University of Bielefeld
Faculty of Technology

RVS
Rechnernetze und verteilte Systeme

CITEC

Safety and Security Issues in Human-Machine Cognitive Interaction

Peter Bernard Ladkin

Artificial agents, hereafter called robots, are machines situated in a real-world environment. In any such situation there arises the safety question: can the situated programmed machine behave in such a way as to cause damage (whatever we may think of as damage)? There is a second, related, question: can we, or how can we, assure the integrity of the machine, so that its intended behavior is indeed how it will behave?

Both of these questions gain importance as the behavior of the robots becomes more complex, interacting with humans through gestures and speech.
I propose here some Lines of Investigation (LOI) within which projects addressing these questions may fit.

Short- and Medium-Term: Due Diligence

There is a basic legal issue of ensuring the safety of the researchers interacting with the robots at all stages of the research, including right now. If a researcher were to be seriously injured, tomorrow or next week, in hisher work with a robot, there will be a legal question of due diligence raised of CITEC. Second, there is a basic, medium-term integrity issue, also concerned with due diligence, in protecting the robots against running malware, and detecting it if they do become so compromised. These questions are not routine and there are no readily-available standard answers, technical or legal. It is as well to have studied and anticipated one's due-diligence obligations before such a case arises. I propose a LOI to investigate such short- and medium-term due-diligence issues and develop a resolution.

Longer-Term

Beyond the short- and medium-term, the task gains in complexity. A robot able to move around and pour you coffee when you ask for it may also be capable of pouring the hot liquid on your head. There are, broadly, two reasons why it may do so. One is that it mistakes the space which your head occupies for space in which it expects a cup to be. The other is that it is programmed (it “intends”, if Bert Dreyfus will forgive me) to pour coffee on your head. I treat these two reasons below and propose LOIs accordingly.

Principles of Design for Safety

The first reason points to what I shall call the ***Rational Cognitive Model Coherence Criterion*** (***RCMCC***, explanation of the terminology follows below): ensuring that all participants in an interaction have mutually coherent understanding of the state of the world. Violations of RCMCC have been causal factors in a number of recent accidents to highly automated commercial aircraft, for example the 1992 Lufthansa Warsaw, 1998 Bacolod, 2003 Tainan, and 2007 Sao Paolo landing overruns (all A320) and the 2002 Überlingen midair collision.

(Our experience: my group, and our associated tech-transfer company, have analysed these accidents in detail, in part on behalf of involved parties.)

RCMCC bears close analogy to cache coherence in parallel computing. People designing memory management systems for highly parallel computers regard cache coherence as a design principle. (There was until 2000 only one known memory-management algorithm around that did not enforce cache coherence. It was proved correct using Lamport's TLA by Lamport and myself, as part of a verification challenge problem, which also resulted in solutions by Turing prize winner Amir Pnueli and other prominent verification researchers, published in 1999. I have not worked with memory-management algorithms since that time.) However, there is one safety-critical multi-agent algorithm implemented in kit required on every high-performance commercial aircraft flying, namely the collision-avoidance system TCAS, which is guaranteed to violate RCMCC, which violation was demonstrably a causal factor in the 2002 collision (there was also a non-related TCAS phenomenon, known since 2000, which also played a direct causal role and which one could arguably characterise as an algorithm-integrity issue).

Analysis of the situations in which RCMCC was violated makes use of the notion of Rational Cognitive Model (RCM, which term first appeared in my Überlingen analysis but was implicit in earlier analyses). The *RCM of an agent A at a time T consists of what the agent thinks/believes/stores is the relevant (partial) state of the world*. Its importance lies in that, if the RCM reflects the true state of the world, RCMCC says that all agents in an interaction shall have coherent RCMs. My experience with RCMs suggests that investigations involve largely (small) finite-state-machine engineering techniques.

The RCMCC is one example of a criterion that one could use as a design principle for situated interaction with robots, that will ensure that certain kinds of safety problems do not arise. Are there others? Most certainly. Consider, for example, that a RCM takes no account of the bounded rationality of agents (human and robotic). Bounded rationality has a number of components. One is *bounded perception*. Humans, for example, can discriminate up to 7 or 8 different sounds concurrently, but more are perceived as cacophony. This is an important restriction in the design of warning tones in sophisticated aircraft, for example. A similar principle for visual perception of arrays of warning lights is not yet known. Another component is *bounded reasoning and decision-making*, which has been comparatively well-studied in AI since it was first addressed by Herb Simon 50 years ago. One can imagine a ***Bounded-Rationality Criterion (BRC)****: in context A there shall arise no state in which a safety-related decision or action to be made by agent A requires more reasoning/decision/executive capability that that available to agent A*. I do not know at this point which techniques are appropriate to address BRC questions.

Are there more such principles? Certainly. For example, at a lower level of detail Thimbleby has proposed design principles, in his recent book Press On, for certain sorts of interactions with programmable-digital devices, say, user-programmable so-called “smart” medical devices, and proposes “UI model discovery” as a verification technique. In recent work, he has pointed to what arguably are failures of due diligence in the design of devices on the market. But in general I do not think we yet know what the high-level principles of user interaction, comparable to RCMCC and BRC, should be.

I propose a LOI to formulate such principles, investigate their consequences and violations, and to develop associated methods and tools to check for their fulfilment or lack of it in specific situations. The persistence of designs that violate RCMCC, even in cases in which its violation has been shown to have led causally to fatal accidents, testifies to the importance of investigating the viability and necessity of such principles.

The only such principles generally available at the time of writing are Asimov's (old) three principles, and principles of action being developed currently in the U.S. for warrior robots to replace human soldiers in war situations. I think it unlikely that warrior principles would be generally applicable to peacetime safety requirements.

The RCMCC and BRC as here formulated are new (although there are precursors).

Security and Integrity

The second reason points to the question of how we secure the integrity of a robot's behavior. This question is broader that that of safety: it arises , say, for MAX, whereas safety issues for MAX are limited, since his behavior consists fundamentally only in changing pixels on a computer screen. On the other hand, the consequences of failures of integrity for MAX are similarly inconsequential.

There may be many reasons why the coffee is poured on your head by a robot (other than a violation of RCMCC, or that you deserved it). It could be that two incompatible versions of agent control modules were inadvertently loaded. It could be that an unverified and dangerously faulty rapid-prototype module was loaded. It could be that malware had infiltrated the loaded control system.

These three possibilities fall in the general area of dynamic access control for robot systems. One can imagine that it should be made impossible to load conflicting control modules simultaneously, or the robot be incapacitated when it happens. One can imagine that only verified modules and configurations be loaded. One can imagine procedures for verifying the continued integrity of loaded modules.

Dealing with such issues requires, of course, that a comprehensive “threat model” (of integrity violations) exists, and that effective principles and verification methods exist for ensuring integrity according to this model. I propose a LOI to formulate such integrity principles and a threat model, not only *in abstracto* but also in detail for direct application to the CITEC and Cor-Lab robot development efforts, as well as effective integrity-verification methods and tools which implement these methods.

(Our experience: members of my group derived such a threat model, using my technique Ontological Hazard Analysis, for the installation path of control software from manufacturer to dealer for configurable automobile control and drive components, within the European Union Integrated Project AC/DC.)