Analyzing Human Trust of Autonomous Systems in Hazardous Environments

Daniel P. Stormont

Department of Computer Science
Utah State University
Logan UT 84322-4205
daniel.stormont@aggiemail.usu.edu

Abstract

Autonomous systems are becoming more prevalent in our everyday lives. From agent software that collects news for us or bids for us in online auctions to robots that vacuum our floors and may help care for us when we grow old, autonomous agents promise to play a greater role in our lives in the future. Yet, there is ample evidence that many people do not trust autonomous systems – especially in environments where human lives may be put at risk. Examples include search-and-rescue operations at disaster sites, military operations in a combat zone, and caregiving scenarios where a human's life depends on the care given by an autonomous system. This paper uses previous work on trust in multi-agent systems as a basis for examining the factors that influence the trust humans have in autonomous systems. Some of the technical and ethical implications of relying on autonomous systems (especially in combat areas) are considered. Some preliminary results from a simulation of a firefighting scenario where humans may need to rely upon robots are used to explore the effects of the different factors of human trust. Finally, directions for future research in human trust of autonomous systems are discussed.

Introduction

We rely on autonomous systems every day – often without realizing it. The Global Positioning System satellites that help us find our way through unfamiliar territory, the bidding agent we entrust with helping us win an auction on eBay® , or the Roomba® robot vacuuming our floors are all examples of autonomous systems. In the near future, autonomous systems are likely to play an even greater role in our lives. For example, in Japan, Honda has made a long-term investment in the development of the Asimo humanoid robot. Honda envisions this robot carrying out a wide range of tasks in Japan, necessitated by an aging population and a shrinking workforce. Among the tasks envisioned for Asimo are tasks in the service industry, menial office tasks, and – most importantly – elder care. In fact, to make Asimo less intimidating, it has been reduced in height from its original adult human size to a more child-like 130 cm in height (Honda 2007).

In the United States, Congress mandated in 2001 that one-third of all combat ground vehicles will be unmanned by 2015 and the armed forces of the United States continue to support active development of unmanned systems for land, sea, and air. While these systems exhibit differing degrees of autonomy, nearly all of them have some autonomous functions in order to reduce the workload of the operators. The United States National Institute for Standards and Technology (NIST) has been a key player in introducing robotic systems (both autonomous and non-autonomous) into search-and-rescue scenarios. NIST holds a number of exercises each year to allow experienced rescuers to evaluate robotic systems and they are one of the primary sponsors of the annual RoboCup Rescue Robot competition.

One factor that is slowing the acceptance of autonomous systems, especially in hazardous environments such as disaster areas and combat zones, is human distrust of autonomous systems. This distrust is illustrated by regulations that require human operators to manually observe the operations of fully autonomous systems and take control of them if the operators are uncertain what the autonomous system is doing to the statement by human rescuers that they will never accept an autonomous system that tells them whether or not there is a victim trapped in a rubble pile. Given the advancing technological capabilities of autonomous systems and the obvious advantages of supplementing limited human resources in a hazardous environment, what would be necessary for humans to trust autonomous systems?

This paper looks at some of the factors of human trust of autonomous systems. It begins with a definition of an autonomous system to ensure that the reader understands what autonomy is and what the implications of an autonomous system are. Then, it considers some of the research that has been published concerning the definition and formalizing of the concept of trust – especially as it applies to autonomous systems. Using this framework of existing research into trust, a computer simulation of a hazardous environment – in this case a brush fire simulation – that simulates both human and robotic firefighters is introduced as a testbed for exploring the factors of trust. The simulation results are reported and interpreted with regard to the components of human trust in autonomous systems. Finally, conclusions about the simulation results are reported and opportunities for future work in this area are identified.

What is an Autonomous System?

The dictionary definition of autonomy is: "independence (Houghton Mifflin 2007)." A good definition of the term autonomous robot is: "Autonomous robots are robots which can perform desired tasks in unstructured environments without continuous human guidance (Wikipedia 2008)." The critical element in this definition is that an autonomous robot (or system) operates in an unstructured environment, that it can perform desirable tasks, and that it does not require human intervention on a regular basis. As seen in figure 1, autonomy falls into a spectrum; from no autonomy (tele-operated or "remote control" robots) to full autonomy (robots that act with no human intervention). In between these extremes are robots with partial autonomy (e.g., an "intelligent" manipulator that can be commanded to a desired position or directed to perform a task without continuous input from the operator) and robots with sliding autonomy (the degree of autonomous operation is selectable by the operator).

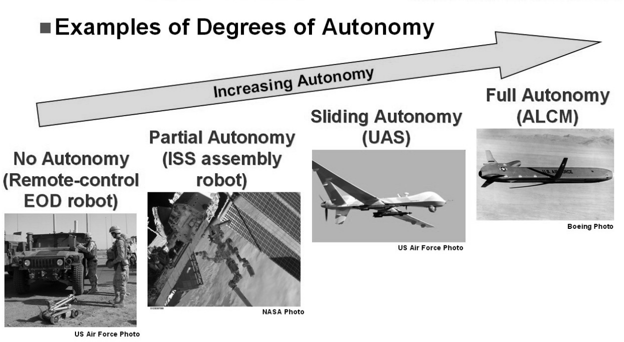


Figure 1. Examples of degrees of autonomy for autonomous systems, from no autonomy to full autonomy.

