Contents

1 Introduction 3

2 Some historical remarks 6

3 Philosophical presuppositions of AI 7

4 Scientific Presuppositions of AI 15

5 Common sense and the common sense informatic situation 17
5.1 Bounded informatic situations 19
5.2 The general common sense informatic situation 20

6 The AI of philosophy—some advice 26

7 Information in contexts and about contexts 34

8 Conclusions and remarks 36

THE PHILOSOPHY OF AI AND THE AI OF PHILOSOPHY

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

25 June 2006

Abstract

The philosophy of X, where X is a science, involves philosophers analyzing the concepts of X and sometimes commenting on what concepts are or are not likely to be coherent. Artificial intelligence (AI) has closer scientific connections with philosophy than do other sciences, because AI shares many concepts with philosophy, e.g. action, consciousness, epistemology (what it is sensible to say about the world), and even free will. This article treats the philosophy of AI but also analyzes some concepts common to philosophy and AI from the standpoint of AI. The philosophy of X often involves advice to practitioners of X about what they can and cannot do.

We partly reverse the usual course and offer advice to philosophers, especially philosophers of mind. The AI point of view is that philosophical theories are useful to AI only if they don't preclude human-level artificial systems and provide a basis for *designing* systems with beliefs, do reasoning, and plan. AI research has particularly emphasized formalizing the actions available in a situation and the consequences of taking each of several actions. In order to do this, AI has mainly dealt with simple approximations to phenomena.

A key problem for both AI and philosophy is understanding common sense knowledge and abilities. We treat the notion of the *common sense informatic situation,* the situation a person or computer program is in when the knowledge available is partial both as to observation and as to theory, and ill-defined concepts must be used. Concepts ill-defined in general may be precise in specialized contexts.

1 Introduction

Richmond Thomason (Thomason 2003) wrote

> The relations between AI and philosophical logic are part of a larger story. It is hard to find a major philosophical theme that doesn't become entangled with issues having to do with reasoning. Implicatures, for instance, have to correspond to inferences that can be carried out by a rational interpreter of discourse. Whatever causality is, causal relations should be inferable in everyday common sense settings. Whatever belief is, it should be possible for rational agents to make plausible inferences about the beliefs of other agents. The goals and standing constraints that inform a rational agent's behavior must permit the formation of reasonable plans.

The relation of AI and philosophy involves many concepts that both subjects include—for example, action, goals, knowledge, belief, and consciousness. However, AI takes what we may call the *designer stance* about these concepts; it asks what kinds of knowledge, belief, consciousness, etc. does a computer system need in order to behave intelligently and how to build them into a computer program. Philosophers have generally taken a more abstract view and asked what are knowledge, etc. The *designer stance* is akin to Daniel Dennett's *design stance*(Dennett 1978) but not the same. The design stance looks at an existing artifact or organism in terms of what it is designed to do or has evolved to do. The designer stance considers how to design an artifact. This may necessitate giving it knowledge, beliefs, etc., and the ability to plan and execute plans.

Philosophical questions are especially relevant to AI when human-level AI is sought. However, most AI research since the 1970s is not aimed towards human-level AI but at the application of AI theories and techniques to particular problems.

I have to admit dissatisfaction with the lack of ambition displayed by most of my fellow AI researchers. Many useful and interesting programs are written without use of concepts common to AI and philosophy. For example, the language used by the Deep Blue program that defeated world chess champion Garry Kasparov cannot be used to express "I am a chess program, but consider many more irrelevant moves than a human does." and draw conclusions from it. The designers of the program did not see a need for this capability. Likewise none of the programs that competed in the DARPA Grand Challenge contest to drive a vehicle knew that it was one of 20 competing programs. The DARPA referees prevented the vehicles from seeing each other by making them pause when necessary. A more advanced contest in which one vehicle can pass another might need some awareness of "other minds".

The 1950s AI researchers did think about human-level intelligence. Alan Turing, who pioneered AI, was also the first to emphasize that AI would be realized by computer programs. Now there is more interest in human-level AI and methods to achieve it than in the last 40 years.

(Nilsson 2005) offers a criterion for telling when for human-level AI has been reached. It is that the system should be teachable to do a wide variety of jobs that humans do—in particular that it should be able to pass the examinations used to select people for these jobs, admitting that passing the exams may be possible without having adequate common sense to do the job. Nilsson is not specific about what kind of teaching is involved, and his criterion is weaker than Lenat's requirement that the system be able to learn from textbooks written for humans. I agree that this is one of the requirements for human-level AI.

(McCarthy 1996a) also discusses criteria for human-level AI, emphasizing the common sense informatic situation.

Even as the work aimed at human-level AI increases, important methodological differences between AI research and philosophical research are likely to remain. Consider the notion of belief. Philosophers consider belief in general. AI research is likely to continue with systems with very limited beliefs and build up from there. Perhaps these are top-down and bottom-up approaches.

We will discuss several of the concepts common to AI and philosophy in connection with the following example.

A policeman stops a car and says,

> "I'm giving you a ticket for reckless driving. If another car had come over the hill when you passed that BMW, there would have been a head-on collision."

Notice that the example involves a counterfactual conditional "if you had passed ..." with a non-counterfactual consequence "...reckless driving." Less obviously perhaps, a system understanding the sentence must jump into a suitable context and reason within that context, using concepts meaningful in the context. Thus a particular hypothetical head-on collision is in question, not, for example, statistics about how often a head-on collision is fatal.

The philosophy of X, where X is a science, often involves philosophers analyzing the concepts of X and commenting on what concepts are or are not likely to be coherent. AI necessarily shares many concepts with philosophy, e.g. action, consciousness, epistemology (what it is sensible to say about the world), and even free will.

This article treats the philosophy of AI, but section 6 reverses the usual course and analyzes some basic concepts of philosophy from the standpoint of AI. The philosophy of X often involves advice to practitioners of X about what they can and cannot do. Section 6 reverses the usual course and offers advice to philosophers, especially philosophers of mind. One point is that philosophical theories can make sense for us only if they don't preclude human-level artificial systems. Philosophical theories are most useful if they take the *designer stance* and offer suggestions as to what features to put in intelligent systems.

Philosophy of mind studies mind as a phenomenon and studies how thinking, knowledge, and consciousness can be related to the material world. AI is concerned with designing computer programs that think and act. This leads to some different approaches to problems considered in philosophy, and we will argue that it adds new considerations or at least different emphases that philosophers should consider. I take the opportunity of this Handbook to present some ideas and formalisms rather brashly.

Some of the formalisms, e.g. nonmonotonic reasoning and situation calculus, are heavily used in AI systems. Others have not yet been used in computer programs, but I think the problems they address will be important for human-level AI.

2 Some historical remarks

Although there were some precursors, serious AI work began in the early 1950s when it became apparent that electronics was advanced enough to do universal computation. Alan Turing recognized in (Turing 1947) that programming general purpose computers was better than building special purpose machines. This approach depended on AI researchers having access to computers, marginal in the early 50s but nearly universal by the late 1950s.[1]

The 1956 Dartmouth workshop, whose 1955 proposal introduced the term *artificial intelligence* triggered AI as a named field.[2]

My (McCarthy 1959) triggered work in logical AI, i.e. using mathematical logical languages and reasoning to represent common sense. Progress in logical AI has been continuous, but is still far from human-level.

The Ernst-Newell-Simon *General Problem Solver* (GPS) (Ernst and Newell 1969) was based on the idea that problem solving could be put in the form of starting with an initial expression and transforming it by a sequence of applications of given rules into a goal expression. Alas, this was an inadequate idea for problem solving in general.

The first chess programs were written in the 1950s and reached world champion level in the late 90s, through a combination of heuristics and faster computers. Unfortunately, the ideas adequate for champion level chess are inadequate for games like *go* that branch more than chess and which require recognition of parts of a situation.

Marvin Minsky's (Minsky 1963) summarized the ideas available in 1963.

(McCarthy and Hayes 1969) got the situation calculus formalism to a large AI audience.

Pat Hayes's (Hayes 1979) and (Hayes 1985) advanced a set of ideas that proved influential in subsequent AI research

David Marr's (Marr 1982) influenced much work in computer vision with its idea of the 2 1/2 dimensional representation.

The Stanford Artificial Intelligence Laboratory introduced the first robotic arms controlled by programs with input from TV cameras. (Moravec 1977)

[1] I began thinking about AI in 1948, but my access to computers began in 1955. This converted me to Turing's opinion.

[2] Newell and Simon, who got started first, and who had definite results to present at Dartmouth, used the term *complex information processing* for some years which didn't do justice to their own work.

described a cart with a TV camera controlled by radio from a time-shared computer.

I will not go much beyond the 1960s in describing AI research in general, because my own interests became too specialized to do the work justice.

