THE WELL-DESIGNED CHILD

John McCarthy, Stanford University

2008 Sep 18, 5:28 p.m.

Abstract

This article is inspired by recent psychological studies confirming that a child is not born a *blank slate* but has important innate capabilities. An important part of the "learning" required to deal with the three dimensional world of objects, processes, and other beings was done by evolution. Each child need not do this learning itself.

By the 1950s there were already proposals to advance artificial intelligence by building a child machine that would learn from experience just as a human child does. What innate knowledge the child machine should be equipped with was ignored. I suppose the child machine was supposed to be a blank slate.

Whatever innate knowledge a human baby may possess, we are interested in a *well-designed* that has all we can give it. To some extent, this paper is an exercise in wishful thinking.

The innate mental structure that equips a child to interact succesfully with the world includes more than the *universal grammar* of linguistic syntax postulated by Noam Chomsky. The world itself has structures, and nature has evolved brains with ways of recognizing them and representing information about them. For example, objects continue to exist when not being perceived, and children (and dogs) are very likely "designed" to interpret sensory inputs in terms of such persistent objects. Moreover, objects usually move continuously, passing through intermediate points, and perceiving motion that way may also be innate. What a child learns about the world is based on its innate mental structure.

This article concerns *designing* adequate mental structures including a *language of thought*. This *designer stance* applies to designing robots, but we also hope it will help understand universal human mental structures. We consider what structures would be useful and how

the innateness of a few of the structures might be tested experimentally in humans and animals.

In the course of its existence we'll want our robot child to change. Some of the changes will be be development, others learning. However, this article mainly takes a static view, because we don't know how to treat growth and development and can do only a little with learning.

Contents

1 Introduction

René Descartes proposed that a philosopher should assume as little about the world as possible and gradually build reliable knowledge using step-by-step reasoning, observation and experiment. John Locke proposed that a baby starts out as a "blank slate". Bertrand Russell [Rus13] proposed starting with sensation and building up a theory of the world on that foundation. Positivist philosophy and behaviorism in psychology advocated the same methodology.[1]

Likewise, the AI learning literature is based on learning to recognize patterns in the inputs to a machine or computer program. A baby that started with its sensations and built a world-model from that might be called a *Lockean baby.* I don't know whether any computer program starting from sensation has ever learned the existence of semi-permanent physical objects that persist even when not perceived.

For a philosopher, starting from sensation and building up from there has the advantage of avoiding *a priori* assumptions, but neither actual science nor common sense works that way. Instead there is almost always a complex structure of ideas that is modified piecemeal.

Evolution solved a different problem than that of starting a baby with no *a priori* assumptions.

Instead of building babies as Lockean philosophers taking nothing but their sensations for granted, evolution produced babies with innate prejudices that correspond to facts about the world and babies' positions in it. Learning starts from these prejudices.[2] Evolution isn't perfect and human babies don't have all useful prejudices. What is the world like, and what are these instinctive prejudices?[3]

This paper studies the problem as follows.

- We ask what the world is like at the level at which people and

[1] The earlier web version of this paper mistakenly ascribed the blank slate doctrine to Descartes.

[2] There is a complication. Appropriate experience is often required for the genetically determined structures to develop properly, and much of this experience can be regarded as learning.

[3] I don't argue that a Lockean baby wouldn't work at all. Only that it would have a much longer babyhood than human babies do. Even *"universal grammar"* might be learned from experience. I just think evolution has learned to build in many of these features. Therefore, it is an empirical question whether a particular ability is learned or innate.

robots interact with it. Particularly important is what we call *the common sense informatic situation.* It relates the partial information about the world that can be obtained and the kinds of results that can be achieved in the world with these actions.[4]

- Next we ask what knowledge would be useful to build into a robot or for nature to have built into babies. We do this without regard to which feature are actually present.
- Now we ask what features seem to be present in babies.
- Finally we consider what experiments have been made and can be made to discover what innate knowledge nature has given us.

In so far as we have an idea what innate knowledge of the world would be useful, AI can work on putting it into robots, and cognitive science and philosophy can look for evidence of how much of it evolved in humans. This is the *designer stance.*[5]

2 What the World is Like

The most straightforward philosophical way of thinking about the world's interaction with a baby or other person is in terms of its input-output relations with its environment. Unfortunately for our philosophical convenience, but fortunately for our survival, this is not the way the world is structured.

The world's structure is not directly describable in terms of the input-output relations of a person. The basic structure of the world involves the interaction of elementary particles on time scales of 10^{-25} seconds, but intelligence did not evolve in structures of small numbers (mere billions) of elementary particles. When intelligence evolved, it was in structures of the order of 10^{26} elementary particles and time scales of the order of 10^{-1} seconds to years and very complex hierarchical structures. Even then only some of the higher level and slower objects and events are directly perceivable. Even bacteria, weighing

[4]We emphasize the effect of the actions on the world and not the new sensations that result from the action.

[5]Designer stance is related to Daniel Dennett's design stance [Den78], but Aaron Sloman has persuaded me that I was not using it quite in the way Dennett used design stance. 1999 note: Dennett has approved this usage of *design stance*, but I still want a distinct term.

one picogram 10^{-12} grams, have about 10^{10} atoms. The mass of a small virus is about 10 attograms [attogram = 10^{-18} grams].

