dimension ones – that's just 8 triads – and then all the cross-dimension ones, they continue to use to get the within dimension ones correct as they did before. But then they *also* get the cross-dimension ones correct – they have made a significant gain in their ability to do the cross-dimension ones. I think what is happening is what we call progressive alignment: they are able to align the structure on the easy ones, because the object sizes or shadings actually…

[MISSING]

Discussion on Gentner

[MISSING]

Stephen Laurence

University of Sheffield

University of Sheffield

This talk is based on joint work with Eric Margolis, who unfortunately couldn't make it. I'll organize the talk around Fodor's published discussions of concept nativism. Here is what I take to be a standard formulation of Fodor's argument for radical concept nativism [slide]. The argument goes, roughly, (1) learning requires hypothesis testing, (2) Hypothesis testing requires conceptual structure, (3) but lexical concepts aren't structured, so (4) they are innate.

We think that premise (1) is the one to question. We think that premise is false; Suggestive of this is the fact that empiricists traditionally are concerned to have not all concepts be innate. One way of having all concepts not be innate that doesn't involve hypothesis testing is to construct them in imagination. This isn't a learning model, but it is a way of acquiring new concepts doesn't involve hypothesis testing. Of course it may very well involve composing concepts, but that is a different matter.

Secondly, and related to this, is a point which a number of people have brought up — including Jean Mandler and Dedre and Tim – about empiricist models not necessarily involving hypothesis testing and confirmation. I agree with that. One way of doing this is to imagine that you form a new representation when you notice that a number of features are correlated reliably in the environment. A mechanism that simply forms a new representation that is the conjunction of those representations of those features, gets you a new representation that isn't acquired by hypothesis formation, which automatically conjoins these things.

Of course, that conjunction again requires *structure*, and we get a reformulation of Fodor's argument [slide: (1) Apart from miracles or futuristic super-science, all concepts are either constructed from primitives or innate. (2) if they are constructed from primitives, they're structured. (3) lexical concepts aren't structured. (4) so lexical concepts aren't constructed from primitives. (5) therefore, lexical concepts are innate.]. In this formulation, the point is that – the problem for empiricists is that all models that empiricists come up with involve concepts having structure – and the models I just gave are no exception to that. So you have a revised formulation that gives Fodor the same conclusions.

Now there are two questions to focus on here. One is the question about lexical concepts – concepts like DOG, COW, BROWN, and so forth. How are these acquired, and are they learned or are they innate? The second question is about the primitive concepts. Of course if lexical concepts are themselves constructed from simpler concepts, then these are completely different questions. Many people, as Jesse pointed out in the previous discussion period, take Fodor to task there, questioning whether lexical concepts can't be constructed from simpler ones. I take the other question, about the primitive concepts, to be the more interesting one, basically because, if we can acquire new primitive concepts, that means we can expand the combinatorial expressive power of our conceptual system. I take it that this is one of the most interesting questions in this area. Our claim is going to be that the new premise (1) in Fodor's argument which I've repeated here – that all concepts are either primitive or innate – is also false – that's going to be the main claim of my talk. So I want to claim that it *is* possible to acquire new primitive concepts, expanding the combinatorial expressive power of the conceptual system.

Needless to say, not everyone agrees with this. Fodor clearly doesn't. Here are some quotes saying there is no such thing as expanding the combinatorial expressive power of a conceptual system. [slide] Here's some other people who don't endorse Fodor's conclusion, also endorsing Fodor's argument in that respect – that primitive concepts are in fact innate. They are people who *don't* endorse Fodor's conclusion, so they are going to look to build lots of lexical concepts out of smaller primitives, so you don't get the conclusion that all the *lexical* concepts are innate.

Here's the strategy – let's suppose along with Fodor that concept acquisition centrally involves acquiring some representational structure with a particular content or meaning – so, for example, to acquire the concept DOG you have to at least acquire a representational structure which refers to all and only dogs. If DOG is a primitive concept, then the theory of content involved needs to be the one that applies to primitives. And, of course, the content of DOG won't be determined compositionally, because DOG by hypothesis is a primitive concept, it doesn't have any parts, and so, you can't get its meaning from the parts.

