Robotson Crusoe – or – What Is Common Sense?

Don Perlis

Computer Science Dept, University of Maryland, College Park MD 20742 USA
perlis@cs.umd.edu

Abstract

I will present a perspective on human-level commonsense behavior (HLCSB) that differs from commonsense reasoning (CSR) as the latter is often characterized in AI. I will argue that HLCSB is not far beyond the reach of current technology, and that it also provides solutions to some of the problems that plague CSR, most notably the brittleness problem. A key is the judicious use of metacognitive monitoring and control, especially in the area of automated learning.

Introduction

Commonsense reasoning (CSR) is a central area of research within AI. Indeed, it might qualify as the original area, at least in the sense that the Dartmouth Conference featured human-level reasoning as a principal goal. Moreover, logic, in one form or another, was seen as a major tool in this endeavor, and has remained so ever since.

Here I wish to challenge several aspects of this paradigm (some of these I have raised before, so the present essay is an attempt to pull together all my objections at once – but also with positive suggestions for antidotes). These fall into three categories:

a. aims
b. logic
c. domains

I will discuss these in turn. In the process I will also describe a different paradigm that addresses these objections, and indicate some successes it has had so far.

Aims

Human-level AI is the study of how to design an artifact that behaves like a human, at least in regard to intelligence: the ability to reason, to imagine and pursue alternatives, solve problems, get things done efficiently, adapt to changing circumstances, improve over time, avoid disaster, survive and thrive.

How much of this is common sense? In order not to turn this into a sterile question of semantics, let me rephrase it this way: what is it that we do when we are performing effectively and yet are not exercising expertise? That is how I often describe common sense as a technical area to outsiders: the sort of thinking that lets us get things done effectively even without training. We do train to do many things: ride a bicycle, play tennis, solve calculus problems, eat with a fork, read, write, practice a profession, etc. But oddly – as was learned some decades ago in the AI enterprise – it is far easier to train automated "expert" systems to perform many (though not all) such feats than to perform what appear far simpler activities: to see, to converse, to infer that your car's flat tire may occasion a change in your plans to fly to California, etc. So, these latter and their ilk by default could be called "commonsense behaviors". But what do they have in common, other than not being the result of deliberate training?

To forestall a counter-objection: To be sure, vision, language, and plan adaptation all involve complex skills (probably a mix of nurture and nature). But these also are major sources of surprises, of things going wrong (we mistake what we are seeing, misunderstand an utterance, find our plans askew) and – most of the time – we resolve the situation (we reinterpret what we see or hear, or we change plans) without any fuss, indeed with such great ease that we often barely notice there was any surprise at all. That is, these finely-honed subsystems (vision, language, planning, etc) often encounter unanticipated situations and yet most of the time we deal with them smoothly and effectively. (Examples of various such situations and experiments with computational models of resolving them – e.g., reinforcement learning, HTN planning, contradiction handling in NLP – are summarized in (Anderson et al, 2005,2008; Perlis, 2010); we will mention some of the underlying design issues below.)

As indicated, one characterization of common sense goes by the name commonsense reasoning (CSR); it is to a large extent concerned with solving puzzles. The mutilated checkerboard, the three wise-men, missionaries and cannibals, monkey and bananas, Yale shooting problem, are some famous examples. True, some of these are puzzles that challenge humans, and others are intuitively obvious and it is the formal treatment that puzzles. Some people are skilled at puzzle-solving and some less so. But that skill does not seem to correlate with our ability to negotiate well with the world on a daily basis. So there is some other ability involved, in just getting on with the needs of everyday life where things often go awry in ways we are not already trained to handle.

Thus I wish to call attention to the particular ability to deal with a situation one is not expecting or prepared for, yet to deal with it effectively even so – what the British call "muddling through". Is this a single ability (or closely connected set of abilities), or is it an evolutionary hodge-podge with no particular concise characteristics that we can come to understand and even use in our artifacts?

I argue that it is the former, and not very complicated. Indeed it is, I suspect what allows us to contemplate – let alone work on (or give up on) – the above puzzles. Give up on? Yes! Recognizing that one is in over one's head and that it is a better use of time to give up than continue in folly, is a mark of common sense.

