Righteousness and Conscience as a Path to Socially Acceptable Autonomous Behavior

Rick Dove, rick.dove@incose.org

Systems engineering is advancing the autonomous decision-making abilities of systems on many fronts: from self-driving cars that platoon on the highways and collaborate on developing traffic conditions, to weapon systems that work alone and in heterogeneous swarms. What will govern the behavior of these unsupervised systems when life, property, or mission is at stake, and when previously unanticipated situations arise? A nearly infinite variety of environmental situations is possible, and impossible to test in advance. Independent of the environment, systems fail, and complex systems can exhibit complex unanticipated behaviors. Systems are also targeted by intelligent and persistent adversaries intent on system intervention.

How can we turn these things loose—with any confidence? Maybe we can take a lesson from the social systems we know. I would argue that society works when an individual conscience governing personal behavior connects to a collective righteousness that governs group behavior.

As Allison George (2012) summarizes Jonathan Haidt's new book *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (2012), in its original meaning, *righteous* means just, upright, and virtuous. But in a colloquial sense, it often means self-righteous, judgmental, moralistic. The evolutionary story Haid tells in his book is one where moral judgment is the ability to create moral matrices and punish, shame, and ostracize those who do not behave rightly. As we will see, judging the behavior of others is not limited to humans in the animal kingdom.

The authors of Wikipedia have defined *conscience* as "an aptitude, faculty, intuition or judgment of the intellect that distinguishes right from wrong. Moral judgment may derive from values or norms (principles and rules). In psychological terms conscience is often described as leading to feelings of remorse when a human commits actions that go against his/her moral values." A common metaphor for conscience is the "voice within."

Artificial autonomous systems present a particular challenge for morality. Work led by Ronald Arkin at Georgia Institute of Technology is concerned with the ethical behavior of unmanned autonomous systems (UAS) used in military operations, and recognizes the potential for peer monitoring: "When working in a team of combined human soldiers and autonomous systems, they [unmanned autonomous systems] have the potential capability of independently and objectively monitoring ethical behavior in the battlefield by all parties and reporting infractions that might be observed. This presence alone might possibly lead to a reduction in human ethical infractions" (Arkin 2007: 7). Here Arkin appears to focus on robots monitoring human behavior; but with this capability they could also monitor the behavior of other robots as well as their own behavior. Arkin does recognize robotic self-monitoring, as in the work of Moshkin and Arkin (2007: 1), who describe their task as "designing a computational implementation of an ethical code within an existing autonomous robotic system, i.e., an 'artificial conscience', that will be able to govern an autonomous system's behavior in a manner consistent with the rules of war."

Arkin's project included a survey "to establish opinion on the use of lethality by autonomous systems spanning the public, researchers, policymakers, and military personnel to ascertain the current point-of-view maintained by various demographic groups on this subject" (Moshkina and Arkin 2007: 1). Figure 1 shows survey results comparing what is expected of robots and soldiers, with applicability of ethical categories ranked from the more concrete and

specific at the bottom of the chart, to the more general and subjective at the top.

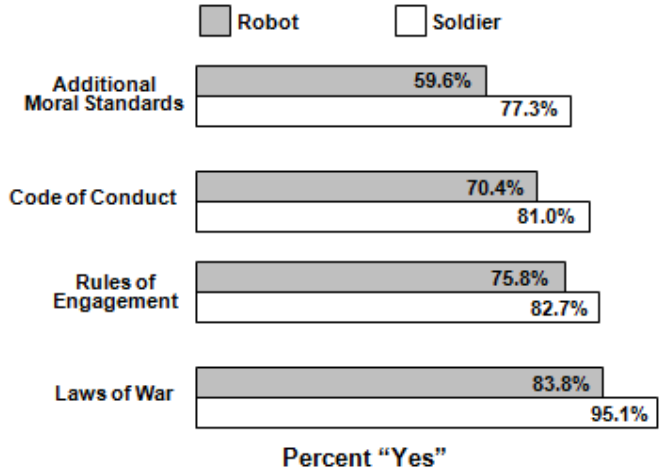

Figure 1. Ethical behavior for soldiers and robots. Reproduced with permission from the survey reported in Moshkina and Arkin (2007).

This article does not attempt a philosophical approach to ethics and morals; but rather suggests that autonomous systems should include embedded agents that serve as judgmental inner voices, which constantly monitor and evaluate the behavior of self and the behavior of others, relative to expected and acceptable behaviors. Monitored behavior would include acceptable patterns of ethics and morals, as well as the expectations of performance and mission. These inner voices need to be implemented as self- and peer-policing behavior recognition agents generally independent of influence by other UAS subsystems.