3 Philosophical presuppositions of AI

That it should be possible to make machines as intelligent as humans involves some philosophical premises, although the possibility is probably accepted by a majority of philosophers. The way we propose to build intelligent machines makes more presuppositions, some of which are likely to be controversial.

This section is somewhat dogmatic, because it doesn't offer detailed arguments for its contentions and doesn't discuss other philosophical points of view except by way of making contrasts.

Our way is called *logical AI*, and involves expressing knowledge in a computer in logical languages and reasoning by logical inference, including nonmonotonic inference. The other main approach to AI involves studying and imitating human neurophysiology. It may also work.

Here are our candidate philosophical presuppositions of logical AI. They are most important for research aimed at human-level AI. There are a lot of them. However, much present AI is too limited in its objectives for it to be important to get the philosophy right.

objective world The world exists independently of humans. The facts of mathematics and physical science are independent of there being people to know them. Intelligent Martians and robots will need to know the same facts as humans.

A robot also needs to believe that the world exists independently of itself and that it cannot learn all about the world. Science tells us that humans evolved in a world which formerly did not contain humans. Given this, it is odd to regard the world as a human construct from sense data. It is even more odd to program a robot to regard the world as its own construct. What the robot believes about the world in general doesn't arise for the limited robots of today, because the languages they are programmed to use can't express assertions about the world in general. This limits what they can learn or can be told—

and hence what we can get them to do for us.[3]

In the example, neither the driver nor the policeman will have any problems with the existence of the objective world. Neither should a robot driver or policeman.

correspondence theory of truth A logical robot represents what it *believes* about the world by logical sentences. Some of these beliefs we build in; others come from its observations and still others by induction from its experience. Within the sentences, it uses *terms* to refer to objects in the world.

In every case, we try to design it so that what it will believe about the world is as accurate as possible, though not usually as detailed as possible. Debugging and improving the robot includes detecting false beliefs about the world and changing the way it acquires information to maximize the correspondence between what it believes and the facts of the world.

correspondence theory of reference AI also needs a *correspondence theory of reference* , i.e. that a mental structure can refer to an external object and can be judged by the accuracy of the reference. The terms the robot uses to refer to entities need to correspond to the entities so that the sentences will express facts about these entities. We have in mind both material objects and other entities, e.g. a plan or the electronic structure of the helium atom. The simple case of verification of correspondence of reference is when a robot is asked to pick up block $B3$, and it then picks up that block and not some other block.

As with science, a robot's theories are tested experimentally, but the concepts robots use are hardly ever defined in terms of experiments. Their properties are partially axiomatized, and some axioms relate terms representing concepts to objects in the world via observations.

A robot policeman would need debugging if it thought a car was going 20 mph when it was really going 75 mph. It would also need debugging

[3] Physics, chemistry, and biology have long been at a level where it more feasible to understand sensation in terms of science than to carry out the project of (Russell 1914) of constructing science in terms of sensation. The justification of common sense and scientific knowledge is in terms of the whole scientific picture of human sensation and its relation to the world rather than as a construction from sensation.

if its internal visual memory highlighted a cow when it should have highlighted a particular car.

A correspondence theory of reference will necessarily be more elaborate than a theory of truth, because terms refer to objects in the world or to objects in semantic interpretations, whereas sentences refer to truth values. Alas, real world theories of reference haven't been much studied. Cognitive scientists and allied philosophers refer to *the symbol grounding problem*, but I'm not sure what they mean.

reality and appearance The important consequence of the correspondence theory is the need to keep in mind the relation between *appearance*, the information coming through the robot's sensors, and *reality*. Only in certain simple cases, e.g. when a program plays chess with typed in moves, does the robot have sufficient access to reality for this distinction to be ignored. A physical robot that played chess by looking at the board and moving pieces would operate on two levels—the abstract level, using (say) algebraic notation for positions and moves, and a concrete level in which a piece on a square has a particular shape, location, and orientation, the latter necessary to recognize an opponent's move and to make its own move on the board. Its vision system would have to compute algebraic representations of positions from TV images.

It is an accident of evolution that unlike bats, we do not have an ultrasonic sense that would give information about the internal structure of objects.

As common sense and science tell us, the world is three dimensional, and objects usually have complex internal structures. What senses humans and animals have are accidents of evolution. We don't have immediate access to the internal structures of objects or how they are built from atoms and molecules. Our senses and reasoning tell us about objects in the world in complex ways.

Some robots react directly to their inputs without memory or inferences. It is our scientific (i.e. not philosophical) contention that these are inadequate for human-level intelligence, because a robot needs to reason about too many important entities that cannot be fully observed directly.

A robot that reasons about the acquisition of information must itself be aware of these relations. In order that a robot should not always believe what it sees with its own eyes, it must distinguish between appearance and reality.

A robot policeman would also need to be skeptical about whether what it remembered having seen (appearance) corresponded to reality.

third person point of view We ask “How does it (or he) know?”, “What does it perceive?” rather than how do I know and what do I perceive. This is compatible with correspondence theories of truth and reference. It applies to how we look at robots, but also to how we want robots to reason about the knowledge of people and other robots.

The interaction between the driver and the policeman involves each reasoning about the other’s knowledge.

science Science is substantially correct in what it tells us about the world, and scientific activity is the best way to obtain more knowledge. 20th century corrections to previous scientific knowledge mostly left the old theories as good approximations to reality. Since science separated from philosophy (say at the time of Galileo), scientific theories have been more reliable than philosophy as a source of knowledge.

The policeman typically relies on his radar, although he is unlikely to know much of the science behind it.

mind and brain The human mind is an activity of the human brain. This is a scientific proposition, supported by all the evidence science has discovered so far. However, the dualist intuition of separation between mind and body is related to the fact that it is often important to think about action without acting. Dualist theories may have some use as psychological abstractions. In the case of a programmed robot, the separation between mind and brain (program and computer) can be made quite sharp.

common sense Common sense ways of perceiving the world and common opinion are also mostly correct. When general common sense errs, it can often be corrected by science, and the results of the correction may become part of common sense if they are not too mathematical. Thus common sense has absorbed the notion of inertia. However, its mathematical generalization, the law of conservation of momentum, has made its way into the common sense of only a small fraction of people—even among the people who have taken courses in physics. People who move to asteroids will need to build conservation of momentum and even angular momentum into their intuitions.

From Socrates on, philosophers have found many inadequacies in common sense usage, e.g. common sense notions of the meanings of words. The corrections are often elaborations, making distinctions blurred in common sense usage. Unfortunately, there is no end to possible elaboration of many concepts, and the theories become very complex. However, some of the elaborations seem essential to avoid confusion in some circumstances.

Robots will need both the simplest common sense usages and to be able to tolerate elaborations when required. For this we have proposed three notions—contexts as formal objects (McCarthy 1993) and (McCarthy and Buvač 1997), *elaboration tolerance* (McCarthy 1999b), and *approximate objects.* (McCarthy 2000)[4]

science embedded in common sense Science is embedded in common sense. Galileo taught us that the distance s that a dropped body falls

[4] Hilary Putnam (Putnam 1975) discusses two notions concerning meaning proposed by previous philosophers which he finds inadequate. These are

> (I) That knowing the meaning of a term is just a matter of being in a certain "psychological state" (in the sense of "psychological state" in which states of memory and psychological dispositions are "psychological states"; no one thought that knowing the meaning of a word was a continuous state of consciousness, of course.)
>
> (II) That the meaning of a term (in the sense of "intension") determines its extension (in the sense that sameness of intension entails sameness of extension).

Suppose Putnam is right in his criticism of the general correctness of (I) and (II). His own ideas are more elaborate.

It may be convenient for a robot to work mostly in contexts within a larger context C_{phil1} in which (I) and (II) (or something even simpler) hold. However, the same robot, if it is to have human level intelligence, must be able to *transcend* C_{phil1} when it has to work in contexts to which Putnam's criticisms of the assumptions of C_{phil1} apply.

It is interesting, but perhaps not necessary for AI at first, to characterize those circumstances in which (I) and (II) are correct.

in time t is given by the formula

$$s = \frac{1}{2}gt^2.$$

To use this information, the English or Italian (or their logical equivalent) are just as essential as the formula, and common sense knowledge of the world is required to make the measurements required to use or verify the formula.

common sense expressible in mathematical logic Common sense knowledge and reasoning are expressible as logical formulas and logical reasoning. Some extensions to present mathematical logic are needed.

possibility of AI According to some philosophers' views, artificial intelligence is either a contradiction in terms (Searle 1984) or intrinsically impossible (Dreyfus 1992) or (Penrose 1994). The methodological basis of these arguments has to be wrong and not just the arguments themselves.

mental qualities treated individually AI has to treat mind in terms of components rather than regarding mind as a unit that necessarily has all the mental features that occur in humans. Thus we design some very simple systems in terms of the beliefs we want them to have and debug them by identifying erroneous beliefs. Its systematic theory allows ascribing minimal beliefs to entities as simple as thermostats, analogously to including 0 and 1 in the number system. Thus a simple thermostat can have as its set of possible beliefs only that the room is too hot or that it is too cold. It does not have to know that it is a thermostat. This led to controversy with philosophers, e.g. John Searle, who think that beliefs can only be ascribed to systems with a large set of mental qualities. (McCarthy 1979a) treats the thermostat example in detail.

rich ontology Our theories involve many kinds of entity—material objects, situations, properties as objects, contexts, propositions, individual concepts, wishes, intentions. Even when one kind A of entity can be defined in terms of others, we will often prefer to treat A separately, because we may later want to change our ideas of its relation to other entities.