This may not sound like much. But I think it is the germ of a different and powerful approach to human-level AI: the ability to notice something is amiss; to assess it in terms of risk and benefit and any known available responses; to choose and enact one or more such responses; and to monitor their success. Common among such responses, of course, are these: giving up, asking for help, trial and error, thorough diagnostic assessment, redefining the situation in terms of higher-level goals, and –last but not least (and this is not intended as a comprehensive list) – initiating a course of training to acquire a perceived lack of expertise. It is important that training is included here – that is, the realization that lack of some skill is getting in the way, and that it can be rectified. So the capacity to recognize the usefulness of expertise (the lacked skill) and to undertake steps to get it, is itself a deeply powerful commonsense ability distinct from that expertise itself.

My group has been making efforts in this direction – investigating what we call the metacognitive loop (MCL) – for some years, and with successes reported in a variety of venues (Anderson and Perlis 2005; Anderson et al 2008; Perlis 2010). What is especially exciting is that as we explore new domains or more complex anomalies, the "available responses" needed do not seem to grow significantly in number or complexity. To give a perhaps overly simple gloss: giving up or asking for help are almost always options, and not any harder (maybe easier!) when the problem is harder.

In essence, it is the ability to step back and assess the situation in high-level terms that then allows a decision as to what to do. Carrying out the chosen action *can* be complicated and time-consuming, but that is another story. For instance, one might decide to learn French, instead of constantly having to struggle with an interpreter to be understood; the decision may be easy, but the follow-on learning may not.

In any event, this is the direction of our MCL work. When suitably integrated with an existing automated system, MCL endows the resulting symbiot with the ability to make decisions as to whether, when, and how to attempt to improve itself. To borrow a stock phrase: fool me once, shame on you; fool me twice, shame on me. Such an agent then can be fooled (by other agents, or simply by the complexity or the world) but sooner or later catches on and tries to do something about it.

My purpose here however is to propose a reconception of *human-level common sense behavior* (HLCSB) – rather distinct from traditional CSR as often understood – that this work seems to suggest. Namely, that HLCSB involves a *concise but powerful set of general repair strategies that allow an agent to get better at what it needs to do by means of assessing how it is doing and what options it has for possible improvement*.

Here then is a vision for future work: design, implement, and test a robot in the guise of a modern-day Robinson Crusoe: this robot – let me call it "Robotson C" – will find itself in circumstances not quite what it expects, and will have to adjust in order to survive, let alone accomplish anything. I will say more about this in a later section.

Logic

Logic is a standard tool in the arsenal of CSR researchers, and it is no less so in the case of HLCSB. But the role is a bit different in CSR and in HLCSB. We have identified three major deficits in traditional logics used in CSR (Anderson and Perlis 2005; Anderson et al 2008; Perlis 1986,1996-7,2010):

(i) time evolves, always, even during thinking. No on-board logic for an agent's use can afford to ignore the fact that – as reasoning proceeds – time is passing

(ii) data will contain errors and outright inconsistencies; there is no way to prevent this if the agent is engaged with the world at large

(iii) semantics – the meanings of the expressions used in a language (formal or natural) – is not fixed for all time, but changes, often rapidly; and even the expressions themselves change – e.g. new expressions or signs come into use.

No traditional CSR logics – including non-monotonic and temporal logics – have mechanisms for dealing with any of these; and even most so-called *paraconsistent* logics employ methods that simply skirt inconsistencies rather than identify and respond to them as potential indications of something amiss. Consequently, to tackle HLCSB we developed so-called *active logics* (Elgot-Drapkin and Perlis 1990) in an attempt to address these deficits; it turns out that a properly-evolving notion of *Now* is a key to all three concerns above (Miller 1993). Such logics have been implemented and are now part and parcel of our MCL work (Anderson and Perlis 2005; Anderson et al 2008; Perlis 2010).