The theory of content that I want to use — just as an illustration — which is useful for rhetorical purposes, is Fodor's own theory of content, and it's the one he endorsed in 1990 and some version of it in 1998 (I don't know about more recently). This is a cartoon version of the theory [slide]: the idea is that content is determined by a mind-world relation, not determined by compositional relations amongst things in the head – what it means for a representation to represent dogs is for there to be a nomic relation between dogs and your dog representation so that DOG is tokened as a consequence of your causal contact with dogs. It may also be tokened due to causal contact with other things – you see a cat run in front of your car in a dark night, and maybe that causes you to think there is a dog there; so Fodor wants to claim that these tokenings are in fact dependent on those, but not vice versa. That's the theory. The details won't be hugely important, because as I say this is only meant to be an illustration for how it could work with various theories of content.

The key notion that we need to introduce here is the notion of a sustaining mechanism, which Frank Keil alluded to. And the idea is that when you have nomic causal theory like Fodor's theory, the sustaining mechanism is whatever mechanism it is in virtue of which the concept stands in that mind-world relation. And, as I note here, this can be generalized to other theories of content so that broader version of a sustaining mechanism is whatever mechanisms need to be instantiated to realize the content-determining properties, that your theory of content says makes representations have the content that they have. Any theory of content that would apply to primitives (or lexical concepts directly) would be relevant to this puzzle.

I'm going give you just a sample type of sustaining mechanism, introduced by Eric Margolis, which we appeal to in our paper together. The idea is that you have got a *kind syndrome* – which is a representation associated with a concept that picks out a collection of properties that are indicative of the kind, but are accessible in perceptual encounters. And then together with that, you have an essentialist disposition which gets you to treat things as instances of the concept just in case you take them to have the same essential properties as the paradigmatic instances picked out by the kind syndrome. The idea is basically that this realizes the theory of content because you get – the kind syndrome gets you in the neighborhood of the property, and the essentialist disposition deals with fakes – things that have the same sorts of perceptual properties as the kind syndrome picks out, but in fact are not instances of that category – things like cats in the dark night.

I think there are a couple of important points that need to be made about this, because in cognitive science, atomistic theories of content like Fodor's are not widely used, and so they are less familiar. One question is, why is DOG still primitive on this account? There are a couple of points to make here. First, the sustaining mechanism that controls the tokening of the concept, that is, gets you to the token as a causal consequence of seeing dogs, doesn't determine the content directly by means, say, of a compositional rule – it is not as though we have other representations which compose by a compositional rule to yield that content. Rather, what they do – the sustaining mechanisms -- sets up the mind-world relation – and that is what gives the content.

Perhaps the best way of seeing this is to note that there are different types of sustaining mechanisms, that will work even for a given theory of content like asymmetric dependence. Frank Keil mentioned several in his talk that would all work for a single theory of content like that. In principle, different people could have different sustaining mechanisms, setting up the same mind-world relation for the concept DOG. They would then all have the same content, and so the same concept. That shows you that the sustaining mechanism is only contingently related to the concept.

Here's some quotes from Jerry Fodor which basically illustrate this point for his theory of content. Here he is saying (as he said in comments earlier) that there are lots of different ways that you can be connected up – your concept can be connected up to the thing in the world that it refers to, and here what matters is that mind-world relation obtains, not the route by which it obtains.

Nonetheless, its worth noting that what does set up the mind-world relation is something cognitive. You're using a theory, you are using representations of properties that instances of the concept have, these are all representational and intentional.

Next point, is well, why is this concept learning? First of all, it seems pretty clear that, its not a triggering model, its not as though you have the concept DOG all wired up and ready to go, you just see a couple of instances of DOG and that gets it going. Rather, what you are doing is tracking properties that dogs have. So you are appealing to the fact that dogs have certain properties that are picked out by the kind syndrome, and you are relying on the essentialist disposition to get you into the right mind-world connection.

