CLARK GLYMOUR

ANDROID EPISTEMOLOGY FOR BABIES: RELECTIONS ON *WORDS, THOUGHTS AND THEORIES*

ABSTRACT. *Words, Thoughts and Theories* argues that infants and children discover the physical and psychological features of the world by a process akin to scientific inquiry, more or less as conceived by philosophers of science in the 1960s (the theory theory). This essay discusses some of the philosophical background to an alternative, more popular, "modular" or "maturational" account of development, dismisses an array of philosophical objections to the theory theory, suggests that the theory theory offers an undeveloped project for artificial intelligence, and, relying on recent psychological work on causation, offers suggestions about how principles of causal inference may provide a developmental solution to the "frame problem".

1.

At its birth in 19th century neuropsychology, the most successful strategy of cognitive psychology was decomposition. Apparently indivisible intelligent capacities were shown to consist of a complex of less intelligent component subcapacities. When parts of our machinery are broken – when our brains are damaged – we behave irrationally, incompetently, and our failings reveal something of the brain's mechanisms. The psychologists of the day allowed that when whole we were still the grand, rational creatures we had taken ourselves to be since the Enlightenment. Freud, who began his professional career as a neuropsychologist, extended the strategy to psychological breakage. But he and his disciples gave a post-Enlightenment twist to abnormal behavior and rationality: as demand for their services increased early in this century, Freudians claimed mental breakage is almost universal, and found further proof in the subsequent decline in demand for their services.

By mid-century, a certain pessimistic parallelism emerged in social and cognitive psychology. Through a series of slightly shocking experiments, social psychologists argued that features of character we think are stable are really artifacts of context. Change the context sufficiently and the kind become vile, the brave servile. At about the same time, Paul Meehl (perhaps not accidentally a psychoanalyst) argued that simple algorithms make

Synthese **122:** 53–68, 2000.

better predictions than do expert clinical psychologists. Meehl and his contemporaries in social psychology anticipated a genre that has now virtually taken over psychology. Cognitive or ethical behavior is compared with some normative standard, and humans are found wanting. Well designed machines would optimize; we are machines that can only satisfice, on a good day. According to received opinion in cognitive psychology we are ill-constructed, incompetent machines, without firm character, unable to act by moral or rational standards, deluded that our conscious deliberations cause any of our actions. The one bit of intelligence left us – science – is an unstable oddity we sustain only through elaborate social mechanisms. The philosophers agree.

We might have guessed most of this from the newspapers. Still, fundamentalists and post-modernists aside, we are smarter than toasters and thermostats. We are a lot smarter than any machine we have been able to build. Most extraordinarily, each of us learns a natural language, and each of us learns everyday physics, spatio-temporal regularities, commonsense psychology, and a wealth of causal relations involving people and things. No machine as yet comes close. The intelligent things about us are not what we do or what we know but that we learned to do or to know. On reflection, the common complaint that Turing set too high a standard for machine intelligence has got it backwards: for intelligence like ours, a computer should not only be able to hold a conversation that imitates a man's, or imitates a man imitating a woman, it should be able to learn to hold such a conversation, in any natural language, from the data available to any child in the environment of that language. Turing thought as much himself.[1] For machines we can build, that would be a dream, if only machines we can build could dream. If we're so dumb, how come we're so smart?

2.

My six year old daughter, Madelyn Rose, has a frog named James. James and Madelyn have rather different worlds. Judged by his behavior, James' world is pretty well described by a Carnapian idealized language with two monadic predicates: brownspot-in-water, fast-large-motion-nearby. James probably has some sort of sex predicate too, but we haven't seen much sign of it. When James was a tadpole his world may have been simpler, but it can't have been a lot simpler. Madelyn's amazing world is filled with things with various powers, all of which she knows about and knows how to use, people, with mental states she matches or contrasts with her own, complex relationships of indescribably many kinds, and a language she can speak

and read and sort of write and tell bad jokes in.[2] She has explanations for her world, pretty good explanations even when off the mark.[3] Six and a half years ago she knew none of this. How did she come to be such a know it all?

