CONSISTENCY AT THE CORE OF COMMONSENSE

Donald Perlis
Computer Science Department, University of Maryland, College Park MD, Maryland, U.S.A.

Keywords: Knowledge base, Inconsistency, Autonomous error-recovery, Reasoning, Commonsense.

Abstract: This paper argues for a "commonsense core" hypothesis, with emphasis on the issue of consistency in agent knowledge bases. This is part of a long-term research program, in which the hypothesis itself is being gradually refined, in light of various sorts of evidence. The gist is that a commonsense reasoning agent that would otherwise become incapacitated in the presence of inconsistent data may – by means of a modest additional error-handling "core" component – carry out more effective real-time reasoning, and also that there may be cases of interest in which the "core" is more usefully integrated into the knowledge base itself.

1 INTRODUCTION

The idea of a knowledge (or belief, or information) base (KB) is central to artificial intelligence. Somehow, an automated agent is to make – and possibly act upon – inferences (such as answers to queries, or plans to achieve goals); and this inferring makes use of whatever information (i.e., whatever KB) may be at the agent's disposal. A great deal of work has gone into characterizing such inferences; most of it assumes that the KB itself is consistent. (In what follows, I shall take "KB" to refer to a dynamic store of beliefs together with a knowledge-representation and reasoning framework.)

The consistency assumption has various advantages, both theoretical and practical. There are many important uses of logics for which the axioms form a consistent set; indeed, this is the standard situation in the history of logic formalisms, including many new logics invented for use in computer science (non-monotonic logics, temporal logics, and so on).

But increasingly there has been interest in the alternative so-called paraconsistent situation: the set of axioms is inconsistent and yet – by some key variation on classical logic – the (paraconsistent) logic behaves usefully. We will not attempt to review this extensive literature here, except to point out that one motivation for this interest comes from artificial intelligence, where the need to reason within inconsistent knowledge bases is a serious reality in many situations. Indeed, there is reason to think that inconsistency is nearly inevitable in large-scale dynamic real-world knowledge bases, and especially in the domain of commonsense agent behavior; see (Anderson, Gomaa, et al, 2008; Anderson, Fults, et al 2008; Grant, 1978; Land & Marquis, 2010; Perlis, 1997) for a sampling of work in this area. Essentially the idea is that mistakes will crop up, leading to eventual contradictions, and this in turn requires a repair mechanism to avoid havoc.

The most notorious form of such havoc is this: reasoning (inference) in classical logics with an inconsistent set of axioms is easy; all too easy. Any formula whatsoever is a theorem, in virtue of the validity of the "ex contradictione quodlibet" inference schema: from A and -A, infer B. This property of classical (and some other) logics goes by the aptly colorful name of "explosivity" (Priest, 1996).

The rest of this paper is organized as follows: we describe a notion of commonsense agent behavior for which consistency is an implausible luxury; then our "core" hypothesis is described, along with various forms of evidence; we conclude with a discussion of whether the core is best viewed as part of the KB, or as a separate module.

2 COMMONSENSE BEHAVIOR

Humans survive (by and large) in a complex and rapidly changing world. Much of our competence is surely due to many finely honed "instincts" suited to distinct specialized circumstances; but we also tend to do well when faced with highly novel or irregular

Perlis D..
CONSISTENCY AT THE CORE OF COMMONSENSE.
DOI: 10.5220/0003158405780581
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 578-581
ISBN: 978-989-8425-40-9

situations requiring action and yet for which we have no specific (instinctive or learned) responses. Our ability not to fall apart or – stated more positively – to carry on in an effective manner even when things are not quite as we are used to, perhaps captures much of what is loosely termed commonsense; here we shall call this ability "commonsense behavior". To put it yet another way: when things are suddenly amiss and a (possibly quick) irregularity-fix is needed (but not already at hand), we often come up with something that allows us to continue making useful progress toward at least some goals and to avoid huge increase in costs (often, but not always: the financial wizards did not manage this as the derivatives market began to collapse, nor the oil rig engineers in the Gulf of Mexico, nor the reactor engineers at Chernobyl).