But those are all properties that you are contingently picking up. As you can see here, for example, imagine that you are acquiring a new concept, like the concept of a kangaroo, you are going to pick up on the contingent properties that you notice that are associated with kangaroos, that will get you the initial linked kangaroos and the essentialist disposition will connect you to the kind more directly.

To my ears, the account clearly sounds like a learning based account: its sensitive to environmental contingencies in a way that is very sensible given the concept that you want to acquire. On the other hand, I'm perfectly happy to have lots of elements of the sustaining mechanism be innate. As Frank Keil alluded to, I am happy to have concepts like ARTIFACT or NATURAL KIND, even more specific concepts like ANIMAL, and help to structure your learning in setting up the sustaining mechanism. That doesn't mean the concept *itself* is already there. Rather you have (I completely agree with what he was saying), you have sort of a general mechanism which allows you to set up a large range of concepts – an *open-ended* range of concepts, within a certain category, like natural kinds or artifacts.

Maybe it's worth, just before ending, to compare this with Fodor's own response, in his most recent work – I don't know if he still endorses this – but in the most recent discussion that I know of this, he adopts an account of concept acquisition which just basically says that concept acquisition simply isn't rational or cognitive at all. And here is a couple of quotes for you: [slide] In his discussion of this, he spends a lot of time discussing what he calls the doorknob-DOORKNOB problem (this is a problem that Fodor himself introduces). This is basically the problem of why we acquire the concept like DOORKNOB as a result of interacting with doorknobs, as opposed to zebras or something like that. This is a problem for Fodor, of course, if the process is non-rational, not cognitive, then it seems a mystery why you should acquire a given concept by relating to its instances. Fodor's own solution to this is a metaphysical solution. He basically says well the reason why you acquire the concept DOORKNOB by interacting with doorknobs is because the property of being a doorknob that you are referring to is partly constituted by the concept that you are acquiring as it were. It is no surprise then that DOORKNOBs are acquired on the basis of interacting with doorknobs.

Here is our solution, by contrast. It's a very cognitive solution. The reason why you acquire the concept DOG by interacting with dogs is that dogs are the kind of things that exhibit the dog syndrome. You are acquiring information about the dog syndrome, and that's what gets you to lock onto dogs. That's a very different solution than Fodor's metaphysical solution.

I'll end by noting that what I'm claiming is that Fodor's argument fails, I think (this is a bold claim), and it fails by the falsity of that premise (1) that I pointed to. That you can in fact acquire a new primitive concept. But of course, this leaves lots and lots and lots of work for a theorist interested in concept acquisition, the details which need to be filled out by psychologists, who are doing the hard work in the field.

I'll end with that.

Discussion on Laurence

Jerry Fodor: Look, in a way, I think what is going on is hardly worth arguing about. Here's a thought experiment. Suppose that we were so constructed that we can acquire the concept DOG in the following way: we learn, let's say, hydrodynamic theory, and we learn it very well, and come to have all those beliefs and so on, and we learn it by all the usual inductive processes, like telling you something, whatever. And then the concept DOG pops into our head. Would that count as *concept learning* in the sense of concept learning that we are trying to reconstruct? Well, obviously not. Learning takes place, but the learning that takes place isn't relevant to the identity of the concept. What that shows is that real honest-to-god concept learning that we all intuit the notion – the concept learned would have to be semantically related in some interesting way to the body of information that occasions the learning.