That seems a good question. What we seem to know from developmental psychology is this: Madelyn was born able to discriminate up-close objects, with the ability to judge whether there were one or several such objects, and with a disposition to reidentify objects that moved continuously in her field of view. She also identified the objects of one sense with the objects of another – the same object was seen and touched. By the age at which she could control her head a bit, she could reidentify objects that had not moved when she had turned her head so that they were out of her field of vision, and then turned it back. By six months she could reidentify objects by predicting a trajectory when they had been out of her sight for part of that trajectory, as long as the total trajectory was very simple, e.g., a straight line. She made lots of mistakes – in particular she thought things that disappear tend to be where they were last seen, even in contexts where that was repeatedly falsified. At about nine months she began to think that people in different positions see different aspects of an object, the details of which she was still working out at 18 months. Using constancy or near constancy of perceptual features, by 12 months she could reidentify objects that had been out of sight for a while, and she was no longer stuck with the mistake of thinking things remained where last seen, although she could still be fooled. By 18 months she was reidentifying objects from perception more or less like an adult, but her understanding of what others perceive was still not correct. By age 3 she had got others' perceptions of object – at least what is visible and what is invisible to whom – right.

Madelyn was born knowing how to imitate some facial expressions. Within a couple of months she had learned that certain of her actions, in certain contexts, produced a result, and that in some cases the result varied with the intensity of the action (as in kicking). She tended for a long while to radically overgeneralize and undergeneralize connections between her actions and consequences. If pulling a blanket with a toy on it brought the toy to her, she would pull the blanket even when the toy was beside, not on, the blanket.

In this same period Madelyn learned to crawl and to walk, and began to learn to talk. According to psychologists, the timing of these skills was not accidental. Crawling improves judgements about reidentification (or "object permanence"), and judgements about objects that are out of sight develop at about the time that a general word for absence ("gone") enters speech.

Madelyn's psychological knowledge went through a similar series of stages. For a while she did not recognize that others' beliefs, or her own, could be false. Her judgements of what was believed were a subset of her judgements of what was true. Eventually she came round to our distinctions.

Now at six, and even before her, knowledge of folk psychology and folk physics and spoken English is essentially complete. She still has some odd false beliefs (she thinks she speaks Spanish because she spent her first year in Costa Rica), but then don't we all?

3.

Developmental psychology has been mostly an account of stages. At certain ages infants do this then that, later the other thing. As with butterflies from caterpillars, going through stages, even amazing stages, even stages that lead to the right answer, may make a thing or person interesting, but not smart. Compare a developmental version of Kevin Kelly's (1997) Einstein machine: the first hundred data points of the right sort you put in, it responds with $E = mc$; the next hundred $E = mc^3$; after that $E = mc^2$. It does nothing else. In this world, the Einstein machine converges to the right answer; in any world in which the energy equation is different in any way, the Einstein machine gets the wrong answer or no answer at all. By increasing or slowing the rate at which relevant data are input, you can change how soon the Einstein machine converges to Einstein's equation; by stopping the right data input before 201 relevant data points are submitted, you can stunt its growth. But that's about all you can do. Nothing could be more different than the Einstein machine and Einstein, at least the popular Einstein: the popular Einstein would have found the truth whatever it might have been (as long as it was beautiful and simple, etc.) The popular Einstein was smart; the Einstein machine is stupid. But from another viewpoint, the two, Einstein and the Einstein machine, differ only in degree, only in the range of different possible circumstances in which they find differing truths.

A lot of psychologists think kids – and therefore all of us – are Einstein machines. We will, given normal stimulation, develop the right cognitive skills and beliefs, no matter what else; speeding up the stimulation may speed up the development timing, and the reverse for slowing the stimulation. Abnormal stimulation in place of normal stimulation just stops development. Put in a world where objects can pass through other objects – or appear to – where people have visual perception out of their line of sight, where objects really vanish when out of perception and don't

reappear, where an unhuman language is spoken, children couldn't adapt their beliefs and skills accordingly. What goes on in development is like data decompression triggered by outside events, just as Plato claimed. Sometimes this is called the modular view of development, which doesn't seem very descriptive.

