An Overview Of Some Unsolved Problems In Artificial Intelligence

Aaron Sloman
Cognitive Studies Programme
School of Social Sciences
University of Sussex
At University of Birmingham since 1991
http://www.cs.bham.ac.uk/~axs

This was first published in the Proceedings of Aslib Informatics 7
Cambridge, March 1983

Introduction

It is rash for the first speaker at a conference to offer to talk about unsolved problems: the risk is that subsequent papers will present solutions. To minimise this risk, I resolved to discuss only some of the really hard long term problems. Consequently, I'll have little to say about solutions!

These long-term problems are concerned with the aim of designing really intelligent systems. Of course, it is possible to quibble endlessly about the definition of 'intelligent', and to argue about whether machines will ever really be intelligent, conscious, creative, etc. I want to by-pass such semantic debates by indicating what I understand by the aim of designing intellingent machines. I shall present a list of criteria which I believe are implicitly assumed by many workers in Artificial Intelligence to define their long term aims. Whether these criteria correspond exactly to what the word 'intelligent' means in ordinary language is an interesting empirical question, but is not my present concern. Moreover, it is debatable whether we *should* attempt to make machines which meet these criteria, but for present purposes I shall take it for granted that this is a worthwhile enterprise, and address some issues about the nature of the enterprise.

Finally, it is not obvious that it is *possible* to make artefacts meeting these criteria. For now I shall ignore all attempts to prove that the goal us unattainable. Whether it is attainable or not, the process of attempting to design machines with these capabilities will teach us a great deal, even if we achieve only partial successes.

Behavioural criteria for intelligence

The following is a list of conditions which may one day be satisfied by computing systems, and which are already satisfied by human beings. Even if they do not constitute intelligence they are very closely related to it.

(1) Having a *general* range of abilities, including
(a) the ability to cope with varied objects in a domain
(b) the ability to cope with a variety of domains of objects
(c) the ability to perform a variety of tasks in relation to any object

The term 'object' here covers such diverse things as physical objects, spoken or written sentences, stories, images, scenes, mathematical problems, social situations, programs, etc. 'Coping' includes such diverse activities as perceiving, interpreting, producing, using, acting in relation to, predicting, etc. In many cases, coping requires the ability to discern the *fine structure* of things, e.g. what the parts are and how they are related. Generality of the kinds listed here requires many sub-abilities. It is not enough that a machine should have all these abilities: it must also have the ability to decide which subset is appropriate in any given situation or for a given task.

(2) Various forms of discovery, learning, or self-improvement, including: qualitative extensions to new domains, new kinds of abilities, and quantitative improvements in speed of performance, complexity of tasks managed, etc. Important special cases include the discovering new concepts, including discerning resemblances or analogies, discovering heuristics or generalisations within a domain, creating new domains, and combining information about several different domains to solve a new class of problems. The more complex examples overlap with what we ordinarily refer to as 'creativity'.

(3) Making inferences, including not only logical deductions but also reasoning under conditions of uncertainty, and making use of default assumptions which may be cancelled by new information, reasoning with non-logical representations e.g. maps, diagrams, networks.

(4) Being able to communicate and co-operate with other intelligent systems, especially human beings. (Most of the other abilities, and this one, also seem to require the ability to communicate with oneself.)

(5) Being able to co-ordinate and control a variety of sensors and manipulators in achieving a task involving physical movement or manipulation.

(6) Coping flexibly with an environment which is not only complex and messy, but also partly unpredictable, partly friendly, partly unfriendly and often fast moving. This includes the ability to interrupt actions and abandon or modify plans when necessary, e.g. to grasp new opportunities or avoid new dangers. The need for speed implies that it will often not be possible to collect all relevant information, or to do all the analysis or inference that might theoretically be required for reliable decision making. In other words, an intelligent agent in a fast moving world is bound to be fallible.

(7) Self-awareness, including the ability to reflect on and communicate about at least some of one's own internal processes. Some form of self awareness is required even for rudimentary actions and planning: one needs to know where one is in order to work out where to go. Awareness of internal states is required for self-debugging and the ability to explain and justify one's decisions -- which expert systems are increasingly expected to do.

(8) Coping with a multiplicity of "motivators", i.e. goals, general principles, preferences, constraints, etc. which may not all be totally consistent in all possible circumstances. This need can arise either because a single high-level goal can generate a set of inter-related sub-goals, or because a system has a collection of independent sources of goals, requirements, etc.