Even on the human size and time scale, the world is not structured in terms of human input-output relations. Moreover, much of the determinism of the world at the microscopic level appears as non-deterministic at the level at which a person can interact with the world. [6]

Animal behavior, including human intelligence, evolved to survive and succeed in this complex, partially observable and very slightly controllable world. The main features of this world have existed for several billion years and should not have to be learned anew by each person or animal. In order to understand how a well-designed baby should work, we need to understand what the world is like at the gross level of human interaction with the environment.

Here are some of the world's characteristics. A baby innately equipped to deal with them will outperform a Lockean baby.

appearance and reality Some properties of the world are stable even though their appearances change. Objects last from seconds to centuries, while appearances change in fractions of a second. Therefore, humans, animals and robots are better off representing information about objects in so far as it can be obtained by observation and inferred from past experience or is innate.[7] [McC99] presents a puzzle in which the subject must experiment to determine the 3-d reality behind the 2-d appearances.

things of interest Some aspects of the world are relevant to an animal's or person's survival or prosperity, and others are not. However, notice that human and animal curiosity concerns many aspects of the world not related to survival or enjoyment. Other details of shape and pattern are not interesting.

semi-permanent objects Much of the world consists of three-dimensional objects that have masses, moments, compliances, hardnesses, chemical composition, shapes, outer surfaces with textures and

[6]Whether a baseball will pass over the plate is approximately deterministic once it leaves the pitcher's hand, but the batter and the spectators have to guess.

[7]Kant distinguished between appearance and reality, but AI and psychology need to study the distinction at a more mundane level than Kantian philosophers have brought themselves to do. I don't believe the study of platonic or neo-platonic forms will help understand the relation between physical dogs and the various ways they could affect the senses of a child or robot.

colors, are often made of identifiable parts and which move relative to the rest of the object. A particular object can disappear from perception and reappear again. The location of an object in the world is more persistent than its location in the visual field.

Objects usually have internal structures that are not apparent to human senses.

A baby seems to have an innate interest in the names of things quite apart from what may be immediately useful. Thinking of it linguistically, it is an interest in semantics, not just in syntax. We'd better build that into our robotic children.

continuity of motion Objects move continuously passing through intermediate points and intermediate orientations.

continuous processes Besides moving objects, there are many continuous processes with intermediate states.

two dimensional world Because of gravity, much of the world is two dimensional with its simpler topology. Paths can block other paths.

specific objects The environmment of a child contains other people, usually including a mother, and parts of people including parts of the child itself. Objects often have parts which are objects. However, often only some of the parts are separately identifiable. The boundaries of the parts are often not definite.

solidity Objects that are solid do not ordinarily penetrate one another. Some are rigid and some are flexible.

gravity Objects require supporting surfaces, and an unsupported object falls to a lower surface.

kinds of objects Objects have kinds, and objects of the same kind have properties associated with the kind.[8] Babies are ready very early to learn what kinds there are.

relations Objects not only have individual properties and belong to kinds, but different objects and kinds have relations with one another. At least some ternary relations such as betweenness

[8] It might be more parsimonious intellectually to have just a relation of similarity between objects. However, the world as it is justifies the bolder attitude that there are kinds, and we should build this into our robots and expect it in our children. The use of nouns in language presupposes more than just similarity relations.

are basic. Also "A is to B as C is to D" seems to be basic. In its numerical use, it reduces to the equality of two fractions, but the quaternary relation seems to be basic in common sense usage.

In philosophy, AI, and computer science, there is an overemphasis on unary relations, i.e. properties.

natural kinds Many of the objects a child encounters, e.g. lemons, belong to *natural kinds*. The objects of a natural kind have yet undiscovered properties in common. Therefore, a natural kind is not usually definable by an *if-and-only-if* sentence formulated in terms of observables.

fundamental kinds Animate objects are to be understood in terms of their desires and actions. Inanimate objects are passive. Some objects are edible by humans and some are not. These kinds pervade the baby's environment.

abstractions Kinds belong to higher kinds and have relations. Red is a color and color is a quality. This is a fact of logic rather than about the physical world, but its usefulness is dependent on objects being naturally grouped into kinds rather than being all completely different.

sets and numbers There are sets of objects and other entities. This includes both sets of objects perceivable on a single occasion and sets organized more abstractly. Sets can often be counted. Some are more numerous than others and this is significant. Sets can be used up, e.g. all the food can be eaten.

situations and kinds of situations Kinds of situations recur.

the body The baby itself and its parts are objects.

movability Some objects can be moved with the arms and legs of a child.

responses Mothers help a baby that cries.

love A mother loves her baby.

unimportant aspects Many aspects of the world are ordinarily unimportant for a human or animal. For many purposes, shadows are mere epiphenomena.

quantitative physics Humans could act more precisely if our senses gave us numerical measures of time, distance, velocity, humidity, temperature, etc.; our minds could do rapid arithmetic with them, and we could give numerical values to the signals telling our muscles how fast to contract. Nature didn't give us this, but we can build it into our robots as an add-on to the kinds of semi-quantitative information human senses give us.

Newtonian physics While the world is not fully determinate at the level at which humans interact with it, many events are related in a simple numerical way. For example, $s = \frac{1}{2}gt^2$ describes the distance a body will fall, and hot bodies cool at a rate proportional to the difference in temperature between a body and its surroundings.

atoms The material world is built up from atoms and molecules. It is more fundamental than most of the above facts but is similar to them. While even ancient Greek philosophers like Democritus could conjecture that the world was built from atoms, John Dalton was the first person to offer scientific evidence for the fact.

mathematics Very complex structures, (e.g. groups, rings and fields), exist in a mathematical sense.

mathematics of the world Very complex mathematics is "unreasonably effective" in understanding and controlling the physical world.