The modular view of development can be traced to Plato, but there are more modern philosophical sources as well. I have in mind some of the writing of Rudolf Carnap, Bertrand Russell, Clarence Irving Lewis and Nelson Uoodman. Carnap and Russell and Lewis had similar philosophical educations, first in the conventional turn of the century neo-Kantianism, second in mathematical logic. Russell hinted at a combination of the two in *Our Knowledge of the External World.* The world delivers to us the matter of sensation (in Kant's terms) or sense data (in Russell's terms) or qualia (in Lewis' terms). We supply the apparatus of logic and an elaborate scheme of definitions, which, applied to the particulars of sense data, define (literally) objects, processes, space, time, relations of all kinds. The world we experience just *is* logical combinations of sense data. Russell doesn't work out the details. C.I. Lewis gave a very similar story in *Mind and the World Order*, again without the details. Carnap was a detail guy. *Der Logische Aufbau der Welt* assumes that what is given in sensation is a gestalt, an entire experience at a moment, not particulate sense data that have to be assembled into a gestalt. What is given in reflection, according to Carnap, is the recollection that two gestalts are in some respect similar. With these primitives, Carnap offered explicit logical schemes to represent sensory modalities, objects, space and time; what's more, he realized (in 1928!) that he was writing a program, and in parallel with the definitions he offered "fictional procedures" to construct an instantiation of whatever entity he was defining. Carnap's effort was revived in the 1940s by Goodman in *The Structure of Appearance*, which explored various logical methods of definition and constructions from different primitive bases. Carnap's hints about procedures were not followed up.

Several things strike me as interesting about this bit of philosophical history, now regarded by most philosophers who know of it as so much logical weirdness. First, it was equivocally substantive psychological theory; Russell and Lewis claimed to be giving an account of how the mind works, rather in the spirit of George Boole. Carnap, who actually did some ingenious mathematical work, typically muddled issues by claiming he was giving a "reconstruction" and a "logical justification" of something, although of just what is unclear. Unlike Boole, Carnap never wrote the plain and obvious thing, that his theory aimed to be an idealized, and therefore approximate, account of how we think. Although his work was

arguably the most ambitious mathematical psychology of the time, psychologists then took (and now take) no notice of it, although the impulse of logical formalization seems to have influenced Clark Hull. Second, none of this work has a reliabilist cast. The view of Russell, Lewis, Carnap, and Goodman is not that there is a world out there of things and properties and processes and minds and relationships, veridical representations of which we are constituted to construct. And there is certainly nothing in this viewpoint about learning, nothing at all.

It requires only a twist of perspective to see these philosophical efforts as attempts to describe a modular mind, a system of Einstein machines, of the kind many contemporary cognitive psychologists think we are. And contemporary philosophy finds the modular view of development remarkably congenial. According to Jerry Fodor, for all but the highest order processing, modules are the end state of development, and these views seem to be shared by a number of philosophers. Modularism fits with English neo-Kantianism. For example, a book length essay by John Campbell (1994), offers a more or less a priori account – but sprinkled with references to the psychological literature – of relations of precedence about conceptions of space and time and personal identity, all notions of importance in developmental psychology.[4]

Artificial intelligence is at least equally friendly to the modular viewpoint, at least partly because it is difficult enough to give a computational account of relatively encapsulated skills. Naive physics, for example, is an interesting descendant of Russell, Lewis, Carnap, Goodman efforts, and it bears on the Einstein machine view of development. Patrick Hayes' idea was to formalize (preferably in a computationally tractable way) the principles of everyday common sense adult knowledge of the identity and behavior of middle sized dry and wet goods, solids and liquids. That must include principles about containment, occlusion, disappearance and reappearance, co-movement of parts or regions, identity through time and through changes of properties, causal interactions that influence shape and motion, and so on, all topics investigated in developmental psychology. So far as I know, Hayes and others working on naive physics have paid no attention to developmental psychology (the inattention is mutual) but his project, if brought to fruition, would imply a procedural characterization of adult (and six year old) competencies. (The line of descent is this: Carnap was Herbert Simon's teacher, and Simon, along with others, founded artificial intelligence, of whose American Association Hayes is a Past President. Besides, look to the logic.)

4.