AI has to consider several related concepts, where many philosophers advocate minimal ontologies. Suppose a man sees a dog. Is seeing a relation between the man and the dog or a relation between the man and an appearance of a dog? Some purport to refute calling seeing a relation between the man and the dog by pointing out that the man may actually see a hologram or picture of the dog. AI needs the relation between the man and the appearance of a dog, the relation between the man and the dog and also the relation between dogs and appearances of them. None need be regarded as most fundamental.

Both the driver and the policeman use enriched ontologies including concepts whose definition in terms of more basic concepts is unknown or even undefined. Thus both have a concept of a car not based on prior knowledge of its parts. The policeman has concepts of and names for offenses for which a ticket is appropriate and those requiring arrest.

natural kinds The entities the robot must refer to often are *rich* with properties the robot cannot know all about. The best example is a *natural kind* like a lemon. A child buying a lemon at a store knows enough properties of the lemons that occur in the stores he frequents to distinguish lemons from other fruits in that particular store. It is a convenience for the child that there isn't a continuum of fruits between lemons and oranges. Distinguishing hills from mountains gives more problems and disagreements. Experts know more properties of lemons than we laymen, but no-one knows all of them. AI systems also have to distinguish between sets of properties that suffice to recognize an object in particular kinds of situation and a general kind.

Curiously, many of the notions studied in philosophy are not natural kinds, e.g. proposition, meaning, necessity. When they are regarded as natural kinds, fruitless arguments about what they really are often take place. AI needs these notions but must be able to work with limited notions of them.

approximate entities Many common sense terms and propositions used successfully in conversation and writing cannot be given agreed-upon if-and-only-if definitions by the participants in a dialog. Examples include "x believes y", which has attracted much philosophical attention but also terms like "location(x)" which have not.

Some people have said that the use of computers requires terms to be defined precisely, but I don't agree. Many approximate entities will have to be considered by computer programs, internally and in communication. However, precision can often be achieved when terms and statements are interpreted in a context appropriate to a particular situation. In human usage, the context itself is not usually specified explicitly, and people understand each other, because the common context is implicit.

Our emphasis on the first class character of approximate entities may be new. It means that we can quantify over approximate entities and also express how an entity is approximate. (McCarthy 2000) treats approximate entities and approximate theories.

The counterfactual "If another car had come over the hill when you passed ..." is very approximate. It is adequate for communication between the driver and the policeman, but attempts by them to define it more precisely would probably not agree.

There is some overlap between the discussion of approximate entities and philosophical discussions of vagueness. However, our point is the need for approximate entities in AI.

compatibility of determinism and free will A logical robot needs to consider its choices and the consequences of them. Therefore, it must regard itself as having (and indeed has) a kind of *free will* even though it is a deterministic device. In the example, a judge might be offered the excuse that the driver couldn't drop back after he started to pass, because someone was right behind him.

(McCarthy 2005) formalizes a simple form of deterministic free will. A robot's or human's action sometimes has two stages. The first uses a non-deterministic theory, e.g. *situation calculus*, to compute a set of choices and their consequences and to evaluate the situations that result from performing the actions. The second stage chooses the action whose consequences are regarded as best. The sensation of free will is the situation at the end of the first stage. The choices are calculated, but the action isn't yet decided on or performed. This simple theory should be useful in itself but needs to be elaborated to take into account further aspects of human free will. The need is both philosophical and practical for robot design. One aspect of human free will that is probably unnecessary for robots is weakness of will.

mind-brain distinctions I'm not sure whether this point is philosophical or scientific. The mind corresponds somewhat to software, perhaps with an internal distinction between program and knowledge. Software won't do anything without hardware, but the hardware can be quite simple, e.g. a universal Turing machine or simple stored program computer. Some hardware configurations can run many different programs concurrently, i.e. there can be many minds in the same computer body. Software can also interpret other software.

Confusion about this is the basis of the Searle Chinese room fallacy (Searle 1984). The man in the hypothetical Chinese room is interpreting the software of a Chinese personality. Interpreting a program does not require having the knowledge possessed by that program. This would be obvious if people could interpret other personalities at a practical speed, but Chinese room software interpreted by an unaided human might run at 10^{-9} the speed of an actual Chinese.[5]

Most AI work does not assume so much philosophy. For example, classifying scenes and other inputs need not assume that there is any reality behind the appearances being classified. However, ignoring reality behind appearance will not lead to human-level AI, and some short term AI goals have also suffered from incorrect, philosophical presumptions, almost always implicit.

Human-level AI also has scientific presuppositions.

4 Scientific Presuppositions of AI

Some of the premises of logical AI are scientific in the sense that they are subject to scientific verification or refutation. This may also be true of some of the premises listed above as philosophical.

innate knowledge The human brain has important innate knowledge, e.g. that the world includes three dimensional objects that usually persist

[5] If Searle would settle for an interaction at the level of Joseph Weizenbaum's (Weizenbaum 1965), a person could interpret the rules without computer aid—as Weizenbaum recently informed me.

even when not observed. This knowledge was learned by evolution. The existence of innate knowledge was not settled by philosophical analysis of the concept, but is being learned by psychological experiment and theorizing. Acquiring such knowledge by learning from sense data will be quite hard but possible.

Indeed it is worthwhile to build as much knowledge as possible into our robots. The CYC project of Douglas Lenat is an attempt to put a large amount of common sense knowledge into a database.

Identifying human innate knowledge has been the subject of recent psychological research. See (Spelke 1994) and the discussion in (Pinker 1997) and the references Pinker gives. In particular, babies and dogs know innately that there are permanent objects and look for them when they go out of sight. We'd better build that into our robots, as well as other innate knowledge psychologists identify. Evolution went to a lot of trouble to acquire knowledge that we needn't require robots to learn from experience. Maybe the childhood preference for natural kind concepts is something robots should have built in.

middle out Humans deal with middle-sized objects and develop our knowledge up and down from the middle. Formal theories of the world must also start from the middle where our experience informs us. Efforts to start from the most basic concepts, e.g. to make a basic ontology, are unlikely to succeed as well as starting in the middle. The ontology must be compatible with the fact that the basic entities in one's initial ontology are not the basic entities in the world. More basic entities, e.g. electrons and quarks, are known less well than the middle entities.

logic level Allen Newell, who did not use logical AI, nevertheless proposed (Newell 1993) that there was a level of analysis of human rationality that he called the *logic level* at which humans could be regarded as *doing what they thought would achieve their goals.* Many of the systems the Carnegie-Mellon group built, e.g. SOAR, were first designed at the logic level.

universality of intelligence Achieving goals in the world requires that an agent with limited knowledge, computational ability and ability to observe use certain methods. This is independent of whether the agent is human, Martian, or machine. For example, playing chess-like games effectively requires something like alpha-beta pruning.

universal expressiveness of logic This is a proposition analogous to the Turing thesis that Turing machines are computationally universal—anything that can be computed by any machine can be computed by a Turing machine. The *expressiveness thesis* is that anything that can be expressed, can be expressed in first order logic with a suitable collection of functions and predicates.

Some elaboration of the idea is required before it will be as clear as the Turing thesis. First order logic isn't the best way of expressing all that can be expressed any more than Turing machines are the best way of expressing computations. However, with set theory, as axiomatized in first order logic, whatever can be expressed in stronger systems can apparently also be expressed in first order logic.

Gödel's completeness theorem tells us that every sentence p true in all models of a set a of sentences can be deduced. However, nonmonotonic reasoning is needed and used by humans to get consequences true in simple models. Very likely, reflection principles are also needed.

We expect these philosophical and scientific presuppositions to become more important as AI begins to tackle human-level intelligence.

5 Common sense and the common sense informatic situation

The main obstacle to getting computer programs with human-level intelligence is that we don't understand yet how to give them human level common sense. Without common sense, no amount of computer power will give human-level intelligence. Once programs have common sense, improvements in computer power and algorithm design will be directly applicable to making them more intelligent. Understanding common sense is also key to solving many philosophical problems.

The logical AI and knowledge representation communities undertake to study the world and represent common sense knowledge by logical formulas. A competing approach is based on studying the brain and how common sense knowledge is represented in synapses and other neurological structures.