Domains

As mentioned earlier, the usual CSR domain is something akin to a math puzzle. Axioms are given, and a query is to be answered, in a manner agreeing with intuition. An alternative approach – the CYC project (Panton et al 2006) – instead takes a more open-ended view of what is an axiom, even allowing a form of crowd-sourcing as input. But in either case, autonomous dynamic real-world physical interaction is absent or kept to a minimum, and then mainly as proof of principle once the system is ready to perform at its best; see (Reiter 2001) for an impressive example of the latter.

By contrast, we propose a system that is maladroit at first (except in highly constrained artificial settings) but that learns from its mistakes. The notion of an *apprentice* is a rough match: an initially unskilled agent decides it should learn a skill, and does so by a mix of happenstance, trial-and-error, advice, and intentional training; and in some cases may even decide to give up, by its own lights in a wider and evolving set of concerns. One is reminded of Nilsson's call for systems that reason and operate in the context of lifetimes of their own (Nilsson 1983), an early instance of the (now more in vogue) harking back to AI's original human-level focus.

So, expertise enters, but as the *result* of HLCSB, not necessarily built in; and learning enters, but under the *initiation* of the HLCSB agent – it is a learner when it wants to be, in order to address a perceived lack of expertise; and reasoning enters in the form of HLCSB's monitoring success and failure and deciding on what remedial action – if any – to take.

Thus the domain I propose for HLCSB is the real (physical) world, where the agent (say, Robotson C) knows whatever it knows (maybe very little) and by hook or by crook has to manage to survive and get better at it, using a few basic ("designed-in") skills plus a lightweight but general-purpose set of anomaly-handling tools (MCL). Among these, as noted, is that of asking for help, so NLP is a big piece of HLCSB. Thus, perhaps, we have come full circle, back to McCarthy's *Advice Taker* (McCarthy 1959); but now – I believe – we have most of the pieces needed to achieve it. But now I must forestall another counter-objection at this point.

NLP

NLP as anomaly-handling tool? Nearly ready to go, not far beyond current technology? How can that be? This is one of the hardest parts of AI! Well, our work indicates that the very same MCL methodology above also allows an agent to improve its language skills. Indeed, such improvements are featured among our successes to date (Gurney et al 1997; Anderson et al 2003) In fact, NLP plays an interesting dual role here: it is a key resource for one of MCL's most important response strategies to repair mistakes (by asking for and understanding advice), and it is also itself a *source* of many mistakes that MCL then has to cope with (Perlis et al 1998).

To be sure, we do not at present have a human-level NLP system! What we have is a detailed vision that, little by little, is being refined, implemented, and applied to more and more domains. This vision includes an apprentice-like approach to developing language skills, whether new word meanings, new grammatical categories, distinctions between word and meaning, and so on. Some of it is largely along the lines of a logical exercise, noting that a word, say, is now being used in a different way than before. But other portions will require substantial training and almost certainly statistical methods as well. (To some extent we are already taking modest steps in the latter direction, such as in the use of Bayes' nets in some of the MCL's automated management of the response choices.)

Related Work

The idea of building knowledge-based agents that deal with novelty across a wide range of dynamic and uncertain domains, and that do so in part by adapting their intentions and actions, is not new. Ideas from the BDI architecture (Bratman 1987; Rao and Georgeff 1995) in fact have been incorporated to a large extent in much of the MCL work. Indeed, (Josyula 2005) developed an extension, BDIE, that incorporates expectations as another first-class entity, in order to better model the reasoning required in MCL.

SOAR (Laird et al 1987) is a very general architecture (and ongoing implementation) intended to allow effective and flexible integration of multiple subsystems. Thus it aims at a different competence than does the MCL work, which "steps in" when expectations are not met. A SOAR-MCL symbiot would be a very interesting item to investigate.

DALI (Constantini and Tocchio 2008) is a logic-programming language facilitating specification (and even performance) of a BDI agent that has substantial knowledge about its actions and their effects including reasoning, over time. As such DALI has much in common with the active logic techniques we have developed. It does not appear that DALI is able to perform a key task of MCL, in "stepping back" from *ongoing* activities to assess and control them at a metalevel.