Now, is there any way of requiring this, on the kind of picture that's been presented? Well, no. In fact, there's no reason why the concept – why the theory that you learn – the learning that locks you onto the concept shouldn't be radically false and totally crazy. So for example, your theory of stars (a favorite example of mine) is that they are holes in the curtain of heaven – that's what a lot of Greeks thought – but it would still be a theory of stars. It will still also be connected to stars by that totally crazy false theory – but that's not the kind of thing we have in mind. The kind of thing we have in mind in concept acquisition is that if concepts are defined as a theory, then the experiences that lead to your forming them have roughly speaking the character of evidence for that theory – prima facie evidence for that theory. That's not guaranteed by this – so the fact that the object that sustains the mind-world relation happens to be a theory is no more interesting than if it were sustained by an expert or something, it is sustained not by a set of beliefs but by another intentional system. That's all very interesting – a theory of sustaining mechanisms, if there could be such a thing, would be nice to have, but *this* kind of theory of sustaining mechanism simply doesn't *bear* on the question of whether concept acquisition is a learning process.

By the way, I should say, I'm still prepared to take quite seriously, what I think is more and more plausible as time goes by, that you can't characterize the mechanisms that sustain your reference relation essentially to dogs or cats or any of the other quotidian concepts at (as one says) the intensional level – in terms of connections between concepts and the world, simply because there are so many kinds of sustaining mechanisms, when you chunk things up that way. You might be able to characterize the sustaining mechanisms at the level of – you have to at least a better chance of being able to characterize them – at the level of (as it were) *brain*-world correlations, but of course, that's not a psychological theory.

Stephen Laurence: Ok, well there's a lot there. First of all, my feeling is that the hydrodynamic stuff is a total red herring. The issue is – ok, let's take the Greeks and stars – the Greeks had a radically false theory of stars, let's suppose. But, if that's enough to get them the concept STAR, that is, if asymmetric dependence theory is in the right ballpark, if having this kind of a relation to stars is enough to get you the concept STAR, then I don't *care* whether its false. Seems to me you've acquired the concept STAR by getting into that relation.

Jerry Fodor: You could acquire it by having your head be re-wired. The problem is --

Stephen Laurence: Yes, but the relation between having a false and relevant theory of stars and acquiring a concept in that way, and getting your head kicked, is a totally different kind of relation. First of all, it's clear that the account we are giving is intentional. The sustaining mechanisms, the kind syndrome, and the essentialist disposition, are all at the intentional level. The information that you collect is relevant to the concept. It is not totally irrelevant. You can imagine a case where you have totally irrelevant information – about hydrodynamic theory or whatever – and that sets up an asymmetric dependence relation. To me, there is an open choice here: Asymmetric dependence theory is false – that may well be, in which case, we have to look for another theory of content, that's, I think, a very real possibility. Asymmetric dependence theory (like any other theory of content) has all kinds of problems. And any theory of acquisition that is completely based around a theory of content is going to inherit the problems of that theory of content. If however, you think that it is enough to stand in those kinds of relations to have the concept – then I think that a theory of acquisition which is intentional and lets you set up a sustaining mechanism of that kind is enough to let you acquire new primitives.

Jerry Fodor: This misses an option, and it's a crucial option. If the sustaining mechanism (if you believe the theory of reference) is set up, then you indeed in fact have the concept. It doesn't follow that you learned it. And it doesn't follow that you learned it, even if some or all of the mechanism by which the mind-world relation was set up is an intensional mechanism. Roughly speaking, what you need is not just that it be intensional, but that be *evidential.* I have no idea how to cash that out, and neither does anybody else. I think it's a suspicion that you probably *can't* cash that out, that you are probably not going to have an intensional theory of concept learning.

Stephen Laurence: Well you can define learning if you like in such a way that its…

Jerry Fodor: Oh, come on …

Stephen Laurence: Well, let me finish – it's true it doesn't *follow* that necessarily the only way to instantiate this is by a learning process – you could imagine somebody going into your head, hooking neurons up and that'll do it -- maybe concepts could be acquired that way. Neither does it follow that it's non-intentional. All that you need -- in order to undermine your argument for radical concept nativism -- is a case where you *can* acquire new primitives, by a process that is plausibly learning (not that the only way to acquire a primitive is by learning). Now the reason why we think that this is a *learning* process is that the agent is sensitive to environmental contingencies in a reasonable way. Now this is why I say you can *define* learning in such a way that the only thing that counts as learning is hypothesis testing if you want to. But that's just a *verbal* quibble. What we want to say is that this is learning because what the agent is doing is sensitive to her environment in a reasonable way. The way that we suggest – who knows whether it will work, but that's the theory.