CYC (Lenat 1995) is a knowledge base with several million common sense facts. Douglas Lenat (Matuszek et al. 2005) has repeatedly emphasized that a key level of common sense will be reached when programs can learn from the Worldwide Web facts about science, history, current affairs, etc. The above cited 2005 paper says

> The original promise of the CYC project—to provide a basis of real world knowledge sufficient to support the sort of learning from language of which humans are capable—has not yet been fulfilled.

Notice the implication that the lack is common sense knowledge rather than the ability to parse English. I agree.

This section is an informal summary of various aspects of common sense. The key phenomenon for both AI and philosophy is what we call the *common sense informatic situation.*

What is common sense?

Common sense is a certain collection of knowledge, reasoning abilities, and perhaps other abilities.

In (McCarthy 1959) I wrote that the computer programs that had been written up to 1958 lacked common sense. Common sense has proved to be a difficult phenomenon to understand, and the programs of 2005 also lack common sense or have common sense in *bounded informatic situations.* In the 1959 paper, I wrote "We shall therefore say that **a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows.**"

Programs with common sense à la (McCarthy 1959) are still lacking, and, moreover, the ideas of that paper are not enough. Logical deduction is insufficient, and nonmonotonic reasoning is required. Common sense knowledge is also required.

Here's what I think is a more up-to-date formulation.

A program has common sense if it has sufficient common sense knowledge of the world and suitable inference methods to infer a sufficiently wide class of reasonable consequences of anything it is told and what it already knows. Moreover, many inferences that people consider obvious are not deduced. Some are made by mental simulation and some involve nonmonotonic reasoning.

Requiring some intelligence as part of the idea of common sense gives another formulation.

A program has common sense if it can act effectively in the *common sense informatic situation*, using the available information to achieve its goals.

A program that decides what to do has certain information built in, gets other information from its inputs or observations; still other information is generated by reasoning. Thus it is in a certain *informatic situation.* If the information that has to be used has a common sense character, it will be in what we call the *common sense informatic situation.*

We need to contrast the general *common sense informatic situation* with less general *bounded informatic situations.* The latter are more familiar in science and probably in philosophy.

5.1 Bounded informatic situations

Current (2006) science and technology requires that to write a computer program in some area, construct a database, or even write a formal theory, one has to bound the set of concepts taken into account.

Present formal theories in mathematics and the physical sciences deal with *bounded informatic situations.* A scientist decides informally in advance what phenomena to take into account. For example, much celestial mechanics is done within the Newtonian gravitational theory and does not take into account possible additional effects such as outgassing from a comet or electromagnetic forces exerted by the solar wind. If more phenomena are to be considered, scientists must make new theories—and of course they do.

Likewise present AI formalisms work only in bounded informatic situations. What phenomena to take into account is decided by a person before the formal theory is constructed. With such restrictions, much of the reasoning can be monotonic, but such systems cannot reach human-level ability. For that, the machine will have to decide for itself what information is relevant, and that reasoning will inevitably be partly nonmonotonic.

One example is the simple "blocks world" much studied in AI where the position of a block x is entirely characterized by a sentence $At(x, l)$ or $On(x, y)$, where l is a location or y is another block. The language does not permit saying that one block is partly on another. Moreover, using $On(x, y)$ does not require a previous analysis of the meaning of the word "on" or the concept it represents. Only certain simple axioms are used.

This works, because within the context of the kind of simple block stacking program being built, one block is definitely on or not on another, assuming the program never makes the robot put a block in an ambiguous position. Patrick Winston extended the blocks world to allow a block to be supported by two others and discussed structures like arches. See (Winston 1977).

Another example is the MYCIN (Davis et al. 1977) expert system in which the ontology (objects considered) includes diseases, symptoms, and drugs, but not patients (there is only one), doctors or events occurring in time. Thus MYCIN cannot be told that the previous patient with the same symptoms died. See (McCarthy 1983) for more comment on MYCIN.

Systems in a bounded informatic situation are redesigned from the outside when the set of phenomena they take into account is inadequate. However, there is no-one to redesign a human from the outside, so a human has to be able to take new phenomena into account. A human-level AI system needs the same ability to take new phenomena into account.

In general a thinking human is in what we call the *common sense informatic situation*. The known facts are necessarily incomplete. [6]

5.2 The general common sense informatic situation

By the *informatic situation* of an animal, person or computer program, I mean the kinds of information available to it and the reasoning methods available to it. The *common sense informatic situation* is that of a human with ordinary abilities to observe, ordinary innate knowledge, and ordinary ability to reason, especially about the consequences of events that might occur including the consequences of actions it might take. Specialized information, like science and about human institutions such as law, can be learned and embedded in a person's common sense information. In spite of almost 50 years of effort, only modest progress has been made towards making computer systems with human-level common sense abilities. Much more progress

[6] As discussed in section 4, we live in a world of middle-sized objects which can only be partly observed. Science fiction and scientific and philosophical speculation have often indulged in the *Laplacean fantasy* of super-beings able to predict the future by knowing the positions and velocities of all the particles. That isn't the direction to speculate. More plausible super-beings would be better at using the information that is available to the senses—maybe having more and more sensitive senses, e.g. ultrasound, permitting seeing internal surfaces of objects. Nevertheless, their ability to predict the future and anticipate the consequences of actions they might choose would still be limited by chaotic processes.

has been made with specialized systems in bounded informatic situations.

No-one has a full understanding of what the common sense informatic situation is. I think understanding it is the single biggest problem for AI, and maybe for philosophy and cognitive science. However, it has at least the following features.

beliefs about actions and other events The policeman believes that one car passed another. His beliefs about the effects of events cause him to believe that if another car had come over the hill, there would have been a head-on collision.

elaboration tolerant theories The theory used by the agent is open to new facts and new phenomena. For example, the driver and the policeman could take possible fog into account, or the driver could claim that if another car had been coming he'd have seen the headlights reflected on a barn at the top of the hill. The cop's theory recommended that he reply, "Tell that to the judge."

Another example: A housewife shopping for dinner is at the butcher counter and thinks that her son coming on an airplane at that afternoon likes steak. She decides to check whether the airplane will be in on time. Suddenly a whole different area of common sense knowledge that is not part of the shopping-for-dinner script becomes relevant, i.e. the flight information number of the airline and how to get it if it isn't on her cell phone's telephone list. Section 6 has more on elaboration tolerance.

incompletely known and incompletely defined entities The objects and other entities under consideration are incompletely known and are not fully characterized by what is known about them. The real cars of the driver and the policeman are incompletely known, and the hypothetical car that might have come over the hill is quite vague. It would not be appropriate for the driver to ask the policeman "What kind of car did you have in mind?" Most of the entities considered are intrinsically not even fully defined. The hypothetical car that might have come over the hill is ill-defined, but so are the actual cars.

nonmonotonic reasoning Elaboration tolerance imposes one requirement on the logic, and this is the ability to do *nonmonotonic reasoning.* The system must reach conclusions that further facts not contradicting the original facts are can alter. For example, when a bird is mentioned, one normally concludes that it can fly. Learning that it is a penguin changes this. There are two major formalisms for doing nonmonotonic reasoning, *circumscription* and *default logic.* Also Prolog programs do nonmonotonic inference when *negation as failure* is used.

Circumscription, (McCarthy 1980), (McCarthy 1986), and (Lifschitz 1993), minimizes the extension of a predicate, keeping the extensions of some others fixed and allowing still others to be varied in achieving the minimum. Circumscription is the logical analog of the calculus of variations in mathematical analysis, but it doesn't so far have as elegant a theory. Here's a basic form of circumscription.

Let a be an axiom with the arguments p (to be minimized), z (which can be varied), and c (which is held constant). Then the circumscription of p, $Circum(a, p, z, c)$ is defined by

$$Circum[a, p, z, c] := a(p, z, c) \wedge (\forall p'\ z')(a(p', z', c) \rightarrow \neg p' < p), \qquad (1)$$

where we have the definitions

$$\begin{array}{l} p' < p \equiv p' \leq p \wedge p' \neq p, \\ \text{and} \\ p' \leq p \equiv (\forall x)(p'(x) \rightarrow p(x)). \end{array} \qquad (2)$$

Taking into account only some of the phenomena is a nonmonotonic reasoning step. It doesn't matter whether phenomena not taken into account are intentionally left out or if they are unknown to the reasoner.

While nonmonotonic reasoning is essential for both man and machine, it leads to error when an important fact is not taken into account. These are the errors most often noticed. [7]

[7] Here's an extended example from the history of science.

Starting in the middle of the 19th century, Lord Kelvin (William Thomson) undertook to set limits on the age of the earth. He had measurements of the rate of increase of temperature with depth and of the thermal conductivity of rock. He started with the assumption that the earth was originally molten and computed how long it would have taken for the earth to cool to its present temperature. He first estimated 98 million years and later reduced the estimate to 20–40 million years. This put him into conflict with geologists who already had greater estimates based on counting annual layers in sedimentary rock.