The views Gopnik and Meltzoff's book present are, so far as I know, the only development in developmental psychology that offers an Enlightenment picture of human capacities.[5] They say that children are more like the popular Einstein than they are like Einstein machines. What Gopnik and Meltzoff think Madelyn Rose did as she grew from zero to six was this: she did science. She formed theories, made observations, conducted experiments, formed empirical generalizations, revised her theories, altered her "conceptual scheme", explained things, collected or ignored anomalies. If she had lived in a world with a different everyday physics (say, for example, she grew up without gravity, the Virginia Dare of space stations), she would have developed a different, but correct, theory of the physics of everyday things. If she had grown up in 'toon land, where even the concrete can talk and buckle and have eyes bug out, she would have had a different theory of kinds, atooned to her environment. Children are scientists, in fact the ideal scientists imagined in old fashioned philosophy of science, with a desire for truth and control unbiased by competition, without need for tenure, without deference to other scientists, with an abundance of data available, with endless leisure. Their inquiry may be unconscious, or only partly conscious, but so is the thinking of individual adult scientists.

Here is a Rousseauian theory of cognitive development that rides on philosophy of science, more or less as philosophers in the fifties and sixties understood science, a theory that offers a radically rational view of each of us at our beginning. Man is born brilliant but is everywhere stupid. If ordinary adults have a huge irrational streak, committed to absurd Gods, alien abductions, and creationism, it is because, unlike children, they deal with issues for which there is a paucity of evidence, or because their native rationality is corrupted by social forces.

There is a nice historical circle to the theory theory. Gopnik and Meltzoff have obviously been influenced by both Kuhn and Piaget. In the preface to *The Structure of Scientific Revolutions*, Kuhn cites Piaget (and *The Child's Conception of Causality*, in particular) as one of his sources of inspiration, and some of his most famous terminology is derived from Piaget.[6] Now the theory theorists want to take Kuhn's philosophy of science back to development, even if only the (real) aged Kuhn, shorn of the theses that made him famous and closer to Enlightenment conceptions of rationality. What do contemporary philosophers make of all this? Not much.

Although it has been elegantly and plausibly developed by Gopnik and Meltzoff, for philosophy the timing was wrong; the theory theory invokes

the kind of philosophy of science already going out of fashion twenty years ago, and there is no error in philosophy more profound than to be out of fashion. Much of contemporary philosophy and philosophy of science has all but abandoned epistemology focused on individuals, and takes adult irrationalities to be norms, or more accurately, takes all norms to be social.

Hegel is said to have given a proof that there are necessarily only six planets. Our contemporaries are no less adept at a priori arguments that the theory theory must be false, and many philosophers who have not given arguments have staked out positions that entail them.[7] One hostile commentator has repeatedly announced that there is a Natural Ontological Attitude, which inclines people erroneously to believe what they say about middle sized dry goods; he finds it unseemly that the NOA is anything but natural, or that we might be right about such things. Another commentator holds that all real mental processes are conscious, theorizing is a mental process, so, since much of the hypothetical theorizing postulated by the theory theory is unconscious, it can't exist at all. An English noncommentator would find theory theory psychologists to be so many social climbers: for any concept X, from a priori reflection Oxford professors will announce necessary and sufficient conditions for a creature to have the concept of X, and psychologists will confine themselves to showing that the philosophical analyses are instantiated in people. Several commentators are quite certain that there is nothing at all for children, or anyone else, to be right about, hence nothing to be reliably right about. Wittgensteinians know the matter is beneath discussion: a theory is linguistic; an infant's theory would necessarily be private: hence if the theory theory were correct there would be a necessarily private language, which is impossible. Really tedious commentators have some take on the notoriously vague notion of "theory", of which they claim Gopnik and Meltzoff's account is in violation. Perhaps the most offended are those philosophers of science who have claimed in recent years that the study of science should take account of cognitive psychology; they have, one and all, gone over to the dark side. They are postmodern. Their motives were to dismiss all that mean and difficult stuff about evidence and inference and put in its place vagaries and banalities about information processing. The theory theory is their worst nightmare.

5.