All the above are facts about the world. All but the last few may or may not be represented innately. We can also imagine that we might have evolved innate knowledge of the above mathematically expressible facts, but alas we didn't. The items listed are certainly not a complete set of facts about the commonsense world that a well-designed child might know about. Moreover, innate mechanisms for dealing with phenomena related to these facts do not always take a form describable as having certain knowledge.

In the next section we consider which of the above facts a child might know about or have special mechanisms for dealing with.

3 Human Mental Characteristics

Here are some human mental characteristics that affect what abilities might be innate.

evolved from animals The human was not designed from scratch. All our capabilities are elaborations of those present in in animals. Daniel Dennett [Den78] discusses this in the article "Can a Computer Feel Pain". Human pain is a far more complex phenomenon than an inventor would design or a philosopher intuit by introspection. As Dennett describes, kinds of pain are associated with levels of organization, e.g. some are in the structures we share with reptiles.[9]

distributed mechanisms We are descended from animals that mostly have separate neural mechanisms controlling separate aspects of their lives. We have these separate mechanisms too but are more capable than animals of observing their state and integrating their effects.

central decision making A mobile animal can go in only one direction at a time. Therefore, animals above a certain level, including all vertebrates, have central mechanisms for making certain decisions. Very likely sponges don't need a central mechanism.

little short term memory Compared to computers, humans have very little short term memory. In writing a computer program it is difficult to restrict oneself to a short term memory of 7 ± 2 items.

slowness Human performance is limited by how slowly we process information. If we could process it faster we could do better, and people who think faster than others have advantages. For this reason we need to perceive states of motion and not merely snapshots. Computer programs often work with snapshots, but even they suffer from slowness when they don't represent states of motion directly.

incompleteness of appearance When a person looks at a scene, only part of the information available seems to go all the way in. There is the blind spot, but there is more incompleteness than that. What seems subjectively to be a complete picture really isn't. The picture has to be smoothed over in such a way that a detailed look at a part of it sees no inconsistency. While the phenomenon is most obvious for vision, it surely exists for the other senses as well.

[9]Here the facts of evolution have an observable payoff akin to Haeckel's "Ontogeny recapitulates phylogeny".

memories of appearance I suppose this opinion will be controversial among psychologists and neurophysiologists, but I state it anyway. What humans remember about the appearance of an object are attached to their more stable memories of its physical structure and maybe even to memories of its function. For example, my pocket knife is in my pocket, and I remember what blades it has. If required to draw it from memory, I would consult this memory of its structure and draw that. Only a small part of the information used would be visual memories. Physical structure is more stable than appearance.

curiosity Humans and animals are curious about the world. Just how curiosity is focussed isn't obvious.

supposed to do It is often asserted that children learn what to do in situations by being rewarded. The innate mechanism may be more powerful than that.

Children and adults have a concept that in a particular kind of situation there are actions "that one is supposed to do". One learns what one is supposed to do and does it without reinforcement of the specific kind of response. Example: I told several people, "See you later," and an 18 month old baby whom I was not specifically addressing said, "Bye-bye". Children who try to learn what they are supposed to do in a situation and do it will survive better than those who need to learn responses by reinforcement. The race was reinforced—or maybe it was our mammalian ancestors.

senses The characteristics of human senses are an accidental consequence of our evolution and our individual development. A blind person lives in the same world of objects as a sighted person. It is just that sighted persons have an advantage in learning about them. A person with an infra-red detecting pit in his forehead like a pit viper (or some computer terminals) would have a further advantage in distinguishing people, warm-blooded animals and stoves. A person with a bat-like sonar might "see" internal surfaces of itself and other people.

This is not the best of all possible worlds—only a pretty good one.

It would be interesting to look more closely at how human mental characteristics differ from those of animals.

4 What Abilities Could Usefully be Innate?

Taking into account what the world is like and what our nervous systems are like, what knowledge and abilities are possibly and usefully innate? Many of them correspond to the facts about the world discussed in section 2.

some objects persist even while not sensed Having this prejudice is fundamental to the survival of humans—and probably to other land vertebrates. A dog chasing a ball will look for it if it disappears behind something.

identify object Identify a part of the current stimulus pattern as coming from an object. Remember aspects of the object as the same as a previous object or as coming from a new object. The task is basically the same whether the stimuli are visual, tactile, auditory or olfactory or a combination. Success involves recognizing repeated instances of the same object or the same kind of object. Present machine learning schemes are more suited to recognizing kinds of objects than for recognizing individual objects. Both are needed.

What innate structures are suited for this? At least some of these structures are independent of the sensory modality.

natural kinds The child is predisposed to name kinds of entities and to expect that the objects of a kind that is recognized by superficial properties will have additional properties in common. For example, adults call some objects lemons, and all lemons turn out to have similar taste and to have similar seeds.

three-dimensional objects The world contains three-dimensional objects, and humans know about them. While non-blind people usually get most of their information about objects from seeing them, what we know about objects should not be regarded as a collection of 2-d pictures. The objects are far more stable than pictures of them can be, because they are seen at a variety of angles and lighting conditions. We learn about objects from 2-d pictures, but they are not constructs from 2-d pictures.