(Koons Spring 2005) contains a good discussion of various kinds of nonmonotonic reasoning.

reasoning in contexts and about contexts In the context of the Sherlock Holmes stories, Holmes is a detective and his mother's maiden name is undefined. In the context of U.S. legal history Holmes is a judge, and his mother's maiden name is Jackson. Bounded theories, usually have a fixed context.

An agent in the common sense informatic situation is often confronted with new contexts. Section 7 is devoted to information in and about contexts as well as relations between information in different contexts.

knowledge of physical objects There is increasing evidence from psychological experiments (Spelke 1994) that babies have innate knowledge of physical objects and their permanence when they go out of sight. Any common sense system should have this built in. (McCarthy 1996c), "The well-designed child" discusses what information about the world should be built into a robot.

composition of objects Consider an object composed of parts. It is convenient logically when what we knew about the parts and how they are put together enables us to determine the behavior of the compound object. Indeed this is often true in science and engineering and is often the goal of the search for a scientific theory. . Thus it is quite helpful that the properties of molecules follow from the properties of atoms and their interactions.

The common sense informatic situation is not so convenient logically. The properties of an object are often more readily available than the properties of the parts and their relations.

Kelvin's calculations were correct but gave the wrong answer, because no-one until Becquerel's discovery in 1896 knew about radioactive decay, the main source of energy that keeps the earth hot.

Kelvin's reasoning was nonmonotonic. Namely, he assumed that all the sources of energy whose existence could be inferred from his scientific knowledge were all that existed.

Nonmonotonic reasoning is necessary in science as in daily life. There can always be phenomena we don't know about. Indeed there might be another source of energy in the earth besides radioactivity.

Experience tells us that careful nonmonotonic reasoning, taking into account all the sources of information we can find and understand, usually gives good results, but we can never be as certain as we can be of purely mathematical results.

For example, a baseball has a visible and feelable surface, and we can see and feel the seams and can feel its compliance and its simplest heat transfer properties. We also know, from reading or from seeing a baseball disassembled, something about its innards. However, this knowledge of structure is less usable than the knowledge of the baseball as a whole.

The phenomenon of often knowing more about the whole than about the parts, applies to more than physical objects. It can apply to processes. The phenomenon even existed in mathematics. Euclid's geometry was a powerful logical structure, but the basic concepts were fuzzy.

knowledge of regions in space I don't know how to formulate this precisely nor do I know of comprehensive discussions in the psychological literature, but some such knowledge can be expected to be innate. Evolution has had almost 4 billion years to make it intrinsic. Knowledge of the space on the highway is common to the driver and the policeman in the example.

localization We do not expect events on the moon to influence the physical location of objects on the table. However, we can provide for the possibility that an astronomer looking through a telescope might be so startled by seeing a meteorite collide with the moon that he would fall off his chair and knock an object off the table. Distant causality is a special phenomenon. We take it into account only when we have a specific reason.

knowledge of other actors Babies distinguish faces from other objects very early. Presumably babies have some innate expectations about how other actors may respond to the baby's actions.

self reference In general the informatic situation itself is an object about which facts are known. This human capability is not used in much human reasoning, and very likely animals don't have it.

introspective knowledge This is perhaps a distinctly human characteristic, but some introspective knowledge becomes part of common sense early in childhood, at least by the age of five. By that age, a typical child can remember that it previously thought a box contained candy even when it has learned that it actually contained crayons.

counterfactuals Common sense often involves knowledge of counterfactuals and the ability to infer them from observation and to draw non-counterfactual conclusions from them. In the example, the policeman infers that he should give the driver a ticket from the counterfactual that there would have been a collision if another car had come over the hill. People learn from counterfactual experiences they would rather not have in reality.

bounded informatic situations in contexts Bounded informatic situations have an important relation to the common sense informatic situation. For example, suppose there are some blocks on a table. They are not perfect cubes and they are not precisely aligned. Nevertheless, a simple blocks world theory may be useful for planning building a tower by moving and painting blocks. The bounded theory of the simple blocks world in which the blocks are related only by the $on(x, y, s)$ relation is related to the common sense informatic situation faced by the tower builder. This relation is conveniently expressed using the theory of contexts as objects discussed in section 7 and (McCarthy and Buvač 1997). The blocks world theory holds in a subcontext *cblocks* of the common sense theory c, and sentences can be *lifted* in either direction between c and *cblocks*.

learning A child can learn facts both from experience and from being told. Quite young children can be told about Santa Claus. Unfortunately, no AI systems so far developed (2006 January) can learn facts expressed in natural language on web pages.

Closer to hand, we do not expect objects not touching or connected through intermediate objects to affect each other. Perhaps there is a lot of common sense knowledge of the physical motion of table scale objects and how they affect each other that needs to be expressed as a logical theory.

The difficulties imposed by these requirements are the reason why the goal of Leibniz, Boole and Frege to use logical calculation as the main way of deciding questions in human affairs has not yet been realized. Realizing their goal will require extensions to logic beyond those required to reason in bounded informatic situations. Computer programs operating in the common sense informatic situation also need tools beyond those that have been used so far.

In contrast to the above view, Nagel (Nagel 1961) treats common sense knowledge as the same kind of knowledge as scientific knowledge, only not systematically tested and justified. This is true of some common sense knowledge, but much common sense knowledge concerns entities that are necessarily ill-defined and knowledge about their relations that is necessarily imprecise.

Shannon's quantitative information theory seems to have little application to the common sense informatic situation. Neither does the Chaitin-Kolmogorov-Solomonoff computational theory. Neither theory concerns what common sense information is.

6 The AI of philosophy—some advice

Van Benthem (van Benthem 1990), tells us that AI is philosophy pursued by other means. That's part of what AI has to do.

AI research attacks problems common to AI and philosophy in a different way. For some philosophical questions, the AI approach is advantageous. In turn AI has already taken advantage of work in analytic philosophy and philosophical logic, and further interactions will help both kinds of endeavor. This section offers reasons why philosophers might be interested in AI approaches to some specific common problems and how AI might benefit from the interaction.

Achieving human-level common sense involves at least partial solutions to many philosophical problems, some of which are long standing in the philosophical, AI, and/or cognitive science literature, and others which have not yet been identified. Identifying these problems is important for philosophy, for AI, and for cognitive science.

To ascribe certain *beliefs*, *knowledge*, *free will*, *intentions*, *consciousness*, *abilities* or *wants* to a machine or computer program is *legitimate* when such an ascription expresses the same information about the machine that it expresses about a person. It is *useful* when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair or improve it. It is perhaps never *logically required* even for humans, but expressing reasonably briefly what is actually known about the state of a machine in a particular situation may require ascribing mental qualities or qualities isomorphic to them. Theories of belief, knowledge and wanting can be constructed for machines in a simpler setting than for humans and later applied to humans. Ascription of mental qualities is most straightforward for machines of known structure such as thermostats and computer operating systems, but is *most useful* when applied to entities whose structure is very incompletely known.

While we are quite liberal in ascribing *some* mental qualities even to rather primitive machines, we should be conservative in our criteria for ascribing any *particular* quality. The ascriptions are what (Dennett 1978) calls taking the *intentional stance.*

Even more important than ascribing mental qualities to existing machines is designing machines to have desired mental qualities.

Here are some features of some AI approaches to common problems of AI and philosophy.

AI starts small. Fortunately, AI research can often make do with small versions of the concepts. These small versions of the concepts and their relations are valid in limited contexts. We discuss three examples here and in section 7, which is about context. These are belief, action in the blocks world, and ownership of purchased objects.

An intelligent temperature control system for a building should be designed to know about the temperatures of particular rooms, the state of various valves, the occupants of rooms, etc. Because the system is not always correct about these facts, we and it should regard them as beliefs. Weather predictions need always be regarded as uncertain, i.e. as beliefs.

It is worthwhile to consider the simplest beliefs first, e.g. those of a thermostat.

A simple thermostat may have just three possible beliefs: the temperature is too cold, okay, or too hot. It behaves according to its current belief, turning the heat on, leaving it as is, or turning it off. It doesn't believe it's a thermostat or believe it believes the room is too cold.

Of course, the behavior of this simple thermostat can be understood without ascribing any beliefs. Beginning a theory of belief with such simple cases has the same advantage as including 1 in the number system. (Ascribing no beliefs to a rock is like including 0.) A temperature control system for a whole building is appropriately ascribed more elaborate beliefs. Ascribing beliefs and other mental qualities is more thoroughly discussed in (McCarthy 1979a).

A child benefits from knowing that it is one child among others. Likewise, a temperature controller might even benefit from knowing that it is one temperature controller among other such systems. If it learns via the Internet that another system adjusts to snow on the roof, it might modify its program accordingly.