There is an enormous AI literature on commonsense reasoning; here we have defined commonsense behavior to be a little broader, in that it need not involve reasoning, or at least not subtle reasoning used to solve tricky puzzles. Here is an example: you are playing in an outdoor checkers tournament, but several of your pieces fall and roll into a storm sewer. You reach, but fail to retrieve them. You could now ponder at length, treating this as a logic puzzle, hoping for a special insight as to how to retrieve the pieces. Or, you could realize that this might take a long time, that in this situation what matters is not those fallen pieces but rather the ongoing tournament, and that you can ask the referees for advice.

Now, this – asking for advice – is itself a ready-to-hand technique, and so in a sense we do have a fix for many novel things, as long as an expert is handy. But using this kind of fix is very different from knowing specifics about a particular situation. It involves awareness of ones failing efforts, of ones lack of a ready solution or even a good chance at personally finding an appropriate one, and of the availability of someone who might help. These have less to do with checkers than with general ways of coping with irregularities. Thus, roughly speaking, we might divide data in a KB into that portion relevant to a highly specific task-at-hand, and general irregularity-fix data that might be applied more or less independent of the situation.

Just to clarify that asking for help is not the only such strategy, here are a few other frequently-handy irregularity-fix strategies: try again; make small random changes; give up. Yes: even giving up is often a very good thing to do; certainly much better than struggling on and on indefinitely if the cost is great and there is little indication that success is likely. I'll mention one more strategy (often highly effective, but not quick): initiate a training regimen in order to improve (or learn) a particular behavior whose lack was impeding progress.

Now, what has all this to do with inconsistency in a KB? It is this: knowing that the situation is irregular – that one does not already have a strategy at hand – amounts to noting a mismatch between the actual situation and anything one expects. The agent's KB has the item B, perhaps since the agent has the expectation (belief) that the situation is one where B holds; and also in the KB is –B, perhaps as observed data. The two are in contradiction, and the agent must treat this not as a run-of-the-mill set of beliefs with which to reason (that would be to blindly brook explosivity), but as a case where the KB itself is to be looked at as a puzzle: what to do about the anomaly of both B and –B being there? Seen in those terms, a new task arises at a meta-level (that of KB management) and the contradiction becomes a possibly important clue to something needing attention rather than a logical nuisance.

3 CORE HYPOTHESIS

We make the following hypothesis: there is a set of general-purpose irregularity-fix strategies that is:

- adequate to a broad range of novel situations
- concise
- implementable
- largely independent of the size of the KB overall or complexity of the task or domain
- consistent in itself

We refer to this as the commonsense core hypothesis. It can be considered as postulating a fragment of an agent's inferential capacity that has the above properties. Here we will very briefly mention some evidence in its favor.

First of all, humans seem firmly possessed of just such a core set of irregularity-fix strategies. We are quite marvellous at dealing on the fly with myriad unanticipated situations, relying on the same few general-purpose techniques over and over (such as those listed earlier). And – in case that intuitive claim is not convincing – controlled empirical studies have produced data on such strategies in laboratory settings where subjects make high-level general judgments as to their progress vis-à-vis time remaining, confidence in their work, expressions they do not understand, and so on; see (Nelson & Dunlosky, 1994; Nelson, Dunlosky, et al, 1994). Further, neurocognitive work suggests particular

brain structures implicated in error-noting; see (Kendler & Kendler, 1962, 1969). Finally, various implementations of the above ideas have shown success in a wide range of domains (Anderson, Fults, et al, 2008); In these cases however the "core" of irregularity-fix strategies was separate from the system's KB.

Irregularity-fix cores for autonomous agents does exist, as the afore-mentioned implementations show. But can they – or some improved future version – do all that has been hypothesized? What sorts of irregularities might lie beyond any given core set of fixes? Are we in a Godelian situation where, for any core set, there are yet more irregularities beyond its reach? And if a core can reach far, will it no longer be concise? Will core effectiveness scale with the KB? Will a powerful core also require a powerfully expressive language and possibly thereby risk inconsistencies within itself?