Advocates of an initial *tabula rasa* have proposed that a baby *learns* that its sensations should be organized around external objects. Maybe a mechanism for learning this could exist, but a baby would learn faster if this much were innate. In fact animal thought also seems to presume external objects. Many specific instincts, e.g. related to hunting, presuppose them.

A baby also has no difficulty with two dimensional representations of three dimensional objects. A baby apparently doesn't have to be taught that a picture of a dog in a book represents some real dog.

objects have colors Our visual system goes to a lot of trouble to ascribe colors to objects in ways that are independent of lighting. When this fails, we notice it.

expect an object to have a location Since a physical object a person has perceived ordinarily continues to have a location even when it is no longer perceived, because it or the person has moved, it is advantageous for the person to expect it to have a location. He might want to look for it or reason about its effects on other objects, e.g. as described in [Spe94].

perceive motion as continuous Although our visual perceptions of objects are discrete because of our saccadic movements, we perceive objects as moving continuously. We evolved to interpret our sense data, and not just visual sense data, in terms of continuous motion. Perceiving motion as continuous may have evolved very early among vertebrates. I suppose this involves an approximate differentiation of the position.

recognize parts Recognize parts of an object and their relations to the others. It would be interesting if there were an ability to recognize certain physical structures, e.g. towers and walls, analogous to the ability to recognize a grammatical sentence.

kind of situation Identify the current situation as being of a certain kind.

focussed curiosity In the Shannon quantitative measure of information, there is just as much information to be obtained from the pattern of saw marks on the boards of my office wall as there is about what is available for lunch or what can be obtained by research on artificial intelligence. Curiosity needs to be focussed on what is potentially relevant to the baby or robot. Notice that human curiosity, as it ought to be, is quite broad—but it is also selective. Part of the answer is that curiosity is focussed on getting more information about kinds of object that have been identified.

noise rejection Certain appearances are usually noise, e.g. shadows. The child may be predisposed to regard shadows as noise, i.e. to regard an object as continuing through a shadow and to ignore the edges of shadows. Elizabeth Spelke [Spe94] considers the recognition of shadows to be non-innate.

grammar of goal regression The recognition that a goal is achievable because it is either already achieved or all the preconditions of an action that achieves it are achievable. This can be regarded as the grammar of a specific language GR, but unlike the grammar of a spoken languages, the grammar of GR is universal.[10]

principle of mediocrity The baby is like other people. It can learn about its own capabilities from observing others, and it can learn about others by putting itself in their places. [11]

introspection Recent work in psychology, [FO99] and [JHFF00], shows that children develop some introspective ability by age 3, and this ability improves with age. [McC96] discusses the introspective abilities required by a robot.

pointer effect When one uses a pointer, e.g. a pencil, to explore or manipulate in a container, one's senses refer to the end of the pointer and not to one's hands. This seems to be innate, but is not a feature of helpless young babies. Maybe there's a standard name for what I've called "pointer effect".

It would be interesting if there were innate non-linguistic human mental abilities that are not present in animals. Nothing appears obvious, but maybe the innate part of human number sense is qualitatively different from that of animals.

Some abilities require early experience to acquire. For example, people blind from birth who gain sight as adults don't acquire an image processing system fully adapted to the world as it is. However, there is no reason to expect that they could acquire an image processing

[10]This may be worth a small pound on the table. It would seem that a person, and maybe even some animals, can test whether a goal is achievable by parsing the goal regression structure. Of course, there are limits in how big a structure can be parsed, but the *competence* puts no limit on the size. It may be that goal-regression memory is in addition to other short term memory and can only be used in connection with remembering goals.

[11] Astronomers use "principle of mediocrity" for the hypothesis that there is nothing special on the average about our own part of the universe or about our own point in time.

system adapted to a quite different visual world. If this is so, then the image processing system is still basically innate.

5 Features of a Language of Thought

Cognitive scientists argue about whether there is a *language of thought*, but its advocates haven't told us much about what it is like. Stephen Pinker, an advocate, only tells us in [Pin94]

> The hypothetical "language of thought", or representation of concepts and propositions in the brain in which ideas, including the meanings of words and sentences, are represented.

A language of thought that might be used for robots or looked for in humans is constrained by the characteristics of the baby's world and the characteristics of the non-linguistic parts of the baby's mind—including its limitations.

Here are some ideas about mentalese.

grammar is secondary While most linguistic studies have focussed on grammar, meaning is more important—in studying spoken language, in proposing a language of thought and in designing robots. A child's first speech consists of words which are attached to things, or to appearances of things, or to sometimes ambiguous combinations of things and appearances. "Doggie" is stimulated by the sight of a dog, a picture, an animal on TV, the sound of barking and conversation about dogs.

maybe language starts with naming A human child starts language learning with names for objects. This desire is independent of having any goal concerning the object. We have the option of designing an artificial child to know a lot of language, e.g. English and/or a logical language from the beginning. Different experimenters will explore different approaches.

parallel information Images are presumably represented in parallel. There is nowhere anything like a television signal processor that handles a picture serially and repeatedly spreads it out. This is obvious for pictures but surely applies to a lot of other kinds of information. On the other hand, our inability to think completely in parallel shows that many higher mental functions are done serially.

logic For a robot, a logical language [12] will be most suitable, but some appropriate ascriptions of beliefs and intentions to robots will refer to information represented non-logically. Humans *probably* don't use quantificational logic at the pre-verbal level, although we can use it when we have to, and formal logic is often helpful when the information is mathematical. Here's why I only say *probably.* Consider the sentence "For every boy there's a girl who loves only him." Its predicate calculus representation, $(\forall b)((\exists g)(Loves(g,b) \wedge (\forall b')(Loves(g,b') \rightarrow b' = b))$, has three embedded quantifiers. We then ask the question, "What can you say about the total number of boys and girls?" A fair number of people uneducated in logic find the correct answer that there must be at least as many girls as boys. I haven't done the experiment thoroughly, but the results suggest that the sentence with the three levels of quantifiers is understandable by many logically uneducated people and therefore its content is somehow internally represented.