(9) Curiosity, and exploration. The previous criteria are highly functional, or task-oriented. It is arguable that a really intelligent system should deploy many of its abilities even when there is no specific task at hand: undirected exploration can yield results which will be useful at some later time. Thus an intelligent system should engage in various types of explorations, investigations, comparisons, attempts at simplifying what has previously been learnt, etc. even when these do not serve any specific purpose.

(10) The ability to generate, or appreciate, aesthetic objects. This is often thought of as distinct from cognitive abilities, but there are reasons for thinking that aesthetic processes are involved in many cognitive processes, and vice-versa. E.g. elegant proofs not only give pleasure: they generally provide more insight than messy ones.

The notion of intelligence is bound up not only with *what* can be done, but also with *how* it is done (i.e. the style, or manner). For example:

(1) When confronted with messy, ill-defined problems and situations, and incomplete or uncertain information; an intelligent system should degrade *gracefully* as the degree of difficulty/complexity/noise/incompleteness etc. increases, rather than merely 'crashing', or rejecting the problem. Degrading gracefully may involve being slower, less reliable, less general, less accurate, or producing less precise or complete descriptions etc. For example when the book requested is not available, see if you can offer an alternative, or give some indication of where it may be obtained.

(2) Using insight and understanding rather than brute force or blind and mechanical execution of rules, to solve problems, achieve goals, etc. E.g. instead of exhaustive trial and error searching there should be selection of alternatives based on some analysis and description of the current state of a problem-solving process. This is closely connected with a requirement for speed and generality.

(3) Plans should not be created simply by applying pre-defined rules for combining primitive actions to achieve some goal, but should rely on the ability to use inference to answer hypothetical questions about 'What would happen if..'. This should also play a role in the ability to make predictions, or test generalisations.

(4) Conflicting goals should not be dealt with simply by means of a pre-assigned set of priority measures, but for example by analysing the reasons for the conflict and making inferences about the consequences of alternative choices or compromises.

Clearly not all these conditions are *necessary* for intelligence. For instance an intelligent thinker does not need to have sense organs and motors. The list is not offered as a definition of 'intelligence'. It merely summarises salient aspects of the most intelligent systems we already know, namely human beings, though many aspects can also be found in other animals. It also summarises kinds of AI research already being pursued in a more or less fragmentary fashion.

I believe these criteria are also relevant to understanding the evolution of intelligence in biological organisms.

No existing AI system fulfils even a subset of the criteria, except in very restricted domains, with rather generous interpretations of concepts like 'generality', 'graceful degradation', 'flexibly', etc. Nevertheless there are many examples of fragmentary progress.

Intelligent retrieval?

Since this is a conference on intelligent retrieval, I should point out the implications of these criteria for the task of designing intelligent retrieval systems.

Making an information retrieval system intelligent according to the criteria listed above will be a very long term task.

Not all the criteria will be relevant to an intelligent retrieval system. For instance, such a system, unlike a robot, need not be encased in a mobile body, and it will not generally be embedded in a potentially hostile and fast-moving environment except in military "command and control" systems, air traffic control systems, and the like. Even in the case of ordinary document retrieval there may be times when some information is needed very urgently, yet establishing exactly where it is might take a long time. In such a case the system should be able to make an informed guess, just as robots acting in the physical world may have to be able to make informed guesses about things they see or feel, when there isn't time to collect more information or to analyse it in detail. Making guesses, whether informed or not, is liable to produce errors. This helps us understand why it is inevitable that intelligent systems will be fallible.

A really intelligent retrieval system might use its spare time to look for new relationships between the items stored, and perhaps devise new and better classification schemes. This

will require far more than textual information. The system would need to be able to understand stored texts, and this will require considerable world knowledge and the ability to make use of it creatively. Thus, an intelligent retrieval system would have to be an intelligent system, period.

This may be a very general point. That is: there may be no such thing as a really intelligent system which performs only one kind of activity, in relation to one kind of information. For intelligence of the kind we are talking about requires the ability to acquire new knowledge by relating different abilities, or different kinds of information.

Steps towards intelligent systems.

Achieving artificial forms of intelligence as defined above is a long way off. Many problems remain to be solved, both long term and short term. A lot of the short term problems are already being addressed, and will yield in due course to continued efforts. Many of them are of the forms:

> What sort of knowledge is required for X?
>
> What algorithms make it possible to use the knowledge, i.e. to achieve X?
>
> How can X be done with reasonable space and time requirements? (This is important because often there is a way of solving problems by using exhaustive combinatorial search - but which is totally impracticable because of astronomical space or time requirements.)
>
> How can X be done intelligently, e.g. without every possible detail having to be specified by the programmer, and with performance degrading gracefully as tasks become more difficult?