What is Trust?

While trust has been discussed by philosophers and theists since before the time of written history, it really became a topic of research as a result of the work of Diego Gambetta (Gambetta 1988). Gambetta's approach to trust tended to focus on human interactions, rather than on autonomous systems, but his definition for trust is a good starting point:

> trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it) *and* in a context in which it affects *his own* action. When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him.

Gambetta's work was extended to consider autonomous agents by Castelfranchi and Falcone. In (Castelfranchi 1999), two critical extensions were identified. First, they identified that only cognitive agents can "trust" another agent. Second, they identified a framework of goals and beliefs that were essential to trust another agent (whether human or agent). The "core trust" element requires two beliefs: that an agent that can perform a task to help you achieve a goal has the ability to perform the task and the desire to perform the task. If both of these beliefs are satisfied, then the "reliance" element comes into play, where you are relying on the agent to perform the task you believe it is capable of and willing to do.

Some additional work in the area of trust for autonomous systems has been performed as part of the ALADDIN project. In (Ramchurn 2004), two principle components of trust were identified: confidence (do I believe the agent can perform the desired task) and reputation (has this agent been successful in the past and have others trusted this agent with good results).

Similarly, Fullam and Barber (Fullam 2007a) explore two sources for trust information: experience and reputation. The paper utilizes the Agent Reputation and Trust (ART) Testbed to investigate reputation exchange and the value of experience with an agent in a simulation of art appraisals. The ART Testbed is also used for an annual competition in agent trust relationships. The ART Testbed interface is shown in figure 2 (Fullam 2007b).

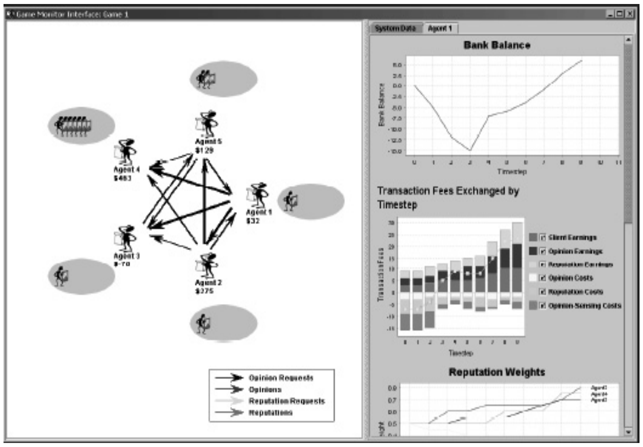


Figure 2. The Agent Reputation and Trust (ART) Testbed interface.

Implications of Distrust of Autonomous Systems

As Robin Murphy of the Center for Robot-Assisted Search and Rescue noted (Murphy 2004):

> One impact of the human side is that rescue workers today refuse to consider fully autonomous systems designed to act as 'yes/no there's something down there' search devices. Scenarios where a swarm of hundreds of robot insects are set loose to autonomously search the rubble pile and report to a single operator not only appear impractical for search (and certainly for structural assessment, victim management, and extrication as well), but also ignore the organizational context of USAR, which has a hierarchy of operators who check and verify any findings from dogs, search cameras, or other sources.

The lack of trust in autonomous systems also manifests itself in regulatory guidance. The United States Federal Aviation Administration does not allow the operation of unmanned aerial vehicles in the national airspace, with the exception of some specially identified operating areas (usually military airspace) and in national emergencies, with the proper coordination. Operation of autonomous air vehicles is never allowed in national airspace (Lazarski 2001). This requirement for maintaining human control over systems capable of operating autonomously has made their employment more difficult than necessary. A recent study found that operators of unmanned aircraft were the most over-utilized and overstressed operators of aircraft

Another critical issue of autonomous systems, especially when operated in combat areas, is the legal ramifications. The Law of Armed Conflict (LOAC) would seem to apply to the use of robotic systems in a combat zone. This means that the commander of a unit utilizing robotic systems could be held legally accountable for civilian deaths, injuries, and loss of property resulting from the use of a robotic system (Lazarski 2001).

More important than the legal ramifications are the moral ramifications. Should an autonomous combat vehicle, such as the Unmanned Combat Air Vehicle shown in figure 3, be allowed to make targeting decisions using its own heuristics? Who is responsible if it erroneously targets innocents? Should life and death decisions be made by an autonomous system? Consideration of the moral aspects of autonomous systems have been lagging the technical aspects.

Figure 3. An unmanned combat air vehicle. (Illustration from DARPA.)