Naive common sense is often right in context. An example is the common sense notion of "x caused y".

There is a context in which "The window was broken by Susan's baseball" is true and "The window was broken, because the building contractor neglected to put a grill in front of it" is not even in the language used by the principal in discussing the punishment of the girl who threw the ball. Such limited contexts are often used and useful. Their relation to more general contexts of causality require study and logical formalization.

theory of action and the frame problem The conditions for an agent achieving goals in the world are very complicated in general, but AI research has developed theories and computer programs of increasing sophistication.

AI has long (since the 1950s anyway) concerned itself with finding sequences of actions that achieve goals. For this AI needs theories of the effects of individual actions, the tree of situations arising from an initial situation, and the effects of sequences of actions. The most used AI formalism for this is the *situation calculus*[8] introduced in (McCarthy and Hayes 1969). Its relations to philosophy are discussed in (Thomason 2003). There are thorough discussions in (Shanahan 1997) and (Reiter 2001), and a new version with *occurrence* axioms as well as the usual *effect axioms* is introduced in (McCarthy 2002). Three problems, *the frame problem*, *the qualification problem*, and *the ramification problem* have arisen and are extensively discussed in the AI literature and also in (Thomason 2003). The frame problem, also taken up by

[8] The event calculus (Mueller 2006) is an alternative.

philosophers, concerns how to avoid stating which *fluents* (aspects of a situation) are unchanged when an action takes place, e.g. avoiding explicitly stating that the color of an object doesn't change when the object is moved.

The basic situation calculus is a non-deterministic (branching) theory of action. AI has also treated deterministic (linear) theories of action. The new formalism of (McCarthy 2002) permits a treatment (McCarthy 2005) of a kind of deterministic free will in which a non-deterministic theory serves as part of the deterministic computational mechanism.

AI has considered simple examples that can be subsequently elaborated. The well-known *blocks world* is treated with logical sentences like $On(Block1, Block2)$ or $On(Block1, Block2, S0)$ in which the situation is explicit. Another formalism uses $Value(Location(Block1), S0) = Top(Block2)$. We may also have

$$\begin{array}{l}(\forall s)(\ldots \rightarrow Location(block, Result(Move(block, l), s)) = l) \\ \text{and} \\ (\forall s)(\ldots \rightarrow Color(block, Result(Paint(block, c), s)) = c\end{array} \tag{3}$$

where ... stands for some preconditions for the success of the action. On one hand, such simple action models have been incorporated in programs controlling robot arms that successfully move blocks. On the other hand, the *frame problem* arose in specifying that moving a block didn't change the locations of other blocks or the colors of the blocks. This problem, along with its mates, the qualification problem and the ramification problem, arose in AI research but arise also in studying the effects of action in philosophy.

Note that in the bounded theory of the blocks world as partly described here, there is only one actor, and a block is never partly on one block and partly on another. Elaborations have been made to study these complications, but the methodology of doing the simple cases first has led to good results. Making a full theory of action from scratch is still only a vaguely defined project.

nonmonotonic reasoning Nonmonotonic reasoning is essentially the same topic as defeasible reasoning, long studied in philosophy. What's new since the 1970s is the development of formal systems for nonmonotonic reasoning, e.g. the logic of defaults (Reiter 1980) and circumscription, (McCarthy 1980) and (McCarthy 1986). There are also computer systems dating from the 1970s that do nonmonotonic reasoning, e.g. Microplanner and Prolog. Nonmonotonic reasoning has been prominent in programs that make plans to achieve goals.

Recent articles in the *Stanford Encyclopedia of Philosophy* have made the connection between AI work in nonmonotonic reasoning and philosophical work on defeasibility. Convenient references are (Thomason 2003), (Koons Spring 2005), and (Antonelli 2003).

elaboration tolerance Explicit formalizations of common sense phenomena are almost never complete. There is always more information that can be taken into account. This is independent of whether the phenomena are described in ordinary language or by logical sentences. Theories always have to be elaborated. According to how the theory is written in the first place, the theory may *tolerate* a given elaboration just by adding sentences, which usually requires nonmonotonicity in making inferences from the theory, or the theory may have to be scrapped and a new theory built from scratch. (McCarthy 1999b) introduces the concept of *elaboration tolerance* and illustrates it with 19 elaborations of the well-known missionaries and cannibals puzzle. The elaborations seem to be straightforward in English but rely on the common sense of the reader. Some of the logical formulations tolerate some of the elaborations just by adding sentences; others don't. One goal is find a logical language in which all the elaborations are additive.

(Lifschitz 2000) accomplishes 9 of the above-mentioned 19 elaborations in the Causal Calculator of McCain and Turner (McCain and Turner 1998). (Shanahan 1997) has an extensive discussion of elaboration tolerance.

I don't know of discussions of the elaboration tolerance of theories proposed in the philosophical literature.

sufficient complexity usually yields essentially unique interpretations A robot that interacts with the world in a sufficiently complex way gives rise to an essentially unique interpretation of the part of the world with which it interacts. This is an empirical, scientific proposition, but many people, especially philosophers (see (Quine 1960), (Quine 1969), (Putnam 1975), (Dennett 1971), (Dennett 1998)), seem to take its negation for granted. There are often many interpretations in the world of short descriptions, but long descriptions almost always admit at most one. As far as I can see, (Quine 1960) did not discuss the effect of a large context on the indeterminacy of translation—of say *gavagai*.

The most straightforward example is that a simple substitution cipher cryptogram of an English phrase. Thus XYZ could be decrypted as either "cat" or "dog". A simple substitution cryptogram of an English sentence usually has multiple interpretations if the text is less than 21 letters and usually has a unique interpretation if the text is longer than 21 letters. Why 21? It's a measure of the redundancy of English (Shannon 1949). The redundancy of the sequence of a person's or a robot's interactions with the world is just as real—though clearly much harder to quantify.

approximate objects and theories The idea that entities of philosophical interest are not always well defined can, if you like such attributions, be attributed to Aristotle's

> Our discussion will be adequate if it has as much clearness as the subject matter admits of, for precision is not to be sought for alike in all discussions, any more than in all the products of the crafts.
> —*Nicomachean Ethics.*

I don't know whether Aristotle pursued the idea further.

I proposed (McCarthy 2000) that AI requires the formalization of approximate entities that sometimes yields firm logical theories on foundations of semantic quicksand. Thus it is definite that Mount Everest was climbed in 1953 even though it is not definite what rock and ice constitute Mount Everest. A much more approximate concept though still useful is *"The United States wanted in 1990"* applied to "that Iraq would withdraw from Kuwait". One proposal is to use necessary conditions for a proposition and sufficient conditions but not to strive for conditions that are both necessary and sufficient. These ideas are connected to notions of vagueness that have been discussed by philosophers, but the discussion in the article (Sorensen Fall 2003) in the Stanford Encyclopedia of Philosophy does not discuss how to formalize essentially vague concepts.

contexts as objects This is an area where, judging from the Stanford Encyclopedia of Philosophy, there is as yet no connection between the rather extensive research in AI that started with (McCarthy 1993) and research in philosophy. Since *information in AI* (and in ordinary language) is always presented in a context, section 7 is devoted to a sketch of a theory of contexts as objects.

concepts as objects In natural language, concepts are discussed all the time. Nevertheless, Carnap wrote

> . . . it seems that hardly anybody proposes to use different variables for propositions and for truth-values, or different variables for individuals and individual concepts.
> ((Carnap 1956) , p. 113.

Perhaps Carnap was thinking of (Church 1951) as the exception. Instead, modal logic is used for expressing certain assertions about propositions, and individual concepts are scarcely formalized at all.

human-level AI will require the ability to express anything humans express in natural language and also to expressions statements about the expressions themselves and their semantics.

(McCarthy 1979b) proposes distinguishing propositions from truth values and individual concepts from objects in a base domain—and using different variables for them. Here are some examples of the notation. The value of $Mike$ is a person, whereas the value of $MMike$ is a concept—intended to be a concept of that Mike in this case, but that it should be is not a typographical convention. Here are some sentences of a first order language with concepts and objects.

$$\begin{array}{l}
Denot(MMike) = Mike,\\
Male(Mike),\\
Denot(MMale(MMike)),\\
Denot(HHusband(MMary)) = Mike,\\
Husband(Mary) = Mike,\\
HHusband(MMary \neq MMike,\\
(\forall x)(x \neq Husband(Mike)\\
\rightarrow \neg Exists(HHusband(MMike)).
\end{array} \tag{4}$$

The sentence $Denot(MMike) \neq Mike$ might be true under some circumstances.

The distinction between concepts and objects makes it convenient to express some assertions that simpler notations find puzzling. Thus Russell's "I thought your yacht was longer than it is" is treated in (McCarthy 1979b).