For X read any of the following: understanding and generating or translating English, French, Urdu,... interpreting visual images, manipulating physical objects, storing and retrieving sentences, images, rules, etc., diagnosing diseases, recommending medical treatment, diagnosing electronic faults, designing circuits, designing programs, deciding on a computer configuration required to suit a particular customer, finding books or articles relevant to some problem, solving mathematical problems, playing chess, being a good teacher of ...., interpreting results of experiments.

All of these are tasks which in their simpler forms can already be or will soon be

performed by suitably programmed computers. The recent explosion of interest in "Expert Systems" is likely to produce many examples of useful, though relatively simple programs.

The shock -- and possibly outrage at unfulfilled promises -- may come when it is found that the techniques and representations do not readily generalise to more complex cases. I'll try to indicate where some of the long term problems lie.

The problem of codifying useable knowledge

Designing intelligent computing systems involves giving them knowledge. What knowledge to give them is often far from obvious. Much of the effort involved in the new technology of 'Knowledge Engineering' (see Michie [1979] for a survey) is concerned with making knowledge explicit, so that it can be stored in computers.

Some kinds of knowledge are hard to get at even though they are shared by all sorts of people including children and adults who are not regarded as particularly intelligent. Some kinds, e.g. the knowledge required for visual and manipulative skills, and for planning actions, are even to be found in other animals. E.g. many mammals and birds have excellent vision and can manipulate physical objects in far more sophisticated and flexible ways than existing robots. No robot that I know of can begin to match the visual and manipulative skills displayed when a bird builds a nest.

Much of this common knowledge is totally inaccessible to introspection, and very hard to get at by doing experiments, including cutting open brains, or implanting electrodes. The difficulty is in part like the difficulty of finding out how a very complex computer program works (e.g. an operating system) merely by interacting with it, or by opening up the computer and measuring electronic processes.

Can we bypass the problems by getting machines to learn things for themselves? This assumes that we can provide the machines with the basic representational powers required in order to represent the results of such learning. As I shall show below, we have not yet surveyed all the important modes of representation relevant to the design of intelligent systems.

If we want the machines not only to work out the specific knowledge required for a variety of tasks, but also to invent good formalisms for expressing different kinds of knowledge, then we may have to wait a long time. It took evolution millions of years even to produce bird-brained intelligence. Can computers significantly speed up this trial and error search if left to invent things for themselves? Individual animals can learn much faster, but that is presumably because they don't start off completely devoid of representations, knowledge, strategies. They have the benefits of all those millions of years of evolution even when they are only a few minutes old. But what this inherited information is, and how it is embodied in a learning system remains an unsolved problem.

More general problems.

But it is one thing to say what intelligent systems should be able to do and quite another to say *how* they should be designed. Our search for mechanisms is hampered by lack of answers to some very general and difficult questions about machines, representations and architectures, namely:

> What classes of machines are there, and what are they good for, bad for?
>
> What classes of representations are there, and what are they good for, bad for?
>
> What classes of system architectures are there, and what are they good for, bad for?

These are technical questions requiring investigations of types which have already begun to emerge from Computing Science. I'll say more about them below.

There is another class of questions more closely connected with philosophical problems, which will need to be addressed in the long run, e.g.:

> What are the requirements for true mentality: consciousness, intentionality, understanding? Is more required than is needed for the behavioural manifestations of intelligence?

I'll have very little to say about these.

At the end, I'll mention another difficult problem, a meta-problem concerning the attempt to list important unsolved problems.

What classes of machines?

We are beginning to understand some of the variety of possible machines. Computer Science includes the study of Turing machines, Von Neumann machines, machines with and without stacks, networks of machines doing things in parallel (e.g. see Treleaven's paper presented to this conference). There has also been a revival of interest in machines modelled in part on the apparent structure of neural nets in animal brains. E.g. see the volume edited by Hinton and Anderson, and recent work by Hinton, Feldman, Ballard and others reported in *7th International Joint Conference in Artificial Intelligence* Vancouver 1981, and in the journal *Cognitive Science* in 1982.

This work has begun to show how some perceptual problems which on conventional machines generate a horrendously slow search for consistent interpretations, may be solved quickly on highly parallel highly-connected machines. The combinatorial explosion reappears in space instead of time. Very large numbers of parallel units with very large numbers of connections seem to be required. Current research on such 'connection'

machines is concerned with ways of re-structuring problems to tame the combinatorial explosion - but if the human brain is taken as a guide, we can expect that very large numbers of units will be required, unless some entirely new techniques turn up.