Abnormal behavior in autonomous systems is likely. Regardless of the degree of autonomous control, it is still possible for design flaws, system malfunction, malevolent control penetration, and emergent situational response to cause abnormal behavior. Simulation and test of individual units with these autonomous capabilities have their own sets of challenges, and cannot predict how these units will behave in group operations. Individual behavior cannot be ignored as simulation and testing advances to group behavior, since it poses an explosive monitoring and evaluation task.

This article suggests that both UAS self and peer evaluation of behavior is necessary when UAS are working unsupervised; and perhaps even more so when these systems are being tested, where they are even less likely to be well behaved. Though we focus here on warfighting machines, because the outcome of bad behavior has the potential for great harm, these behavior-monitoring concepts are equally applicable to an autonomously driven car that exhibits rude driving behavior, or a mobile ad-hoc network node that interferes with overall network performance.

The trend toward increased autonomy in unmanned weapon systems has raised concerns about methods for testing these devices both individually and in tactical group maneuvers (US Department of Defense 2011). Increased autonomy is generally enabled and permitted by increased intelligence of the artificial kind in UAS. Intelligent systems, be they human or artificial, exhibit behaviors in response to situational conditions. Situational conditions are unpredictable and infinite in potential variety, leading to emergent behaviors at both the individual and group level. For UAS in warfighting, emergent behavior is necessary and desirable when it is appropriate and useful, and potentially a major problem when inappropriate.

Such unexpected system behaviors can be good as well as bad. In fact, the goal of fully autonomous intelligent behavior is creative problem solving in situations without precedents. It is unlikely that unleashing a swarm of UAS that are only capable of dealing with well-defined cataloged situations will be effective.

UAS will necessarily be tested and fielded in situations that have no database of cataloged responses. How will we constrain the outcomes to those we can live with? More to the point of this article, how will we detect and evaluate behavior in time to intervene if unacceptable consequences are the likely outcome?

Range testing can never duplicate the situational variety that will arise in warfighting, anymore than prequalifying the capabilities and performance of Michael Vick as an NFL team player was able to prevent his later behaviors, which reflected poorly on all players by association (subsequent rehabilitation not withstanding). UAS that run amok in any way will reflect poorly on all UAS—eroding necessary public trust.

Moshkina and Arkin (2007) identified the important need of UAS conformance to rules of ethics, rules of war, and related high-level behaviors expected by the public of weapons-toting UAS. Infractions can be devastating to continued public acceptance as well as to life and property. Range testing alone cannot assure safety under warfighting conditions. This article suggests that verification of performance and mission behaviors and validation of social behaviors become continuous processes throughout the operational life of UAS, carried out by UAS-embedded self and peer behavior monitoring agents.

Peer-behavior monitoring occurs naturally and constantly in social animals. Each member of the group evaluates the others for adherence to social norms and threats to social coherence and security. Rogue elephants, for instance, are the result of banishment for unacceptable behavior. Social insects are known to restrain and even kill members of the group that overstep certain social bounds.

Humans monitor the behavior of others in ways more sophisticated and more complex than animals of lesser cognitive capability. The process is often carried out formally as a test for granting new candidates membership in a group. Initial tests are typically for similar values, compatible behaviors, acceptable capabilities, and even for synergy in mission-based groups such as sports teams or special forces.

The training and vetting of special-forces operatives provides an experience base. One point to note is that we appear comfortable moving operatives into field status after some "testing" period, even though we know they will face situations that have not been tested. Another point to note is that these operatives on mission are always being evaluated by their peers, who rely on the integrity of each and every member of a team.

Behavior Monitoring and Evaluation

Unlike traditional approaches to sophisticated behavior detection and classification through reasoning, this article suggests an approach inspired by cortical reverse-engineering studies, where it appears that a vast quantity of simultaneously accessible "experience" patterns drives an immediate conclusion, rather than a compromising sequential search or reasoning process. This pattern approach is suggested here, with working examples outlined in Dove (2009a), made possible with new processor architectures offering simultaneous pattern-recognition on a virtually unbounded number of behavior patterns (Dove 2009b).

Research indicates that human expertise (extreme domain-specific sense-making) is strongly related to meaningful pattern quantity. People considered truly expert in a domain (such

as chess masters and medical diagnosticians) are thought to be unable to achieve that level until they have accumulated some 200,000 to a million meaningful patterns. The accuracy of their sense-making is a function of the breadth and depth of their pattern catalog. Of note in biological entities, the accumulation of large expert-level pattern quantities does not manifest as slower recognition time. All patterns seem to be considered simultaneously for decisive action. There is no search, evaluation, or reasoning activity evident.