This example and others use functions from objects to concepts of them. Thus we might write $CConcept1(Cicero) = CCicero$. If we also have $Cicero = Tully$, we'll get $CConcept1(Tully) = CCicero$. While we would not ordinarily want $TTully = CCicero$, but since concepts are not characterized by the typography used to write them, this would not be a contradiction.

Some objects have standard concepts, e.g. numbers. We'd like to write $Concept1(3) = 33$, but this conflicts with decimal notation, so it is better to write $Concept1(3) = 3'3$. Consider the true sentences

$$\begin{array}{l} \neg Knew(Kepler, CComposite(NNumber(PPlanets))) \\ \text{and} \\ Knew(Kepler, CComposite(CConcept1(Number(Planets)))). \end{array} \tag{5}$$

The first says that Kepler didn't know the number of planets is composite. The second says that Kepler knew that the number, which happens to be the number of planets, is composite. See also (Maida and Shapiro 1982) and (Shapiro 1993) for another AI approach to representing concepts.

These considerations are only a small step in the direction, necessary both for AI and philosophy, of treating concepts as first class objects. (McCarthy 1997) argues the inadequacy of modal logic for a full treatment of modality. The article incited some vigorous replies.

correspondence theory of reference This is more complicated than the correspondence theory of truth, because the entities to which a term can refer are not just truth values. We recommend that philosophers study the problem of formalizing reference. There isn't even an analog of modal logic for reference.

appearance and reality Science tells us that our limited senses, and indeed any senses we might build into robots, are too limited to observe the world in full detail, i.e. at the atomic level. AI in general, and robotics in particular, must live with this fact and therefore requires a theory of the relations between appearance and reality. This theory must accomodate different levels of detail in both. I haven't got far with this, but (McCarthy 1999a) gives a small example of the relation between two-dimensional appearance and three-dimensional reality. Realist, especially materialist, philosophers also need to formalize this relationship.

consciousness, especially consciousness of self Humans have a certain amount of ability to observe and reason about their own internal states. For example, I may conclude that I have no way of knowing, short of phoning her, whether my wife is in her office at this moment. Such consciousness of one's internal state is important for achieving goals that do not themselves involve consciousness. (McCarthy 1996b) discusses what consciousness a robot will need to accomplish the tasks we give it.

7 Information in contexts and about contexts

Information is always transmitted in a context. Indeed a person thinks in a context. For the philosophy of information, information in contexts and the relations among contexts are more important than the Shannon entropy of a text.

This section discusses formalizing contexts as first class objects. The basic relation is $Ist(c, p)$. It asserts that the *proposition* p is true in the *context* c. The most important formulas relate the propositions true in different contexts. Introducing contexts as formal objects will permit axiomatizations in limited contexts to be expanded to *transcend* the original limitations. This seems necessary to provide AI programs using logic with certain capabilities that human fact representation and human reasoning possess. Fully implementing *transcendence* seems to require further extensions to mathematical logic, i.e. beyond the nonmonotonic inference methods first invented in AI and now studied as a new domain of logic.

The expression $Value(c, term)$ giving the value of the expression $term$ in the context c is just as important as $Ist(c, p)$, perhaps more important for applications.

Here are some of the features of a formalized theory of context.

1. There are many kinds of contexts, e.g. the context of Newtonian gravitation and within it the context of the trajectory of a particular spacecraft, the context of a theory formalizing the binary relations $On(x, y)$ and $\mathit{Above}(x, y)$, a situation calculus context with the ternary relations $On(x, y, s)$ and $\mathit{Above}(x, y, s)$, the context of a particular conversation or lecture, the context of a discussion of group theory in French, and the context of the Sherlock Holmes stories.

2. There must be language for expressing the value of a term in a context. For example, we have

   $$C0 : Value(Context(\mathit{ThisArticle}), \mathit{Author}) = JohnMcCarthy.$$

3. The theory must provide language for expressing the relations of contexts, e.g. that one context specializes another in time or place, that one context assumes more group theory than another, that one discusses the same subject but in a different language.

4. There must be language for expressing relations between sentences true in related contexts and also for expressing relations between terms in related contexts. When $c1$ is a specialization of $c0$, such rules are called *lifting rules.*

5. Here's an example of a lifting rule associated with databases. Suppose GE (General Electric) sells jet engines to AF (U.S. Air Force) and each organization has a database of jet engines including the price. Assume that the AF context (database) assumes that the price of an engine includes a spare parts kit, whereas the GE context prices them separately. We may have the *lifting formula*

   $$\begin{array}{l} Ist(\mathit{Outer}, Value(\mathit{AF}, Price(\mathit{engine})) = Value(GE, Price(\mathit{engine})) \\ +Value(GE, Price(\mathit{Spare\text{-}Parts\text{-}Kit}(\mathit{engine})))), \end{array}$$

   expressing in an outer context *Outer* a relation between an expression in the AF context and expressions in the GE context. Others call such formulas *bridging formulas.*

   (McCarthy 1993) has an example of lifting a general rule relating predicates $On(x, y)$ and $\mathit{Above}(x, y)$ to a situation with three argument relations $On(x, y, s)$ and $Above(x, y, s)$, in which the third argument s is a situation.

6. We envisage a reasoner that is always in a context. It can *enter* specializations and other modifications of the current context and then reason in it. Afterwards, it can *exit* the inner context, returning to the outer context. In human-level AI systems there will be no outermost context. It will always be possible to *transcend* the outermost context so far named and reason in a new context in which the previous context is an object.

(McCarthy 1993) and (McCarthy and Buvač 1998) present a more detailed theory of formalized contexts. See also (Guha 1991).

Not included in those papers is the more recent idea that what some AI researchers call "toy theories" may be valid in some contexts, and that a reasoner may do an important part of his thinking in such a limited context.

For example, consider a simple theory of buying and owning. From the point of view of a small child in a store after he has learned that he may not just take something off the shelf, he knows that it is necessary for the parent to buy something in order give it to the child. Call this context $Own0$. The details of buying are unspecified, and this simple notion may last several years. The next level of sophistication involves paying the price of the object. Not only does this notion last longer for the child, but an adult in a grocery store usually operates in this context $Own1$, which admits a straightforward situation calculus axiomatization. Outside of supermarkets, ownership becomes more complicated, e.g. buying a house with a mortgage. Certain of these ownership contexts are understood by the general public and others by lawyer and real estate investors, but no-one has a full theory of ownership.

8 Conclusions and remarks

Artificial intelligence is based on some philosophical and scientific presuppositions. The simplest forms of AI make fewer presuppositions than AI research aimed at human-level AI. The feature of human-level AI we emphasize is the ability to learn from its experience without being further programmed.

The concreteness of AI research has led to a number of discoveries that are relevant to philosophy, and these are only beginning to be noticed by philosophers. Three of the topics treated in this chapter are formalized nonmonotonic reasoning, formalized contexts, and the need to deal with concepts that have only an approximate meaning in general. Besides what's in this chapter, we particularly recommend (Thomason 2003) by Richmond Thomason.

References

Antonelli, A. 2003. Non-monotonic logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Carnap, R. 1956. *Meaning and Necessity*. University of Chicago Press.

Church, A. 1951. The need for abstract entities in semantic analysis. *Proceedings of the American Academy of Arts and Sciences* 80(1):100–112. Reprinted in *The Structure of Language*. edited by Jerry A. Fodor and Jerrold Katz, Prentice-Hall 1964.

Davis, R., B. Buchanan, and E. Shortliffe. 1977. Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence* 8(1):15–45.

Dennett, D. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: Bradford Books/MIT Press.

Dennett, D. 1998. *Brainchildren: Essays on Designing Minds*. MIT Press.

Dennett, D. C. 1971. Intentional systems. *The Journal of Philosophy* 68(4):87–106.

Dreyfus, H. 1992. *What Computers still can't Do*. M.I.T. Press.

Ernst, G. W., and A. Newell. 1969. *Gps: A CASE Study in Generality and Problem Solving*. New York: Academic Press.

Guha, R. V. 1991. *Contexts: A Formalization and Some Applications*. PhD thesis, Stanford University. Also published as technical report STAN-CS-91-1399-Thesis, MCC Technical Report Number ACT-CYC-423-91, and available as http://www-formal.stanford.edu/guha/guha.ps.

Hayes, P. J. 1985. The second naive physics manifesto. In H. J.R. and M. R.C. (Eds.), *Formal Theories of the Commonsense World*, 1–36. Ablex.

Hayes, P. J. 1979. The naive physics manifesto. In D. Michie (Ed.), *Expert systems in the microelectronic age*. Edinburgh, Scotland: Edinburgh University Press.

Koons, R. Spring 2005. Defeasible reasoning. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.*

Lenat, D. B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11).

Lifschitz, V. 1993. Circumscription[9]. In *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning.* Oxford University Press.