It should not be assumed that the units out of which highly parallel intelligent systems need to be constructed will be at all like existing digital computers with programs and data stored in a memory accessed by a central processor. Neural nets don't seem to be built up of such units. At the lowest level they don't manipulate symbols but states of activity of individual processing units, and weights, thresholds, etc. Behaviour is probabilistic, instead of deterministic like modern computers.

Units do not transmit symbols to one another, but rather excite or inhibit one another, i.e. change levels of activity, via links whose influence seems to be variable - i.e. the links are 'weighted' and the weights can change over time.

The main use of such machines seems to be to defeat combinatorial explosions in time. All rival interpretations or solutions coexist and compete via constraints represented by the links. The best cluster wins *quickly,* by elements exciting each other and inhibiting elements of rival clusters, through feedback loops. Where there are rival clusters of approximately equal merit, the behaviour of such a machine can be sensitive to small random perturbations, or to the initial state in which it begins to work on a problem.

Little is understood about how to use or program such machines, or how relatively abstract concepts, knowledge and inference rules can be represented in terms of patterns of activation and inhibition of the units. It is also not clear what kinds of manufacturing methods are capable of producing very large numbers of processing units (e.g. many millions) each connected to a large number of other units. The current technology can produce two-dimensional arrays of processors, whereas rich 3-D connectivity can at present only be found in biological systems.

Probably some of the tasks of intelligent systems will not be performed directly at this level. Instead a different sort of virtual machine will need to be implemented on top of connection machines, or perhaps alongside them.

All that is clear at present is that there is a very large space of possible machines and we have only begun to explore a small subspace. That a certain class of machines can support particular aspects of intelligent behaviour does not mean that they are suitable for designing systems meeting the full set of criteria listed above. We don't know whether important entirely new kinds of physical machine organisation are waiting to be discovered.

What sorts of representations?

Besides exploring types of machines it is important, perhaps more important, to understand the sorts of representations that can be used by intelligent systems. Concepts, knowledge, strategies, goals must somehow be represented. Thinking, planning, deciding, learning, perceiving all involve the construction and manipulation of internal representations. Almost certainly different sorts of representations are required for different purposes, just as we need both maps and verbal descriptions. So far the following sorts of representations have been explored in AI and Computer Science.

Declarative:

Quantitative measures

Data-structures e.g. semantic nets

Logical assertions

Analogical representations: maps, lists, image arrays

Procedural

feedback loops

programs: sequential and parallel

Activation/inihibition/relaxation networks

logical inference

Although debates about which sorts of representation to use can get quite heated, we should remember that as with machines, the space of possible formalisms or notations is very large, and we have only begun to understand a few subspaces. In particular, we do not know what forms of representation can be implemented fruitfully using ’connection’ machines, though it is clear that very different representations are possible from those currently understood.

Does spatial reasoning underly other forms?

The representation of spatial structures, properties, relations probably needs to be understood before we can do other things really well. (Animals with superb spatial abilities evolved before logical/mathematical thinkers, or speakers. So perhaps spatial abilities are pre-requisites for the others?).

It may turn out that good spatial reasoning abilities are bound up with the use of analogical

representations i.e. representations in which relationships between parts of a representation represent relations between things denoted - unlike logic, where relationships are explicitly named.

It is also possible that human and animal spatial abilities depend on the use of highly parallel connection machines.

One piece of evidence that more abstract forms of intelligence may depend on the ability to represent and manipulate spatial structures is that we frequently use spatial metaphors in discussing more abstract domains. For instance, we talk about family trees, networks of relationships, and the exploration of a search space. However, none of this is conclusive, and even if the conjecture is correct it is of little use until we know how to give machines new spatial abilities.

What sorts of system architectures?

Besides questions about what can be done with particular forms of physical hardware there are also many unanswered questions about the logical organisation of intelligent systems. For example, what sorts of major components are required and how should they be related to one another? How should overall control of the system be organised? If different sub-systems have different but incompatible needs how should the conflict be resolved?