Jerry Fodor: What is reasonable is left open. Learning is not just a matter of being sensitive to the environment. It's a matter of adopting a hypothesis, *on* the basis of information that evidential for the hypothesis in question. That's what everyone has *always* meant.

Stephen Laurence: That's not what *we* meant. Look, we're looking for a way out of a puzzle. When you're doing that, you may have to question one of the things that everybody has always thought, right? So here is a case where what it is to have a concept is to stand in a certain mind-world relation. It isn't a matter of representing to yourself certain information. Now, if having a concept isn't a matter of representing certain information to yourself, then acquiring a concept isn't going to be a matter of representing certain information to yourself that's necessarily associated with the concept. So you need something else. So what we're claiming is that on the model that we suggest the information that you collect about dogs to get the dog concept is syndrome based information that's perfectly relevant and reasonably related to acquiring the DOG concept.

Jerry Fodor: It's that rational relationship that's carrying the weight of the notion of "learning". Now what you need to do, to make this run, is to show how learning a theory on the basis of evidence could give you a concept. You are not going to be able to show that, and the reason you are not going to be able to show that is the one I pointed out sometime back, the notion of a concept is *prior* to the notion of a theory, not posterior.

Stephen Laurence: Ok, I'll give you the word "*learning"*. Fine, this isn't a theory of concept "learning". But neither are these concepts innate. Neither are they acquired in anything like in the irrational way that they are acquired, say, by being kicked in the head. So obviously there is something *else* here. And if you don't want to call it concept learning, that's fine….

Jesse Snedeker: I'd like to butt in and throw in some of questions that we collected in preparing for this. And this one has obviously come up – about what the nature is, of sustaining mechanisms, and *possible* sustaining mechanisms. One thing that we were wondering about is, Prof. Fodor, you opened up the door, in *Concepts,* to a variation of sustaining mechanisms, which might limit you to not having to consider all concepts that are atomic to be innate.

Jesse Snedeker: In particular, in chapter 7 – right -- you suggest --

Jerry Fodor: You actually read that thing? Nobody else does.

Jesse Snedeker: (Yes, I read every single page. No, I can tell they have) In chapter 7 you suggest actually that the sustaining relationship for natural kind concepts in latter day people (people in the last 100 years adults, people who are well-educated) might actually be a fairly complex one – a theory-based one, in fact. One thing I was wondering was the decision to limit this more complex sustaining mechanism to the sort of peripheral case didn't necessarily seem well-motivated by your own theory, right? So as long as the concepts were available that composed the theory that created the sustaining mechanism –as long as you were able to chop the world into chunks on the basis of some prior concepts this was available to you. Since many of the panelists today actually believe that children have very scientific theorizing available to them, would you consider this as a possible sustaining mechanism in other domains, into a wider range of humans, and into a wider range of situations.

Jerry Fodor: No, for reasons previously suggested. And, obviously, this needs arguing. The reason I wouldn't consider it an option is that if you want the learning to be based on a rational relation, in some sense or other, between the experiences and what's learned, which I take it is really part of the notion of learning – if you want that, then I think you are going to find that any theory rich enough to sustain the locking relation for a concept like GOLD or TIGER or something essentially uses the concept gold or tiger. So for example, it uses -- like chemistry –the fact that gold has a certain position in the periodic table and so on. This is another way of making the point I keep coming back to. Namely, that the priority is from concepts to theories, not from theories to concepts. When you try to go in the opposite direction, you get theories that don't give you learning, because you already endorse the concepts that you are trying to learn.

