Serious Games: Simulations for Robot Risk Assessment and Communication

by Aaron Mannes

Along with new opportunities, new technologies bring new and often unpredictable risks. Simulations, war-games and tabletop exercises (TTX) can be useful mechanisms, not only for assessing and managing risks, but also for the equally vital task of risk communication. Failures to properly assess risks and engage in risk communication have stymied technological development in the past.

As AI, whether virtual or embodied as robots or IoT, becomes ubiquitous the potential for accidents and failures, both mundane and dramatic will become increasingly likely. These failures can include tangible harms such as an autonomous vehicle causing an accident or algorithmic bias denying an individual benefits. Failures may also be more subtle, but still harmful, such as incidents that undermine individual dignity.

Simulations can be used both to identify risks, but also to consider how best to manage these risks. Wargames, in which teams compete against one another, can be used to consider how criminals might manipulate AI. TTX can be used to test a crisis management plan, so that an organization can prevent a minor failure from becoming a larger one. Simulations have been used for these purposes, as well as to teach general concepts, across a wide variety of domains including disaster response, national security, and corporate crisis management.

Given the scale at which AI is being deployed and the unpredictable nature of both AI itself and of how it will interact with individuals, organizations, and society more broadly, some failures and accidents are inevitable. Properly conducted risk communication can help build trust between communities using and affected by AI so that when failures and accidents occur they can collaborate effectively to address the situation. When risk communication does not focus on building this relationship of trust, technological failures can lead to a strongly negative public response that can stymie technological development. (The Three Mile Island nuclear incident is the classic case of poor risk communication leading the general public to turn against a technology.)

Simulations in which representatives of stakeholder communities participate can help build this trust. On the one hand, stakeholders will have the opportunity to observe decision-making by those creating and deploying AI. On the other hand, the creators and implementers of AI may make assumptions about stakeholder attitudes and reactions. Bringing the stakeholders into the simulation can elicit their attitudes and values so that AI can be developed and implemented accordingly. By bringing communities together in an environment that can be stimulating and sometimes fun, trust can be developed so that when accidents and failures occur they do not derail promising AI applications.