It is not always realised that the type of overall system architecture required is not determined solely by the external criteria. In particular, limitations of processing speeds and memory may make certain choices. For example, given unlimited speed and storage a chess playing program could be written with a very simple structure, since all it would have to do at every stage is exhaustively search the game tree. But since that is not possible, a far more complex organisation is required, including provision for learning of new heuristics, mechanisms for recognising patterns, strategies for choosing relevant heuristics, etc. If neurones are the basic processing units in the human brain, then since their operations are relatively slow this will constrain possible architectures, given the need for decisions often to be taken very quickly. For tasks which can be dealt with easily by decomposition into large number of relatively independent processes a neural net is a very suitable architecture. Where the task essentially requires deep exploration of the consequences of some set of assumptions (for instance playing chess, or proving theorems in number theory), then a different organisation is required, and it will probably be necessary to make heavy use of large numbers of not necessarily completely reliable heuristics, which in turn will have to be learnt through trial and error. We don’t yet know which sorts of tasks intelligent systems have to perform are suitable for which sorts of architectures.

A very simple sort of system architecture is used by many existing ’expert systems’. They often consist of a database, in which a collection of facts and rules is stored, and an interpreter or ’inference engine’ which works out implications of the facts and rules, possibly also interacting with a user.

It is clear that a much more complex system organisation is going to be required for intelligent systems meeting the criteria listed above. For example the requirement to be able to monitor the environment for unexpected dangers and opportunities implies a need for quite sophisticated perceptual processes to run in parallel with plan execution processes. Moreover, if the system has not just one top-level goal but a whole collection of motivators, and the ability to generate new motives in new situations, then it will need an architecture capable of dealing with conflicts between different goals, desires, principles, etc. These conflicts can occur in connection with different sub-processes. For instance, the process of selecting which goals or purposes to pursue, the process of deciding priorities among accepted goals, the process of reacting to new information relevant to one high priority goal whilst in the course of attempting to achieve another. (Further analysis of these problems shows that the mechanisms required for really intelligent systems will also make them susceptible to emotional states.)

Besides the need to explore global system organisations, we need also to find out what sorts of organisations are possible for many of the subsystems. For example AI researchers are already investigating different kinds of language understanding systems, and several different kinds of image interpreting systems.

The space of possible computational architectures is infinite, and we therefore need to find a good way of imposing some structure which will help us understand which architectures are good for which purposes.

The problem of mind

Puzzle: in a digital computer it is easy to see how *meaning* could emerge. The elements are already there at the lowest level: instructions, addresses (reference), conditionals and booleans, arithmetic, counting operations, etc.

In a neural net machine there doesn’t seem to be the right set of building blocks - only high and low levels of activation. Quantities don’t seem to be able to support meaning, inference, deciding.

So should we infer that only computers, not brains, can understand, refer, reason???

No such conclusion follows: at the lowest level computers also have the wrong sorts of building blocks - namely sub atomic particles and they are no more suitable than neurones are to support intelligence. The moral is the familiar point that we need to think in terms of different sorts of "virtual machines" at different levels co-existing in one physical system.

Even so, suppose we manage somehow to design a machine which passes all the *behavioural* tests for intelligence listed previously: would it *really* be conscious, aware, intelligent? My own view is that such questions cannot be answered by rational discussion, since they are ultimately ethical questions, not factual ones. But I could be wrong.

However, if I am right, it follows that if we do ever produce really intelligent machines, then we shall have to face up to the ethical questions about the rights of such machines.

But there is plenty of time yet!

In any case, arguing about which sorts of machines would be 'really intelligent' seems to be a much less valuable activity than trying to understand exactly what sorts of machines can be built and what the implications are of different sorts of hardware, different sorts of representation, and different sorts of system architecture.

Conclusion

I have rambled at a rather high level of abstraction over some very general unsolved problems about the space of possible machines, the space of possible representations, the space of possible system architectures and the general requirements for intelligence. Much theoretical analysis, and computational experimentation remains to be done.

The meta-problem

I said earlier that I would end with a meta-problem, a problem about the attempt to list unsolved problems. Here it is:

Do we really know what the *important* unsolved problems are?

Do we even have the right concepts to formulate them?

Could Newton have listed the important unsolved problems of physics?

It may be that attempting to list important unsolved problems is misguided because the central unsolved problem is to identify the important unsolved problems and this would require prophetic insight. My defence is that so long as *most* people put *most* of their effort into tackling the relatively short-term problems, it can do no real harm if some of us occasionally step back and try to see where we are going, and assess our progress.

References

G. Hinton, J. Anderson (eds) *Parallel Models of Associative Memory* Erlbaum 1981.

D. Michie *Expert Systems in the Microelectronic Age,* Edinburgh University Press, 1979.