Stephen Laurence: If I could just jump in. One thing that is interesting is if you compare the solution that Fodor offers in the *Concepts* book with the type of solution that Margolis and I are offering, you see that Fodor claims that what you need is a *metaphysical-based* solution, to the doorknob-DOORKNOB problem. And that the theory is essentially at the level of neurology, which gives you the account of how concepts are acquired. What Margolis and I claim is that much of the story can be told at the *cognitive* level. And I think that's the essential difference.

Jerry Fodor: That's because you are not giving us an account of what the rational relation is. You're saying it's got to be an appropriate relation. So my case learning ORANGE from a theory of hydrodynamic theory doesn't count. But if you *were* to give an account of what the constraints are it would seem that you are not going to -- If you were to give an account of what makes the constraints rational, then --

Audience: Are you talking to us or are you talking to him?

Jerry Fodor: Oh, I was talking to him [Laurence], actually. (chuckles) But if you do give an account of the sustaining mechanisms that requires rational relations between the experience that leads to having the concept and the concept itself, then the sustaining mechanisms will really be *theory-learning*, and what I predict (though this needs a lot of argument) is the theory will *have* to be already rich enough to contain the concept.

Dedre Gentner: Let me just dive in on the old natural kind point, because I think it actually bears out a little of what I've been trying to say, that Jerry wants to say that you have to have the concepts before you have the theories. Another argument says you have to have the theories first, and then the concepts. It's seems to me messier than either of those.

First of all, having a concept, if that's to mean having a rational view, that's to mean being able to explicitly think about the concept, its going to leave out the majority of cognitive content. And like most psychologists, I'm really interested in not only what we are able to think about (whatever we choose to), but also in the stuff that happens to us, cognitively to that we don't manipulate out of choice but rather because we recognize and respond to it, it may take a long time to actually be able to articulate it and do something with it. It seems to me that natural kinds are a really nice example, because they are not at all natural, in general. That is gold is confused with fool's gold and all sorts of things, and figuring out the periodic table was, a huge effort of a lot of smart people, and a lot of blind alleys. What happens is you have some kind of klutzy thing, that you can recognize in the world, and explore its relation with other things, and do compare with other things --

Jerry Fodor: Well you're the guys who think that concept gold is given in terms of the periodic table. *I'm* not.

Dedre Gentner: Actually, I'm more on your side on this one. Arriving at something like the periodic table – or the belief that the periodic table (whatever it is) is probably right – its not something we start with –its an iterative process that changes the concepts as well as the relationships between the concepts.

Jerry Fodor: I don't see any reason to think it changes the concept of GOLD. Nor do I think it is remotely plausible that no one was able to think about GOLD – that's my test for concept possession – until they had the periodic table.

James McClelland: There is one feature of connectionist networks with hidden layers in them that I would like to place on the table in the context of this discussion about whether you have to *logically* have the concepts before you learn something about them *or not*. This is the notion that – if you think of a 3-layer connectionist network – you have your input units, you have your hidden layer, you have your output layer, so in some sense if the network is going to learn some mapping from input to output, somehow or other the hidden layer is going to have to "have the concepts" that then map to the appropriate outputs, like, you know, you have to have DOG to know that BARK is the thing that is going to happen. So you *think* you have got this situation where you've got logical priority, where the input has to know how to code the input into the concept layer for you to then learn how to map the concept onto the output. But its an *illusion*, it's *not* a *fact*, about how the brain works – that there has to be this notion of logical priority. The reason is this. A connectionist network provides the opportunity to learn *any* of the conceivable mappings, from patterns on the input to patterns on the output, so long as there's a few hidden units in there that randomly provide a initial exploratory basis from which to begin to sort out what the possible appropriate internal representations are, that can then be used to map onto the concept.

So what I take issue with fundamentally in the debate here – is the notion of logical priority of the concept, *before* you can learn anything about it.