Philosophers throughout this century have believed that even with social complexities aside, the process of inquiry could not be algorithmic, or as they put it, there is no logic of discovery. As machine learning has

advanced in the last decades, and automated methods have seeped into many sciences, like Hegel's proofs these philosophical cavils have become increasingly quaint. Theory theorists, steeped in the computational conception of mind, suggest that infants and children embody algorithms for inquiry, but they give no hints about the content of learning algorithms, or how they can reliably succeed. While the data on stages of development may not determine a unique algorithm of inquiry, there is enough data to constrain algorithms, and to make a computational theory of development an interesting project. The project seems to me right at the logical center of the most ambitious aspect of artificial intelligence, android epistemology. The center is nearly empty; to the best of my knowledge, there is no serious movement in artificial intelligence to fill it, and no funding for such work. That is too bad.[8]

There are more bad philosophical objections to the theory theory – and to the kind of philosophy of science it invokes – that are best considered in the context of baby android epistemology. According to the philosophers, theory is radically underdetermined by evidence; according to the real Einstein (who claimed to be a determinist!) scientific theory is a "free creation". How come, then, aimost all babies converge on Madelyn's view of things, processes and people? The project of baby android epistemology helps make sense of the theory theorists' reliabilism. If baby scientific theorizing isn't a free creation, but is the application of algorithms that (as theory theorists suggest) start with an initial theory and have rules for elaborating, retracting or revising theory in the light of data, and for acquiring new data, and for attending to some of the data while neglecting other parts, and meta-rules for revising rules, and all babies share relevantly similar data, then convergence is what one would expect. If the data are sufficiently overwhelming with respect to the theoretical options available to the baby, then the algorithms need not even be deterministic or entirely invariant from individual to individual.

Reliable convergence is one thing, reliable convergence to the truth another. According to the philosophers (from Plato to Popper and after) there can't be an algorithm that uses singular data only and that has the following properties: In all worlds in which a universally quantified proposition is true, the algorithm converges to asserting the proposition, and in all worlds in which it is false, the algorithm converges to asserting its denial. The claim is in developmental psychology's philosophical source, *The Meno*; the proof is in Sextus Empiricus. To the two thousand year old argument, contemporary philosophers of science have added only anecdote: even the best confirmed and accepted scientific theories often turn out to be false; witness Newton's. But the proof, and the relevance of the

anecdotes, depend on an unnecessarily stringent criterion of convergence to the truth. The philosophers require that the algorithm be equivalent to a procedure that, after receiving some finite array of evidence, gives a single conjecture, and in every possible world that conjecture is correct.

There are two dimensions of alternative success criteria. The algorithm need not succeed in all possible worlds, but only in a large and interesting set of possible worlds (the theory theorists do not assert that babies would learn the essentials in every consistent world in which they survived; they claim there is an ill characterized range of worlds in which babies would do so). And the algorithm need not be equivalent to one that gives the truth and only the truth in each of these possible worlds; we might only require for example, that in each possible world there comes a time after which the algorithm ceases making erroneous conjectures and ever after conjectures the truth, which is equivalent to a standard Bayesian criterion of reliable convergence to the truth, and is the criterion used in mathematical studies of language learning, or we might require any of a hierarchy of still weaker criteria. Weakening the success criteria in either dimension strengthens the logical content of learnable hypotheses (for details and references, see Kelly 1997).

There is more. Theory theorists claim that babies undergo internal conceptual revolutions; whole groups of theoretical notions dominant at one stage of development are abandoned and replaced by others at later stages of development. At every stage including the last, the categorizations that evolve seem to have an element of artifice; they are conceptual schemes about the mental and physical into which particular events are fitted and shoved. In C. I. Lewis' terminology, the babies evolve different "pragmatic a priori", conceptions; in Carnap's, they evolve different languages; in Kuhn's different paradigms. (So easy to find novelty by not reading long dead white men.) For the philosophical tradition, "conceptual revolution" carries a Kuhnian burden of which the theory theorists take no notice. Conceptual changes alter the meanings of sentences; truth is part of meaning; so conceptual changes alter the truth values of sentences. Right or wrong (I think wrong) the philosophers' picture of conceptual change, more or less explicit from Lewis to Putnam, is that truth is fixed by the world and by the conceptual scheme together. Surely, there can't be any notion of an algorithm reliably converging to the truth if the very output of the algorithm changes what is true.

Yes there can. Actually several interesting notions. The learning algorithm can eventually converge to a single conceptual scheme within which it converges to the truth; or the learning algorithm can vacillate among conceptual schemes, within each of which it converges to the truth.

There is a well worked out abstract theory of relativistic convergence to the truth, and characterizations of algorithms that do so, all of course unread by those who sought in relativism escape from the oppression of reliability. Even Kuhn's own radical relativism, in which the beliefs of the community determine the truth, admits a reliability analysis.[9]

6.