Let's design it into our child.

a word at a time Sentences uttered by humans are usually not preformed in entirety before being uttered. A human starts a sentence and thinks how to continue and finish it as he continues talking. (The obvious argument is from introspection, but I suppose experiments would confirm it.) Humans can preform sentences with some effort. Vladimir Bukovsky tells about having composed a whole book in prison while denied paper.

chemical state Suppose a person is hungry—a condition humans share with dogs. This can perfectly well be only be represented by the chemical state of the blood stream. There is no reason to have anything like the sentence, "I am hungry" anywhere in the brain until the fact has to be communicated. Similarly we don't need anything like a sentence in the memory of a computer to represent the voltage of its battery.

virtual sentences We may regard information that is directly represented by the chemical state of the bloodstream or by a voltage as expressed in *virtual sentences* along the lines of [McC79] or [New82]. We may then sometimes be able to explain some

[12] Logicians do not consider logic itself to be a language, but rather consider a language to be defined by the predicate and function symbols that are used with the logic. This is a valuable distinction and AI and cognitive science researchers should maintain it.

actions as involving logical inference involving the virtual sentences.

immediate reference Thinking about an object before one's eyes does not require that it have a name. Something like a pointer to a structure will do as well. We can see this, because when we have to mention an object in speech we have to think of a name that will enable the hearer to establish his own pointer to his mental structure representing the object in question. Purely internal symbolic names as in Fodor's proposed *language of thought* may be useful even if they aren't communicable.

short thoughts Thoughts are not like long sentences, although a long sentence may be required to express a thought to another person because of a need to translate internal pointers into descriptions.

communication When the fact of hunger or low battery voltage has to be communicated something like a sentence is needed. Let's call it a pseudo-sentence until we find out more. However, a pseudo-sentence isn't needed to stimulate eating. It also isn't necessary to represent the rule, "if hungry, then eat". In view of evolution, one would expect the fact of being hungry to be represented both chemically and in the language of thought.

future There are other uses besides communication for sentence-like forms. Very likely, the expectation of being hungry by dinner time needs something different from a substance in the blood for its representation.

reasoning The language of thought is used for reasoning.

not like spoken languages English and other spoken languages won't do as languages of thought. Here are some reasons.

- Much mental information is represented in parallel and is processed in parallel.
- Reference to states of the senses and the body has to be done by something like pointers. Natural languages use descriptions, because one person can't give another a pointer to his visual cortex.[13]
- We don't think in terms of long sentences.

[13] A robot might tell another robot, "Look through my eyes, and you'll see it.

- Much human thought is contiguous with the thought of the animals from which we evolved.
- For robots, logic is appropriate, but a robot internal language may also include pointers.
- A language of thought must operate on a shorter time scale than speech does. A batter needs to do at least some thinking about a pitched ball, and a fielder often needs to do quite a bit of thinking about where to throw the ball. Pointers to processes while they are operating may be important elements of its sentences.

I think there are additional reasons, but I haven't been able to formulate them.

The language of thought may undergo major reorganizations. This may be one reason why there is so little memory of early life. Almost no-one can remember nursing or drinking from a baby bottle.

6 Experimental Possibilities

6.1 AI

There are two kinds of AI experimental possibilities. The first is to use the ideas of this article to try to break AI systems intended to deal with the common sense world that lack some of the capabilities discussed in this article. The second is to use the ideas to build an AI system. Since the ideas are not advertised as complete enough to serve as a design, the first option seems more fun to pursue.

6.2 Psychological Experiments

Elizabeth Spelke [Spe94] describes a number of experiments that she and others have done to discover and verify innate mental abilities. The basic technique uses the fact that a baby will look longer at something surprising than at something that seems familiar.

Here's one that was first done in 1973 [Bal73] and was repeated by Spelke in 1993. There are experimental babies and control babies and the experiment has two phases. In the first phase the control babies are shown nothing. The experimental babies see an object go behind a screen and shortly another object emerges on the other side of the screen. The timing is such as would be appropriate if the first object struck the second object and knocked it from behind the screen. The babies are shown the phenomenon enough times to get bored with it and stop paying attention.

In the second phase of the experiment the screen is removed. There are two variants. In the first variant, the first object strikes the second and knocks it onward. In the second variant the first object stops short of the second, but the second object takes off as though it had been struck. The control babies look at both variants for the same amount of time. The experimental babies look longer at the second variant.

The conclusion is that the experimental babies inferred that the first object had struck the second when the event occurred behind the screen. When the screen was removed, they were not surprised when the expected event was shown to occur but were surprised and looked longer when this expectation was not met.

The conclusion is that babies have innate expectations about dynamics, i.e. are well-designed in that respect. For details see [Spe94].

That was an actual experiment. Now consider some possible experiments.