Lifschitz, V. 2000. Missionaries and cannibals in the causal calculator. In A. G. Cohn, F. Giunchiglia, and B. Selman (Eds.), *KR2000: Principles of Knowledge Representation and Reasoning,Proceedings of the Seventh International conference*, 85–96. Morgan-Kaufman.

Maida, A. S., and S. C. Shapiro. 1982. Intensional concepts in propositional semantic networks. *Cognitive Science* 6(4):291–330. Reprinted in R. J. Brachman and H. J. Levesque, eds. Readings in Knowledge Representation, Morgan Kaufmann, Los Altos, CA, 1985, 170-189.

Marr, D. 1982. *Vision.* New York: Freeman.

Matuszek, C., M. Witbrock, R. Kahlert, J. Cabral, D. Schneider, P. Shah, and D. Lenat. 2005. Searching for common sense: Populating cyc from the web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence,* Pittsburgh, Pennsylvania, July 2005.*

McCain, N., and H. Turner. 1998. Satisfiability planning with causal theories. In *KR*, 212–223.

McCarthy, J. 1959. Programs with Common Sense[10]. In *Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory*, 77–84, London, U.K. Her Majesty's Stationery Office. Reprinted in (McCarthy 1990).

McCarthy, J. 1979a. Ascribing mental qualities to machines[11]. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence.* Harvester Press. Reprinted in (McCarthy 1990).

[9] http://www.cs.utexas.edu/users/vl/mypapers/circumscription.ps
[10] http://www-formal.stanford.edu/jmc/mcc59.html
[11] http://www-formal.stanford.edu/jmc/ascribing.html

McCarthy, J. 1979b. First Order Theories of Individual Concepts and Propositions[12]. In D. Michie (Ed.), *Machine Intelligence*, Vol. 9. Edinburgh: Edinburgh University Press. Reprinted in (McCarthy 1990).

McCarthy, J. 1980. Circumscription—A Form of Non-Monotonic Reasoning[13]. *Artificial Intelligence* 13:27–39. Reprinted in (McCarthy 1990).

McCarthy, J. 1983. Some Expert Systems Need Common Sense[14]. In H. Pagels (Ed.), *Computer Culture: The Scientific, Intellectual and Social Impact of the Computer*, Vol. 426. Annals of the New York Academy of Sciences.

McCarthy, J. 1986. Applications of Circumscription to Formalizing Common Sense Knowledge[15]. *Artificial Intelligence* 28:89–116. Reprinted in (McCarthy 1990).

McCarthy, J. 1990. *Formalizing Common Sense: Papers by John McCarthy.* Ablex Publishing Corporation.

McCarthy, J. 1993. Notes on Formalizing Context[16]. In *IJCAI93*.

McCarthy, J. 1996a. From Here to Human-Level AI[17]. In *KR-96*, 640–646.

McCarthy, J. 1996b. Making Robots Conscious of their Mental States[18]. In S. Muggleton (Ed.), *Machine Intelligence 15.* Oxford University Press. Appeared in 2000. The web version is improved from that presented at Machine Intelligence 15 in 1995.

McCarthy, J. 1996c. The well-designed child. http://www-formal.stanford.edu/jmc/child.html.

McCarthy, J. 1997. Modality si! modal logic, no! *Studia Logica* 59(1):29–32.

[12] http://www-formal.stanford.edu/jmc/concepts.html

[13] http://www-formal.stanford.edu/jmc/circumscription.html

[14] http://www-formal.stanford.edu/jmc/someneed.html

[15] http://www-formal.stanford.edu/jmc/applications.html

[16] http://www-formal.stanford.edu/jmc/context.html

[17] http://www.formal.stanford.edu/jmc/human.html

[18] http://www-formal.stanford.edu/jmc/consciousness.html

McCarthy, J. 1999a. Appearance and reality[19]. *web only for now, and perhaps for the future*. not fully publishable on paper, because it contains an essential imbedded applet.

McCarthy, J. 1999b. Elaboration tolerance[20]. *web only for now*.

McCarthy, J. 2000. Approximate objects and approximate theories[21]. In A. G. Cohn, F. Giunchiglia, and B. Selman (Eds.), *KR2000: Principles of Knowledge Representation and Reasoning,Proceedings of the Seventh International conference*, 519–526. Morgan-Kaufman.

McCarthy, J. 2002. Actions and other events in situation calculus[22]. In B. S. A.G. Cohn, F. Giunchiglia (Ed.), *Principles of knowledge representation and reasoning: Proceedings of the eighth international conference (KR2002)*. Morgan-Kaufmann.

McCarthy, J. 2005. Simple deterministic free will. See http://www-formal.stanford.edu/jmc/freewill2.html.

McCarthy, J., and S. Buvač. 1997. Formalizing context (expanded notes). In A. Aliseda, R. v. Glabbeek, and D. Westerståhl (Eds.), *Computing Natural Language*. Center for the Study of Language and Information, Stanford University.

McCarthy, J., and S. Buvač. 1998. Formalizing Context (Expanded Notes). In A. Aliseda, R. v. Glabbeek, and D. Westerståhl (Eds.), *Computing Natural Language*, Vol. 81 of *CSLI Lecture Notes*, 13–50. Center for the Study of Language and Information, Stanford University.

McCarthy, J., and P. J. Hayes. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence[23]. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*, 463–502. Edinburgh University Press. Reprinted in (McCarthy 1990).

[19] http://www-formal.stanford.edu/jmc/appearance.html

[20] http://www-formal.stanford.edu/jmc/elaboration.html

[21] http://www.formal.stanford.edu/jmc/approximate.html

[22] http://www-formal.stanford.edu/jmc/sitcalc.html

[23] http://www-formal.stanford.edu/jmc/mcchay69.html

Minsky, M. L. 1963. Steps towards artificial intelligence. In E. A. Feigenbaum and J. Feldman (Eds.), *Computers and Thought*, 406–450. McGraw-Hill. Originally published in *Proceedings of the Institute of Radio Engineers*, January, 1961 **49:**8–30.

Moravec, H. P. 1977. Towards automatic visual obstacle avoidance. In *IJCAI*, 584.

Mueller, E. T. 2006. *Common Sense Reasoning*. Morgan Kaufmann.

Nagel, E. 1961. *The structure of science*. Harcourt, Brace, and World.

Newell, A. 1993. Reflections on the knowledge level. *Artificial Intelligence* 59(1-2):31–38.

Nilsson, N. J. 2005. Human-level AI? be serious! *The AI Magazine* 26(4):68–75.

Penrose, R. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Pinker, S. 1997. *How the Mind Works*. Norton.

Putnam, H. 1975. The meaning of "meaning". In K. Gunderson (Ed.), *Language, Mind and Knowledge*, Vol. VII of *Minnesota Studies in the Philosophy of Science*, 131–193. University of Minnesota Press.

Quine, W. V. O. 1969. Propositional objects. In *Ontological Relativity and other Essays*. Columbia University Press, New York.

Quine, W. v. 1960. *Word and Object*. MIT Press.

Reiter, R. 1980. A Logic for Default Reasoning[24] *Artificial Intelligence* 13 (1–2):81–132.

Reiter, R. 2001. *Knowledge in Action*. M.I.T. Press.

Russell, B. 1914. *Our knowledge of the external worrld*. Open Court.

Searle, J. R. 1984. *Minds, Brains, and Science*. Cambridge, Mass.: Harvard University Press.

[24] .

Shanahan, M. 1997. *Solving the Frame Problem, a mathematical investigation of the common sense law of inertia.* M.I.T. Press.

Shannon, C. 1949. Communication theory of secrecy systems. *Bell System Technical Journal* 28:656–715. http://www.cs.ucla.edu/ jkong/research/security/shannon.html.

Shapiro, S. C. 1993. Belief spaces as sets of propositions. *Journal of Experimental and Theoretical Artificial Intelligence* 5:225–235. http://www.cse.buffalo.edu/tech-reports/SNeRG-175.ps.

Sorensen, R. Fall 2003. Vagueness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.*

Spelke, E. 1994. Initial knowlege: six suggestions. *Cognition* 50:431–445. http://www.wjh.harvard.edu/ lds/pdfs/Spelke1994.pdf.

Thomason, R. 2003. Logic and artificial intelligence. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/archives/fall2003/entries/logic-ai/.

Turing, A. M. 1947. Lecture to the london mathematical society. In *The Collected Works of A. M. Turing*, Vol. Mechanical Intelligence. North-Holland. This was apparently the first public introduction of AI, typescript in the King's College archive, the book is 1992.

van Benthem, J. 1990. Kunstmatige intelligentie: Een voortzetting van de filosofie met andere middelen. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte* 82:83–100.

Weizenbaum, J. 1965. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery* 9(1):36–45.

Winston, P. H. 1977. *Artificial Intelligence.* Reading, Mass.: Addison Wesley Publishing Co.