Another of Pat Hayes' innovations is the frame problem, which Hayes and John McCarthy formulated as a technical problem about logical descriptions of the consequences of changes in logically described world states – Carnapian state descriptions Hayes has adamantly resisted other formulations of the frame problem, but they have proved irresistible, and to much of the artificial intelligence community the problem has become how an agent can feasibly isolate the features of the world at a time that will determine the consequences of an action if performed, so that planning and prediction are possible. The problem is that there is an infinity of features, many of which change as an action is performed, or vary from action to action of the same kind, and an endless variety of possible circumstances in which a customary regularity between action and consequence does not hold; generating and testing the relevance of properties, or the possible violations of ceteris paribus would paralyze an android, no less a baby android. Learning about the consequences of action s is learning about causation, and learning about causation eventually has implications for knowing the consequences of actions. Transformed rather far from Hayes and McCarthy's original formulation, but rather close to how it is nowadays often understood, the frame problem is about how an android can feasibly acquire and use causal knowledge.

Madelyn solved the frame problem, somehow. According to the developmental psychologists, babies come equipped with a set of natural kinds, how large a set one doesn't know. Presumably they must also come equipped with mechanisms – procedures – for guessing causal connections. (Bertrand Russell's last, and most underrated, philosophical work, *Human Knowledge, Its Scope and Limits*, offered some guesses at the prior causal knowledge and inference principles we must start with in order to learn the causal structure of the world.) Somehow the two, kinds and causal inference procedures, are combined to connect action and consequence. We would expect developmental and cognitive psychologists to tell us how. They haven't.

The developmental literature – and the psychological literature more generally – has not been kind to questions about learning causal relations.

Piaget gives accounts of children's causal beliefs, but says almost nothing about how they are arrived at. Pavlov and Skinner avoided talk of learning causes in favor of learning associations, although the salient difference between classical and operant conditioning is that the former teaches associations while the latter teaches causal connections. The neural network model, which is hidden beneath a lot of twentieth century psychology, from Freud to Thorndike and after, promoted the study of associations. Most recent psychological models of causal inference are derived from a neural network model (the Rescorla–Wagner model), and explicitly confound learning associations with learning causes. One aspect of the frame problem is exactly how an android can sort out the difference: every variable feature of the world is associated with an enormous number of other features, but most of these associations are not causal, and actions that alter one feature will not alter the other. When she was hungry or uncomfortable, Madelyn the infant kicked and cried, but only the crying brought Mom or Dad and food or comfort. When she kicked and cried at night, lights came on before Mom or Dad arrived, but the lights did not cause Mom or Dad to come. (Even so, any number of cognitive psychologists just look puzzled when asked about how causal relations are learned; what does causation have to do with learning, they ask.)

No surprise, in experiments supposedly about causal inference, cognitive psychologists find that adults fail to live up to associationist norms, conclusions offered as further demonstration that humans are subrational. (Better evidence about psychologists, I think.) In refreshing contrast, Patricia Cheng (1997) and her collaborators and students have developed quite sensible models of causal relations, and have found that adult judgements are in qualitative accord with them. In related unpublished work, Lien and Cheng have shown that adults use causal and statistical information in forming (or attending to) new, more abstract kinds and use that information in judging causation in new contexts.

Cheng's models turn out to be a particular parameterization of a class of causal models widely used in artificial intelligence. The models go by various names – directed graphical models, Bayes nets, belief net– but what they share is a representation. For a first formal approximation, causal relations between features look like a two-place, asymmetric, antireflexive, transitive relations. Such relations determine directed acyclic graphs. In graphical causal models features or properties or variables are represented by vertices; the (direct) influence of one feature on another is represented by a directed edge. In artificial intelligence work, as (implicitly) in Cheng's models, the directed graph is associated with a joint probability distribution over all possible vectors of values of the features or

variables. The pairing of directed graphs and probability distributions may be subject to various restrictions, the most common of which is called the Markov condition, essentially a generalization of Han's Reichenbach's notion of "screening off". The Markov condition requires that the topological structure of the directed graph is reflected in the conditional independence relations that hold in the probability distribution. In combination with information about the time order of events, or with stronger assumptions connecting the probability distribution and the graph, the Markov condition permits algorithmic inferences about the structure of causal relations from observations of the associations among features.