Jerry Fodor: You may be right, but this is a prime example of connectionism viewed as pragmatism with a computer. The view of concepts I am taking – and if you don't like it, you have to justify not taking it, I suppose – is *not* they are mapping inputs to outputs. That's the very kind of pragmatist view I want to get away from. The view of concepts I am taking is to have a concept is to be able to think about the kind of concept that it is a concept of. And it is *because* you can think about what it is the concept of, that you can perform the various input-output manipulations.

James McClelland: Just to follow up, my belief is, that the things that get learned in the hidden layers *become* the things that you can then think about.

Jerry Fodor: Well there is nothing in your theory which sustains that belief. What the theory rests on, is so far anyway, is the ability to learn the input-output mappings.

Jesse Snedeker: We're going to take questions from the audience now.

Roundtable II
Led by Jesse Snedeker, Harvard University

Rochel Gelman: I guess I have the same question I asked before the break. Anybody who takes the view that you might have something like DOG will then go searching for more information about dogginess -- be it a sustaining mechanism or the kind of theory Jay is discussing – has to deal with the following question: what in the environment will be focused on? In other words, what is the psychological theory of the environment, such that the mind will cut up the environment at the relevant joints, to be picked up by a sustaining mechanism. Or, anything like that, that isn't carrying in advance a whole lot of information. I'm not taking a position on this, but it seems to me that this is being avoided.

Jerry Fodor: It's a fact (among others) that this question is unanswerable unless you assume an innate repertoire of concepts, that drives people to nativism. That's the perfectly straightforward answer to this. What determines which things you attend to is which concepts you innately have!

Rochel Gelman: Jerry, I know your answer.

Jerry Fodor: But it's a good answer!

Rochel Gelman: What I don't have. What I do care about is hearing alternatives. It cannot be the following form: Since it's a cat, you pay attention to its whiskers. That presupposes catness --

Jerry Fodor: Think about Goodman's grue-green problem. The answer to the question why is it that it is the greenness of emeralds that you attend to and not the grueness of emeralds – is: the concept GREEN is accessible to you, presumably innately, and the concept GRUE isn't. That's the only possible answer to that.

Stephen Laurence: One alternative is that you have a lot of innate concepts guiding you that are not identical to CAT or DOG. So for example you could have ANIMAL, you could have NATURAL KIND, or OBJECT, and so forth, and those could all guide you toward relevancies.

Rochel Gelman: Another way of putting it is how do I know which sustaining mechanisms gets hooked with which innate concepts.

Paul Smolensky: This is not a question for Jerry Fodor. This is an observation of someone trying to make sense out of this discussion, from a cognitive science point of view, not a philosophy point of view specifically. It seems to me there is the following argument, I'm not going to attribute to any particular person or the responses. Here's the argument: You think you have a learning system, fine there is the learning system, what I want to do is imagine all the possible ways that this learning system could end up after doing what you call learning. There is a space of possible outcomes call it X, and there is a system that leads the system to some point in X, which we will call learning. What I am going to do is I am going to call X a hypothesis space, and I am going to call the fact that this particular hypothesis was the result of your so called learning process, and that it survived the testing procedure, and I am going to say that everything else was rejected by the testing procedure. So I am going to say that whatever you have, you have a hypothesis testing system. And you might think that what the system ends up with is something new. But of course, that' can't be true, because from the very beginning, there was this space X.

Jerry Fodor: Well that's just wrong.

Paul Smolensky: This is not a question for you, Jerry.

Jerry Fodor: Well I am going to answer it anyway, because its very tempting, and thoroughly wrong. The reason why --

Paul Smolensky: Jerry, I am not finished.

Jerry Fodor: But it's only going to get worse.

Paul Smolensky: Well you'd better get used to that, because it is going to get worse. So you're going to think (mistakenly) that your learning process is produce something new, because that space was out there to begin with, there isn't anything new to find, it was all there to begin with. Now this is argument A. Now this is presented, or something like it, or something mistakable for it.