Suppose we want to determine whether some abilities concerned with a specific fact about how the world is organized is innate. We compare a baby's ability to use this fact compared to its ability to learn a fact about an environment constructed differently from our world but logically no more complex.

Here are some possibilities. Since I am rather innocent of the psychological literature some of them may already have been tried.

three-dimensional objects I'm skeptical that a person's notion of a physical object is fundamentally visual. Here's an informal experiment I actually did. The subjects attempted to draw a statuette in a paper bag. They could put their hands into the bag and feel it as much as they wanted to. The quality of a subject's drawing, except for surface colors, was similar to what that subject would have produced looking at the object except in one case. The object was a statuette of an owl, and the subject who misperceived it as an angel produced an inferior drawing.

It would be worthwhile to use this and analogous techniques to explore people's concepts of three-dimensional objects. I would think that it is possible to investigate how babies perceive objects they are only allowed to touch and then see. The baby could feel an object in a paper bag and then see either the same object or a different object. The hypothesis is that the subject would regard seeing the same object as less surprising than seeing a different object.

anticipating the future To eat when hungry doesn't require having in mind anything like a sentence. However, to know that one will be hungry 4 hours from now may require it. Maybe this is where humans and apes part company. Can an ape that is not hungry perform a non-habitual action, e.g. putting a key by an empty food box, in anticipation of being hungry later?

unethical experiment A Lockean baby would do as well in flatland as in our space. Imagine arranging that all a baby ever sees is a plan of a two-dimensional room and all his actions move around in the room. Maybe the experiment can be modified to be safe and still be informative.

continuity of motion The Lockean baby is brought up in an environment in which motion is discrete. Imagine that the baby's world is a Macintosh screen. Objects move without passing through intermediate points. The baby moves an object by clicking on the initial and final locations. The experiment is to determine how well a baby will do in such a world. This one might be tried with an animal.

attention experiment If a baby is built to expect objects to behave as solids, then it will be surprised when objects appear to interpenetrate. It might pay longer attention to such a scene.

inconsistencies Babies might or might not find Escher-type drawings surprising.

geographical representation Consider a maze with a glass top. Does it help an animal find food if it can walk around on the top of the maze before entering it? The top could have small holes that the smell of the food could get through. One psychologist opined that dogs would be helped and rats would not.

The experiment would test whether the animal can represent a scene by something like an image.

goal regression in animals An animal seeks a goal but discovers that a precondition must be achieved first and undertakes to do it. Then it discovers a precondition for the precondition, etc. Suppose the animal has been trained to achieve B as a precondition for achieving when A isn't already true. It has been trained to achieve C as a precondition to achieving B when B isn't already true, etc. We ask how far the animal can carry the regression. Say the animals are dog which vary in intelligence, or at least vary in the ability to learn the tasks that humans teach dogs. We ask is there an innate limit for dogs or can smart dogs carry it farther than dumb dogs.

Susan McCarthy informs me that when a performing animal is taught a new trick, the trainer starts with the bow at the end and works backwards. I don't know if this is related to goal regression.

grammar via meaning Many of the discussions of a child learning its native language seem to assume that the child learns grammar solely by observing grammatical regularities in speech and having its grammar corrected. Consider a child raised by an English speaking nanny whose native language is Spanish and is addicted to Mexican soap operas. It seems to me that this happens often enough so that observations could be made. The child would then hear a lot of idiomatic Spanish. It would be interesting to observe whether the child would be able to tell grammatical from ungrammatical Spanish sentence.

My conjecture is that grammar is learned as an auxiliary to meaning and is not separately represented in the brain.

7 The Well-designed Logical Robot Child

Ever since the 1950s, people have suggested that the easy way to achieve artificial intelligence is to build an artificial baby and have it learn from experience. Actual attempts to do this have always failed, and I think this is because they were based on the Lockean baby model.

This section concerns the design of a robot child that has some chance of learning from experience and education. We do not mean reprogramming, which is analogous to education by brain surgery. The instructor, if any, should have to know the subject matter and very no more about how the program or hardware works than parents know about the physiology of their children.

Consider designing a logical robot child, although using logic is not the only approach that might work. In a logical child, the innate information takes the form of axioms in some language of mathematical logic. [14]

[McC79] and [New82] both discuss using logical sentences to represent the "state of mind" of a system that doesn't use sentences directly. We don't mean that here. We are discussing a system that uses logical sentences explicitly. If you don't like this approach, read on anyway and then decide how your favorite approach would handle the problems we propose to solve with logical axioms.

We will deal with just four innate structures among those mentioned in Section 2. These are *the relation of appearance and reality*, *persistent objects*, the *spacial and temporal continuity of perception* and the *language of thought*. They are all difficult, and we can't yet go beyond sketching the kinds of sentences that might be used by the robot child. The design of the child robot requires many more.

7.1 Appearance and Reality

It is the essence of our approach that appearance and reality are quite distinct and the child is designed to discover information about reality via appearance. See [McC99], and solve the problem it presents of finding the reality behind and appearance. We take a rather brute force logical attitude by making their relations explicit.

In our formalism, both appearances of objects and physical objects will be represented as logical objects, i.e. as the values of variables and terms. Thus the ontology includes both appearances and objects.

Our examples of appearances will mainly be visual appearances, because we understand them better than auditory, tactile or olfactory appearances. However, we would like a language that applies to combinations of all kinds of appearances—whatever happens to be available.