One of the striking things about the Markov condition is that, outside of the perversities of philosophers, it seems almost irresistible. Experimental design in the social and biological sciences depends on it; analyses of instrument design flaws in physics presuppose it; social science and biological models (without feedback) assume it; Bayesian statisticians who would not truck with graphical models implicitly apply it. One of the recent, much cited, experiments that purports to show that human causal judgement is sub-normative instead appears to show – fairly dramatically – that adults use conditional independence information and the Markov condition in causal inference.[10] If babies are small scientists, they "use" the Markov condition. That does not suffice to show the principle is innate. In a "closed world" in which there are no unobserved common causes of observed features, and in which the inquirer could directly manipulate any feature, the Markov condition could be learned empirically. It does not seem to me very plausible that the baby's world is like that.

How might a baby android solve the frame problem? Starting with a comparatively small array of features, it notes associations either produced by its actions or otherwise, and the time order of associated events. From that information it infers that some associated features are not causally connected, or are connected by a sequence of steps involving other features, or are more or less directly causally connected. It forms new, more abstract features of causal significance, and applies the same strategy to them. In predicting the outcome of an action (whether its own or others'), or a sequence of events following some salient event, the android baby need only attend to a very restricted collection of possible causes or effects. As the baby android learns about new features of the world, it applies the same procedures to them, one by one. Statistical sampling issues aside, the procedure is reliable only so long as a form of "closed world" assumption holds, namely that the associations the baby android observes are not produced by unobserved or unnoticed common causes. The closed world assumptions the baby android implicitly makes are often false, but

their falsity is eventually revealed, for example by finding new properties conditional on which previously associated features are independent, or by intervening to manipulate one feature and finding that another feature, associated in observation, is unchanged.

Clearly, this can't be all there is to children's discovery of the causal structure of the world. Somehow, Madelyn put together a more or less coherent, intensely structured, body of causal knowledge. I haven't a clue as to how.

7.

How could it be that children are Enlightenment rational agents, while adults are slaves of irrationality? Perhaps that is the wrong question; what the theory theory claims is that children conduct inquiry, and they do so by methods that are reliable over a range of possible circumstances. It doesn't follow that children's methods are those that methodological moralists claim are "rational". One of the startling lessons of computational learning theory' is that reliable inquiry has little to do with philosophers' dogmas about rationality. Very well, how could it be that children conduct inquiry in a reliable way while, for many purposes, adults do not? Many of the supposed irrationalities of adults are patterns of inference that are irrational by certain norms but in many circumstances do not interfere with forming reliable judgements. For example, experiments with adults showing that in judging probabilities they tend to ignore base rates. But when there are huge amounts of data, ignoring base rates usually does not prevent giving the truth a large probability. And a preference for testing a hypothesis by looking for confirming instances, supposedly an irrational bias, is in many circumstances an efficient strategy (see Klayman and Ha 1987). Other supposed irrationalities depend on controversial construals of the agent's representation of the problem or the goals, for example, Peter Turney (private communication) has pointed out that even in the Wason Selection Task, if subjects understand "if D then 3" as Prob(3 | D) > Prob(3), then the modal behavior on the task is rational. And, quite seriously, the very young may be aided in inquiry by the fact that they are not part of a community in competition. Philip Kitcher has recently illustrated with simple economic models how competition among scientists with different resources may facilitate scientific progress. But the reverse is also true. For example, from what I read, the psychological literature on causal judgement is, on average, a mess of bad experiments, badly interpreted, relieved now and then by real effects, often soon overwhelmed by more bad experiments that fail to reproduce the effect. Adult science

may require a modicum of democracy, but too much scientific democracy drowns good science in the noise of incompetence. Baby scientists, at least, don't have that difficulty, which may be part of why they succeed so well.

8.

The differences between modular theories and the theory theory are not always clear, and are always matters of degree, and I certainly don't know the true degrees. But a big part of scientific progress is showing a path of exploration that, if true, would yield rich results, and the theory theory does so. At the very least it kindles a project that ought to have been near the center of artificial intelligence. Still less do I know if human babies solve the frame problem as I have suggested. I do know that the philosophical objections to the theory theory are largely drivel.