Philip Ross (2006) talks about the expert mind, and Herb Simon's "chunking" explanation for how chess masters can manage and manipulate a vast storehouse of patterns. Ross ties this chunking discussion into the common understanding that the human mind seems limited by seven plus-or-minus two elements in working memory: "By packing hierarchies of information into chunks, Simon argued, chess masters could get around this limitation, because by using this method, they could access five to nine chunks rather than the same number of smaller details."

Concluding Remarks

The requirement for new concepts of verification and validation, and for test and evaluation, is outlined in the recently released *Unmanned Systems Integrated Roadmap* (US Department of Defense 2011): "As unmanned systems become more complicated, more integrated, more collaborative, and more autonomous, establishing test-driven development constructs and infrastructure for supporting early-onset test and evaluation (T&E) and life-cycle T&E will become increasingly critical" (10). "The rapid acquisition of quickly evolving unmanned systems will require an unmanned systems T&E capability that evolves at a pace that exceeds this evolution." (40). "For unmanned systems to fully realize their potential, they must be able to achieve a highly autonomous state of behavior and be able to interact with their surroundings. This advancement will require an ability to understand and adapt to their environment, and an ability to collaborate with other autonomous systems, along with the development of new verification and validation techniques to prove the new technology does what it should" (45).

This article suggests that a promising, perhaps necessary, basis exists in the recognition and evaluation of social behavior, with technology that can manage and learn a vast quantity of stored reference patterns, which are structured and accessed in a "feedforward," chunked hierarchy.

Social-comparison theory guides us to a comparison of an agent's behavior pattern against behaviors of others on the team, against mission plans, against defined patterns of normal behavior, and against defined patterns of aberrant behavior. And then there is the question of behavior never seen before—emergent behavior that must be evaluated for consequence, and remembered thereafter as newly-learned and classified behavior. Expertise theory, if it can be called that, guides us to a need for an extremely large number of reference patterns that can be simultaneously compared relative to a dynamic situation, eliminating time for sequential evaluation and reasoning steps, and eliminating much of the otherwise selective monitoring and pattern simplification that increases uncertainty.

Verification that requirements are met and validation that system intent is met are not sufficient as currently practiced. Test and evaluation cannot possibly present all possible scenarios that will be encountered in confrontation with the real world. Two things as a minimum must be added to the systems engineering repertoire. First, there should be UAS-integrated operational behavior evaluation throughout the system lifecycle. Second, there should

be effective detection and evaluation of emergent behavior. I suggest that independent but integrated agents of conscience and righteousness are two necessary elements for well-behaving autonomous systems.

It is time to learn how to craft these inner-voice agents; and how to integrate them into the systems engineering processes of verification and validation, test and evaluation, and the operational lifecycle.

References

Arkin, Ronald C. 2007. "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." Technical Report GIT-GVU-07-11. Atlanta US-GA: Mobile Robot Laboratory, College of Computing, Georgia Institute of Technology. http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

DiRose, S., and R. Dove. 2012. "Peer Policing: A Pattern for Detecting and Mitigating Aberrant Behavior in Self-Organizing Systems." Paper to be presented at IEEE GLOBECOM Conference (Security Track), Disneyland, US-CA, 3–7 Dec.

Dove, R. 2009a. "Paths for Peer Behavior Monitoring Among Unmanned Autonomous Systems." *International Test and Evaluation Association Journal* 30 (3): 401–408.

———. 2009b. "Methods for Peer Behavior Monitoring Among Unmanned Autonomous Systems." *International Test and Evaluation Association Journal* 30 (4): 504-512.

George, Allison. 2012. "What Righteousness Really Means." Interview with Jonathan Haidt. *New Scientist Magazine* 2854 (8 March): 30–31.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* New York, US-NY: Pantheon.

Moshkina, L., and R. C. Arkin. 2007. "Lethality and Autonomous Systems: Survey Design and Results." Technical report GIT-GVU-07-16. Atlanta, US-GA: Mobile Robot Laboratory, College of Computing, Georgia Institute of Technology. http://www.cc.gatech.edu/ai/robot-lab/online-publications/MoshkinaArkinTechReport2008.pdf.

Ross, P. 2006. "The Expert Mind." *Scientific American* 295 (2): 64–71.

US Department of Defense. 2011. *Unmanned Systems Integrated Roadmap FY2011-2036*. Washington, US-DC.