Natural language is better at describing objects than appearances. When it has to describe appearances, it often uses objects to describe them—as in "a cloud shaped like a lion's head." This is for two reasons. First, appearances are represented in thought by something like pointers to the appearance itself and thus not readily communicated. Second, appearances are fleeting and can't be fully re-examined. Our

[14] There also has to be a program using the logical sentences, and efficiency will very likely require it to use declaratively expressed heuristics to guide its search. Very little progress has been made in that direction, so we will ignore heuristic control in this article.

robot's language of thought could use pointers to pictures, e.g. gifs. These would be communicable.

Show a hungry child a picture of a hamburger and ask "What's that?"

Answer: "A hamburger".

"So eat it."

"Don't be silly. It's not a real hamburger."

The most obvious predicate in our logical language relates an appearance to an object. Thus we may have a sentence

$$Appears(appearance, object), \tag{1}$$

in a simple context, but this simple formula requires several elaborations.

- Truth of (1) depends on the situation. We can write a situation calculus formula

$$Holds(Appears(appearance, object), s), \tag{2}$$

  but we are more inclined to use the context mechanism of [McC93], although it is somewhat more complex to explain.

- Both the appearance and the object are made up of parts, and the correspondence of these parts often must be stated.

- The correspondence is usually not complete. Some parts of the appearance are artifacts or irrelevant, and some parts of the object are not perceived.

- It is common that the appearance changes during the lifetime of the language of thought sentences asserting the correspondence.

- If the correspondence is to be used to guide motor activity, we need not merely to state that a given part of the appearance corresponds to a leg of a certain chair but also to tell how the orientation in appearance space of the appearance of the leg corresponds to the orientation in physical space of the leg itself.

We need logical formulas for expressing these kinds of facts. It is more straightforward to do when the appearance is visual than when the appearance is tactile. How do we describe the appearance of an object to a blind person who has not yet felt it with his hands? We share with the blind Euclidean geometry extended to what we may call Euclidean physics.

7.2 Persistence of objects

Maybe this part isn't so difficult now that objects are distinguished from appearances. Objects have properties, parts and relations to one another. They also have situation dependent locations and orientations.

Using a *situation calculus* formalism, $Location(object, s)$ gives the location of the object *object* in the situation s. However, the orientation of an object often needs to be stated, usually quite imprecisely.

7.3 Conservation

According to Piaget, notions of conserved quantity come fairly late. Piaget's classical example is asking a child whether a tall glass or a short glass has more liquid in it just after the liquid has been poured from one to the other. Piaget's classical result is that children younger than about seven pick the tall glass, citing the height.[15] Siegler in his textbook [Sie98] asserts that conservation arises more gradually with different conservation laws being learned at different times.

Suppose something appears and disappears. There are two kinds of mental models a person or robot can have of the phenomenon—flow models and conserved quantity models. Flow models are more generally applicable and apparently are psychologically more primitive. Thus water flows from the tap onto the hands, and water flows down the drain. This model does not require a notion of quantity of water. The same is true of a child's early experience with money. A parent gives you some and you buy something with it. When there is no way of quantifying the substance, as with the water flowing from the tap, the notion of conservation of water is of no help in understanding the phenomenon.

Siegler considers various conserved quantities—physical objects, numbers and liquids. Conservation of physical objects comes first. An object that has disappeared is regarded as being somewhere, and if the object is wanted, it is worthwhile to look for it. Conservation of number is not apparent to first graders, and they give silly answers to questions like $4+? = 7$.[16]

[15] Someone tried this on the child of a prominent AI researcher, eliciting the answer, "Oh, I'm not old enough to have conservation yet".

[16] Conservation of heat wasn't apparent to 17th century Italian experimenters such as Toricelli who used a flow model of heat. It wasn't until 1750 that Thomas Black discovered

Let's take the designer stance. It would be good if the notion of conservation law were innate, and experience taught which domains it applied to. Alas, we aren't built that well.

The notion of conserved quantity is more abstract than other early notions. The actor has to believe in there being a quantity of the entity in question, e.g. of water. [Sie98] suggests, p. 42, op. cit., that the child learns conservation of water via taking into account the cross section of a glass as well as its height. My opinion, which a suitable experiment might test, is that the abstract notion of conserved quantity is learned, and talking about the width of the glass is only window dressing, because not even Archimedes could do the geometry needed to confirm the conservation.[17]

A mathematical description of a conservation law may be interesting. Here's a situation calculus axiom saying that the amount of a quantity q is normally conserved.

Let $Amount(q, a, s)$ denote the amount of quantity q in reservoir a in situation s. We wish to say that if the occurence of the event e in situation s is not abnormal, then the amount of q in all the reservoirs together remains constant, i.e. q is conserved.

$$\neg ab(e, s) \rightarrow \sum_{a \in A} Amount(q, a, s) = \sum_{a \in A} Amount(q, a, result(e, s)). \quad (3)$$

The axiom (3) is probably too elaborate and general to express what real children know about conservation.

7.4 Continuity

Some philosophers and philosophically minded psychologists regard it as odd that we perceive our experience as continuous in time even though our nervous systems work discretely and our senses even more discretely. The spacial continuity of our spacial perceptions, e.g. visual, should be just as problematical. What tools should we give our logical robot child for dealing with this? What terms should we put in its language of thought?

---

that heat could be regarded as a quantity that moved from one object to another by conduction.