There is a wealth of experiments that might be done to give some indication, one way or the other, as to whether the theory theory is correct, and by what procedures infants and young children make causal inferences. The decisive experiments that would give some evidence for the theory theory as against modular views of cognitive development are unlikely to be done until we have to raise children in a very different physical environment from the Earth – they are possible, but probably impermissible, in simulated environments. Experiments that investigate by what principles infants and children form a causal understanding of the world are more feasible, but very few experiments have been done to investigate adult use of conditional independence and other frequency constraints in constructing causal explanations, and babies are a lot tougher to work with than adults. Heritability studies are possible. I can at least hope that psychologists influenced by Gopnik and Meltzoff's arguments will take up the questions.

NOTES

[1] As Susan Sterrett kindly pointed out to me, along with the differences in Turing's imitation games.

[2] Madelyn Rose joke: What two Presidents were named Rose? (Answer: Teddy and Franklin.)

[3] Last night's Madelyn explanation: Venus moves faster than Earth which moves faster than Mars. Why? Because Venus is closer to the sun than Earth and Earth is closer than Mars. Planets closer to the sun are hotter than planets farther from the sun. Hotter things move faster than cold things.

[4] In Carnap's spirit Campbell never makes clear whether the priorities he claims, if not to describe at least to intimate, are causal or temporal (in the developmental sense) or logical

(if that the logic is mysterious), or if something else, what. Carnap's spirit is otherwise missing.

[5] I do not mean to suggest only Gopnik and Meltzoff present them. Related ideas may be found, for example, in many papers by Susan Carey and by Henry Wellman.

[6] Again, I am indebted to Susan Sterrett for reminding me of Kuhn's passage.

[7] The quality of the objections reminds me of the infamous inductive proof, attributed to David Kaplan, that an eminent and overpublished philosopher has written an infinity of books: he has written a book, and for every book he has written, he has written a worse.

[8] There is a sort of pseudo, quasi, crypto AI literature on cognitive development written by cognitive psychologists, but it is so much cold fusion. For a recent example, Thelen and Smith (1993) offer a "dynamical systems" account of development marked by buzz words rather than differential equations. Madelyn's explanations are better.

[9] I am sorry to have reached the age at which I have anecdotes. One day in the 1970s Hartry Field and I were talking at lunch about Kuhn's views. Hartry said (I think, maybe it was me) that Kuhn believes that when scientists started believing in electrons, electrons popped into existence. I (or maybe it was Harty) denied that Kuhn believed any such thing. Just then Kuhn walked by our table and I stopped him and asked him. "Of course", he answered, "I think electrons came into existence when scientists began to believe in them". Kuhn was always, always serious. For relativistic reliability analyses see Kelly (1997) and Kelly and Glymour (1992).

[10] The first experiment in Baker et al. (1993) reports adult behavior which the authors characterize as sub-normative, but which is a perfectly correct application of the Markov principle given the data and information available to the subjects. For a critical assessment making essentially my point, see B. Spellman (1996).

REFERENCES

Baker et al.: 1993, 'Selective Associations and Causality Judgements: The Presence of a Strong Causal Factor May Reduce Judgements of a Weaker One', *Journal of Experimental Psychology: Learning, Memory and Cognition* **19**, 414–432.

Campbell, J.: 1994, *Past, Space and Self*, MIT Press, Cambridge, MA.

Cheng, P.: 1997, 'From Covariation to Causation: A Causal Power Theory', *Psychological Review* **104**, 367–405.

Kelly, K.: 1997, *The Logic of Reliable Inquiry*, Oxford University Press, Oxford.

Kelly, K. and C. Glymour: 'Inductive Inference from Theory Laden Data', *Journal of Philosophical Logic* **21**, 391–444.

Klayman, J and Y. Ha: 1987, 'Confirmation, Disconfirmation and Information in Hypotheis Testing', *Psychological Review* **94**, 211–28.

Spellman. B.: 1996, 'Acting as Intuitive Scientists: Contingency Judgements Are Made while Controlling for Alternative Potential Causes', *Psychological Science* **7**, 337–42.

Thelen and Smith: 1993, *A Dynamical Systems Approach to Development*, MIT Press, Cambridge, MA.

Department of Philosophy
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213-3890
U.S.A.