Accountable Autonomous Agents: The next level

A Position Paper for the
DARPA Total Intelligence Workshop

Ronald C. Arkin
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA, 30332
arkin@cc.gatech.edu

AI has been questing and foundering in its quixotic search for truly intelligent behavior for over fifty years[1]. Nonetheless we are finally reaching the point where intelligent systems are having significant impact within our society, and vis-à-vis DARPA's mission, the battlefield. Robotic systems and other intelligent battlefield fighting and management systems are already at work in Iraq, Afghanistan, and elsewhere around the world with more on the way. Significant Department of Defense funding has supported these numerous research efforts and they are finally yielding tangible results.

It is not enough, however, for responsible scientists to simply be generating new capabilities for intelligent agents. We need to hold both ourselves and the machines we create accountable for their actions. It is my contention that as we move further and further down this highway towards human level intelligence, we need to invest in machine morality and ethics as a core component of intelligent systems architectures. For those who have been involved in the design of AI architectures, we know it is far easier to incorporate fundamental principles of behavior and intelligence from the onset rather than trying to retrofit them in afterwards.

This new direction cries out for interdisciplinary research involving people with whom computer scientists and roboticists have had little interaction in the past: philosophers, social scientists, political scientists, psychologists, neuroscientists, and the like. We need to now focus on a machine's ability to better address moral and ethical capabilities in context: i.e., among human beings and subject to the presence of social norms. The outcome of this approach can drive a broad range of AI research. These are daunting problems that are easily classified as DARPA-hard:

- The transformation of International Protocols and battlefield ethics into machine-usable representations and real-time reasoning capabilities for bounded morality using modal logics.
- Mechanisms to ensure that the design of intelligent behaviors only provide responses within rigorously defined ethical boundaries.
- The development of effective perceptual algorithms capable of superior target discrimination capabilities, especially with regard to combatant-noncombatant status.

[1] Assuming that the formal advent of AI began in the 1955 Dartmouth Workshop [1]. A cursory search of the original proposal made no mention of ethical reasoning or morality.

- Practical planning (tine-constrained, anytime, explanation-based, consraint-based, etc.) in the presence of moral constraints and the need for responsibility attribution.
- The creation of techniques to permit the learning and adaptation of an embedded ethical constraint set and the agent's underlying behavioral control parameters that will ensure moral performance, should those norms be violated in any way, involving both reflective and affective processing.
- A means to make responsibility assignment clear and explicit for all concerned parties regarding the deployment of a machine with a lethal potential on its mission.
- The establishment of benchmarks, metrics, and evaluation methods for ethical/moral agents, either in some absolute sense or in contrast to the performance of humans in similar situations.
- The design of complete intelligence architectures that incorporate ethical reasoning and behavior as a core principle as opposed to an afterthought.
- Real-time situated ethical operator advisory systems embedded with warfighters to remind them of the consequences of their actions when appropriate.

This calls for an immediate investment in human-robot interaction and more specifically machine/robot ethics. It is a serious mistake to keep plowing ahead without consideration for the consequences of the technology being created in the global context of societal use and acceptance. We must find novel ways to ensure that it is and will be integrated into what we as a society and nation feel is appropriate.

The nascent machine ethics community has begun to study this problem [e.g., 2-4], but it has often been concerned with an artifact developing its own sense of right and wrong through interaction with humans. While this can provide guidance in terms of understanding what morality is, and in some instances this is an acceptable strategy. In others, such as warfare, there is little place for learning basic ethical principles involving the control of an autonomous combat agent in situ. These systems must be inculcated, as our soldiers are more or less, with the correct ethical responses required for given situations. The army has modestly sponsored this author with research in this area, some results of which are reported in [5-8].

In summary, the time is right for a programmatic thrust into comprehensive moral agents capable of ethical action in the broadest sense. Not as replacements for human soldiers bur rather as adjuncts that can ensure that our nation conducts itself as we truly aspire it to.

References

1. McCarthy, J., Minsky, M., Rochester, N., and Shannon, C., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence", 1955.
2. Anderson, M., Anderson, S., and Armen, C., "An Approach to Computing Ethics*", IEEE Intelligent Systems*, July/August, pp. 56-63, 2006.

3. Bringsjord, S. Arkoudas, K., and Bello, P., "Toward a General Logicist Methodology for Engineering Ethically Correct Robots", *Intelligent Systems*, July/August, pp. 38-44, 2006.

4. Turilli, M., "Ethical Protocols Design", *Ethics and Information Technology*, Vol. 9, pp. 49-62, March 2007.

5. Arkin, R.C., *Governing Lethal Behavior in Autonomous Systems*, Chapman and Hall Imprint, Taylor and Francis Group, to appear Spring 2009.

6. Arkin, R.C., "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture", GVU Technical Report GIT-GVU-07-11, GVU Center, Georgia Institute of Technology, 2007. Also derivative papers published in three parts: Part I: Motivation and Philosophy, *Proc. Human-Robot Interaction 2008*, Amsterdam, NL; Part II: Formalization for Ethical Control, March 2008, *Proc. 1st Conference on Artificial General Intelligence*, Memphis, TN, March 2008; and Part III: Representational and Architectural Considerations, *Proceedings of Technology in Wartime Conference*, Palo Alto, CA, January 2008.

7. Arkin, R.C., Wagner, A.R., and Duncan, B., "Responsibility and Lethality for Unmanned Systems: Ethical Pre-Mission Responsibility Advisement", GVU Technical Report GIT-GVU-09-01, GVU Center, Georgia Institute of Technology, 2009.

8. Arkin, R.C., Ulam, P., and Duncan, B., *"An Ethical Governor for Constraining Lethal Action in an Autonomous System",* GVU Technical Report GIT-GVU-09-02, GVU Center, Georgia Institute of Technology, 2009.