There are grounds to think that reach and conciseness and effectiveness are well within the capacity of an implementable commonsense core: much the same grounds cited above for the existence of a core in the first place. But scalability? As humans are faced with more and more information, our effectiveness can sometimes degrade in two ways. Not only can it take us longer consider all the data (though in some cases of course, the extra information makes things go faster), but also there is a heightened likelihood that we will mess up: we'll forget something, lose track of where we are, confuse or conflate similar notions, etc. However, this is not the issue; rather it is whether we cope as well, whether we still notice things amiss and bring corrective strategies to bear, as well as we do when we have a smaller set of facts to deal with. Here I simply state an opinion (in the current absence of empirical data): yes, we do notice our confusion, our lack of progress, and so on, whether working with a large or small KB, on a simple or complex task, and we also respond actively as well: we start over, or ask for help, give up, etc. But we do not rotely go on and on oblivious to the mess we are in.

4 SHOULD THE CORE FIT INTO THE KB?

We now address the last hypothesized item: consistency within the core. As claimed, the commonsense core can be implemented and included as part of an autonomous system. Having the core sit outside the KB – for instance as a monitor-and-control Bayesian net apart from the agent's world model – is an effective design for many purposes. Further, its isolation then protects it from possible infection from a contradiction in the KB. While the KB may be in the throes of explosive inference, the core is not. Even the beginnings of an explosive KB inference process are readily noted by such a core, which in turn then can redirect KB inference in more productive ways. If the core fixes are expressed in propositional language, and together form a concise set, and if each fix is of the simple sort we have described (ask for help, give up, etc) it is plausible that there may well be no internal inconsistency between them.

Yet there are situations in which it may make less sense to separate the core from the KB. Here are four such situations: (i) over time the core trains a new item into the KB so that what had been a particular kind of anomaly handled directly by the core becomes encoded as a familiar event: the core strategy that had been handling these events is now largely replicated in the KB as a standard piece of knowledge about how the world works; (ii) the query "why did you do that?" may require reference to the core, and so the KB reasoner must have some ability to monitor facts about the core: "I did that because I got confused and had to start over"; (iii) "how/why did I do that?" can be asked as an exercise in self-improvement (maybe it can be done better), which suggests bidirectional monitoring and control between core and KB; and (iv) the core itself may behave in an anomalous manner (and if an infinite regress of anomaly-handling meta-cores is to be avoided then we might as well have the all the anomaly-handling inside a single KB at the outset).

On the other hand, combining core and KB raises the danger of inconsistency infecting the core; how serious a problem this may be is currently under investigation.

REFERENCES

Anderson, M., Gomaa, W., Grant, J., Perlis, D., 2008a. Active logic semantics for a single agent in a static world. *Artificial Intelligence 172: 1045-63.*

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

Grant, J., 1978 Classifications for inconsistent theories. *Notre Dame Journal of Formal Logic, 3: 435-444.*

Lang, J., Marquis, P., 2010. Reasoning under inconsistency: A forgetting-based approach. *Artificial Intelligence, 174: 799-823.*

Nelson, T, Dunlosky, J., 1994. Norms of paired-associate recall during multitrial learning of Swahili-Engush translation equivalents. *Memory, 2:325-335.*

Nelson, T., Dunlosky, J., Graf, A., Narens, L., 1994. Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 4: 207-213.*

Kendler, H., Kendler, T., 1962. Vertical and horizontal processes in problem solving. *Psychological Review, 69: 16*

Kendler, H., Kendler, T., 1969. Reversal-shift behavior: some basic issues. *Psychological Bulletin, 72:229-232.*

Perlis, D., 1997. Sources of, and exploiting, inconsistency: preliminary report. *Jounral of Applied Non-classical Logics, 7:13-24.*

Priest, G., 1996. Paraconsistent logic. In *Stanford Encyclopedia of Philosophy* (online version: http://plato.stanford.edu/entries/logic-paraconsistent/)