And a bunch of empirical scientists respond -- they don't take the argument to be what I just said, because it seems to me unimaginable that anyone could object to that argument. It is a tautology. They hear something else. It could be that the argument is stated in such a way so as to invite disagreement, it could be of some Gricean maxim that says well, the argument can't be that obvious, there must be something that I can consider meaningful in this argument. What is that? Well, what are the empirical questions? What is the space X? What could possibly be the shape of this that could do justice to human cognition? What could possibly be the right process that would lead us to the right place? Those are interesting, important questions. Those are the questions that most cognitive scientists want to answer.

What response could you have? You could say, those empirical investigations are beside the point because they don't address Argument A. Or you could say, Argument A is beside the point.

Jerry Fodor: That's false. I really want to answer that because it's a trivializing argument that I really find quite insulting.

Paul Smolensky: [audio failure]

Jerry Fodor: Let me finish – I stopped you, you do too. The reason you call it hypothesis formation and testing is not that it provides a function from data states to terminal states. The reason is that anyone who has thought about what the function must be *like* can see that it goes via something that has *logical form*. That is, you can't say what the confirmation conditions are for P, unless you know more about it than that its P! You have to know what its structure is. Just like you can't say what it entails, unless you know what its logical form is.

That being so, what you've got to show is that these intermediate states wouldn't be hypotheses, even if they had logical form. Now I'll tell you how connectionists deal with that: they ignore it! They have no notion of logical form, because they have no notion of constituency. So of course, if you ignore the notion of logical form, you don't see any difference between a real and trivial hypothesis formation model.

Alison Gopnik: Just a response to that, which I think is relevant to some of the other responses – think about the situation with vision. So I think vision is a very good case where you can turn out to have a vision system that can have certain kinds of assumptions and certain kinds of procedures, you can say well look, the fact that you can see all this incredibly wide variety of things means all the wide variety of things you can you must be innately determined to see. And in some sense that's going turn out to be true, but what's interesting is you could have had a sort of notion that you have all the possible shapes in your head to begin with, and what happens when you see is that various kinds of retinal patterns trigger concepts of those particular kinds of shapes. It just turns out that's not the way that it works. The way that it works is that there's these much more abstract principles that are designed to actually take particular structural patterns of evidence and draw particular kinds of conclusions from them.

I don't think that would have been obvious had you been sitting a priori and trying to give an account of vision. In fact I think this was just as much an argument of the empiricists in the 17th century as arguments about concepts were – you wouldn't necessarily come up with that kind of answer to the problem. I think that's the kind of answer to problem that's going to turn out to be the right answer for conceptual development as well. And its just the fact that we haven't had that kind of answer that's kept the problem in place.

Frank Keil: There's lots of ways to go from sets of environments to smaller sets of mental representations. Chomsky talked about how we acquire grammar, and it didn't involve hypothesis testing, so I think we have to be very careful about what you set up as hypothesis testing vs any other kind of mapping from sets of environments to smaller sets of mental representations.

Andy Clark: I think Jerry Fodor is getting away with something by slipping it in very early into the argument. And what comes in very early is this rejection of pragmatism. The whole argument is just going to go through flawlessly and beautifully, if you agree that there is no essential connection between grasping a concept and being able to do things in the world. There's all sorts of way to act in the world and learning to do things in the world, which are completely immune to the worries that Jerry Fodor is raising.

So that the only way that this argument goes through is if you buy something that is very counterintuitive, that is to say, that there is no essential relation between the concepts that you grasp, and the things that you can actually do in the world. And that's the price that Jerry has to pay for the very lovely argument.

Jerry Fodor: There's no price for me, I'm a Cartesian!

Look, you got to keep all the balls in the air. Keeping one of them in the air doesn't count. One of the balls is: whatever you choose as concept constitutive has to compose. *Skills* don't compose. So they can't be concept constitutive, and pragmatism has got to be false!

Hmm? Sorry?

Jesse Snedeker: We have to leave now.

Sourabh Niyogi: Thanks, everybody, for coming – I'd like to thank all the participants for coming. We'll hope to have interesting discussions afterwards. Thanks very much!