[17] Archimedes assumed that the volume of the king's crown was equal to the volume of the water it would displace, so he didn't need to make detailed measurements on the crown.

I think it is to our advantage that we don't perceive the discontinuities in our perception of continuous motion or the discontinuous frames of a movie.

This allows us to get velocities by differentiating positions or having the speedometer in a car do it for us. Going 50 mph usually lasts longer than being at any particular location and is therefore part of the situation, e.g. $Speed(Car1, s)$.

A robot child might well be designed to perceive discontinuous frames of a TV image as continuous. Its high speed computation would allow it to also perceive the discrete sequence of frames.

7.5 Fragments of a Language of Thought

In designing our logical robot, we can choose whether to represent and process certain information in a serial way or a parallel way. [18] In a parallel representation, a part of the information is represented by which wire the information is on in a computer system or where in the nervous system the information is located in a human or animal. When all the information is in a single memory, it has to contain labels. Our robot child will use a single memory, and therefore its *language of thought* will be more explicit in representing certain information than a human *language of thought* has to be.

The simplest way to represent a visual image in a computer is by a pointer to a pixel array, e.g. in lisp as $(GIF"72385.gif")$. This can then be the value of a variable or constant and programs can communicate it in this form. This has advantages and disadvantages.

- The information is readily displayable to a human or transmissible to another robot.
- Suitable programs are required if information is to be extracted from the image.
- If relations among parts of the image are to be expressed in l.o.t., a suitable language for this is required. It is also necessary to relate parts of the image to other things such as objects. For example, we need a relation that asserts that a certain part foo

[18] Some aspects of computing are going from serial to parallel in order to achieve greater speed, but a lot of communication within and among computers is going the other way for greater simplicity and reliability. Multi-wire cables are being replaced by single Ethernet or fiber optic cables. Truth and beauty are not to be found in a single direction.

of a certain picture taken from top to bottom represents Tom's left arm. It might be written

$$(represents \\ \quad (Top\text{-}to\text{-}bottom\ (Part\ foo\ (GIF\ \text{"72385.}gif\text{"}))) \\ \quad (Left\text{-}arm\ Tom)).$$

- This representation is unsuitable for mentally generated images, whether they be invented *ab initio* or modified from previous images. They will not be complete pictures.

These considerations suggest that robot l.o.t. should not represent images primarily by pictures, although pictures might be an auxiliary data type. Instead, curiously enough, the robot child will need something that is closer to what we imagine human image representation to be.

7.6 Consciousness

My opinion is that self-consciousness, i.e. the ability to observe some of ones own mental processes, is essential for full intelligence. Whether it is essential for babies and young children is another matter. [McC96] treats the question for robots from the designer stance, i.e. asks what self-conciouness it is useful to build into robots.

8 Remarks

The title of this essay comes from Stephen Pinker. In fact, the essay was stimulated by his book [Pin94]. While he expresses the opinion that the mind has many built-in characteristics and favors the idea of a language of thought, he elaborates neither idea. His chapter on language learning is exclusively devoted to learning grammar. I decided to see what I could do with a language of thought, and this led to other considerations.

This isn't the best of all possible worlds.

References

[Bal73] W. A. Ball. The perception of causality in the infant. 1973. Paper presented at the Society for Research in Child Development, Philadelphia.

[Den78] Daniel Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology.* Bradford Books/MIT Press, Cambridge, 1978.

[FO99] John H. Flavell and Anne K. O'Donnell. Development of intuitions about mental experiences. *Enfance*, 1999. in press.

[JHFF00] Frances L. Green John H. Flavell and Eleanor R. Flavell. Development of children's awareness of their own thoughts. *Journal of Cognition and Development*, 1:97–112, 2000.

[McC79] John McCarthy. Ascribing mental qualities to machines[19]. In Martin Ringle, editor, *Philosophical Perspectives in Artificial Intelligence.* Harvester Press, 1979. Reprinted in [McC90].

[McC90] John McCarthy. *Formalizing Common Sense: Papers by John McCarthy.* Ablex Publishing Corporation, 1990.

[McC93] John McCarthy. Notes on Formalizing Context[20]. In *IJCAI-93*, 1993.

[McC96] John McCarthy. Making Robots Conscious of their Mental States[21]. In Stephen Muggleton, editor, *Machine Intelligence 15.* Oxford University Press, 1996. Appeared in 2000. The web version is improved from that presented at Machine Intelligence 15 in 1995.

[McC99] John McCarthy. **appearance and reality**[22]. *web only for now, and perhaps for the future*, 1999. not fully publishable on paper, because it contains an essential imbedded applet.

[New82] A. Newell. The knowledge level. *AI*, 18(1):87–127, 1982.

[Pin94] Steven Pinker. *The Language Instinct.* Morrow, 1994.

[19]http://www-formal.stanford.edu/jmc/ascribing.html
[20]http://www-formal.stanford.edu/jmc/context.html
[21]http://www-formal.stanford.edu/jmc/consciousness.html
[22]http://www-formal.stanford.edu/jmc/appearance.html

[Rus13] Bertrand Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26, 1913.

[Sie98] *Children's Thinking, Third edition*. Prentice Hall, third edition, 1998.

[Spe94] Elizabeth Spelke. Initial knowlege: six suggestions. *Cognition*, 50:431–445, 1994.

/@sail.stanford.edu:/u/jmc/w95/child.tex: begun 1995 Mar 19, latexed 2008 Sep 18 at 5:28 p.m.