Conclusion

Human-level commonsense behavior (HLCSB) is different from commonsense reasoning, in that the former involves a set of general-purpose anomaly-handling strategies that can be used when an agent's ready-to-hand methods (whatever forms of expertise, including CSR) are not producing expected results. Among these strategies are graceful surrender (not all problems are worth solving), asking for help, and training for a new skill.

Our work to date has only made token use of deliberate skill-training strategies. Our focus beginning now will be on autonomous decisions concerning whether a useful skill is lacking, whether it can likely be learned in a useful time frame, which learning methods are most suited, and whether – once initiated – learning is progressing satisfactorily. An exciting venue for this work is a new laboratory being constructed at NRL, a large-scale multi-environment facility for testing physical systems in widely varying realistic condition (terrain, fire, flood, etc).

Acknowledgment

The work reported here was supported in part by grants from AFOSR (FA95500910144), NSF (IIS0803739), and ONR (N000140910328), and by the University of Maryland Institute for Advanced Computer Studies.

References

Anderson, M., Josyula, D., and Perlis, D. 2003. Talking to computers. *Proceedings of the Workshop on Mixed Initiative Intelligent Systems, IJCAI-03.*

Anderson, M., and Perlis, D. 2005. Logic, self-awareness and self-improvement: The metacognitive loop and the problem of brittleness. *Journal of Logic and Computation.* 15(1).

Anderson, M., Fults, S., Josyula, D., Oates, T., Perlis, D., Schmill, M., Wilson, S. and Wright, D. 2008. A self-help guide for autonomous systems. *AI Magazine*, 29(2): 67-76.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. CSLI Publications.

Costantini, S. and Tocchio, A. 2008. DALI: An architecture for intelligent logical agents. *Proc. of the Int. Workshop on Architectures for Intelligent Theory-Based Agents (AITA08), AAAI 2008 Spring Symposium Series*, Stanford, USA.

Elgot-Drapkin, J., and Perlis, D. 1990. Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence,* 2(1), 75-98.

Gurney, J., Perlis, D., and Purang, K. 1997. Interpreting presuppositions using active logic: from contexts to utterances. *Computational Intelligence.*

Josyula, D. 2005. A unified theory of acting and agency for a universal interfacing agent. PhD dissertation, Department of Computer Science, University of Maryland.

Laird, J., Newell, A., and Rosenbloom, P. 1987. Soar: An architecture for general intelligence. *Artificial Intelligence,* 33(1).

McCarthy, J. 1959. Programs with common sense. In*: Proceedings of the Teddington Conference on the Mechanization of Thought Processes,* 756-91. London: Her Majesty's Stationery Office.

Miller, M. 1993. A view of one's past and other aspects of reasoned change in belief. PhD dissertation, Computer Science Dept, Univ of Maryland.

Nilsson, N. 1983. Artificial intelligence prepares for 2001. *AI Magazine*, 4.

Panton, K., Matuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N., and Shepard, B. 2006. Common sense reasoning – From CYC to intelligent assistant. In: Yang Cai and Julio Abascal (eds.), *Ambient Intelligence in Everyday Life*. 1-31, LNAI 3864, Springer.

Perlis, D. 1986. On the consistency of commonsense reasoning. *Computational Intelligence,* vol 2. 180-190.

Perlis, D. 1996-7. Sources of, and exploiting, inconsistency: preliminary report. 1996 Workshop on Commonsense Reasoning (Stanford). Alsoappeared in: *Journal of Applied Non-Classical Logics* 7:1 + 7:2 (1997).

Perlis, D., Purang, K., and Andersen, C. 1998. Conversational adequacy: mistakes are the essence. *International Journal of Human Computer Studies.*

Perlis, D. 2010. To BICA and beyond: How biology and anomalies together contribute to flexible cognition. *International Journal of Machine Consciousness*, 2(2).

Rao, A. S., and Georgeff, M. P. 1995. BDI-agents: From Theory to Practice, In: *Proceedings of the First International Conference on Multiagent Systems* (ICMAS'95), San Francisco.

Reiter, R. 2001. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems.* MIT Press.