Evidence of concerns about utilizing armed robots occurred recently with the deployment of the Special Weapons Observation Remote Direct-Action System (SWORDS) to Iraq. Three of these armed robots (which are tele-operated, not autonomous – see the photograph in figure 4) were sent into the theater, but the weapon on the robot has never been fired in a combat situation and there have been a number of rumors about the robots being undependable and taken out of service – even though there is no evidence to back up the rumors (Popular Mechanics 2008).

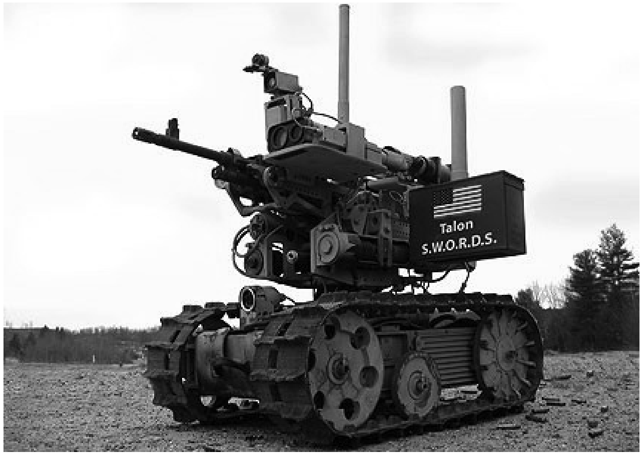


Figure 4. The SWORDS armed reconnaissance robot.

Factors Affecting Trust

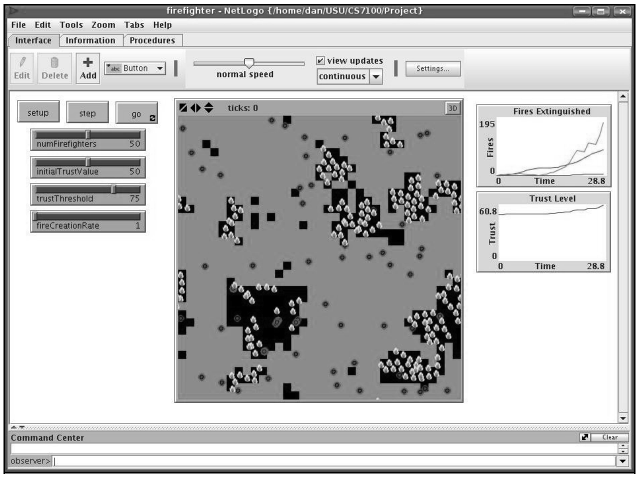
Using the two components of trust identified in (Ramchurn 2004) as a starting point, it is apparent that autonomous systems (or for that matter, robots in general) tend to not have a good reputation. While the reasons for this are not entirely clear, given the obvious advantages of being able to use robots in hazardous environments instead of risking human lives; it is apparent that the other trust component – namely, confidence – plays a key role in the poor reputation of robots.

One reason for a lack of confidence in robotic systems is their lack of reliability in field conditions. A 2004 study of commercially available ruggedized robots operating in field conditions showed a mean-time-between-failures (MTBF) of 12.26 hours and an availability rate of 37% (Carlson 2004). Obviously, robots need to become much more reliable if humans are going to have confidence in them.

However, a more important reason for lacking confidence may be the unpredictability of autonomous systems. When working with humans, we usually can anticipate their actions in a wide range of circumstances – especially if we have trained with them, as is the case in rescue crews and combat teams. But autonomous systems have a tendency to surprise even their creators. This is one of their strengths, as they can often come up with an unanticipated solution that is better than any solution that could be programmed in to the system, but this unpredictability can be very disconcerting in hazardous environments.

A Simulation of Human-Robot Interactions

In order to better understand distrust of autonomous systems in hazardous environments, a firefighting simulation was created in the NetLogo simulation environment (Wilenski 1999). A screenshot of the simulation is shown in figure 5.


firefighter - NetLogo {/home/dan/USU/CS7100/Project}

numFirefighters 50
initialTrustValue 50
trustThreshold 75
fireCreationRate 1

ticks: 0

Figure 5. A simulation of a firefighting scenario developed in NetLogo.

The simulation has two types of agents at present: human firefighters and robotic firefighters. An equal number of firefighters are randomly distributed in a grassland environment. The number of firefighters is determined by the value of a number of human firefighters variable that is selected by the user of the simulation using a “slider bar” on the interface. The variable values are all set before a run of the simulation and can not be changed during a run. At each turn of the simulation, fires start randomly according to the setting of a rate of fires variable on one of the sliders. A fire can be extinguished if a firefighter is in the same location as the fire for one turn of the simulation. Any fires that are not extinguished will spread into an adjacent location, if that location has not already been burned (which is indicated by the black patches on the map). The direction of the spread of the fire is determined by a random wind direction variable, which is set at the start of the turn and affects the spread of all existing fires in the same manner. Another slider can be used to determine the trust threshold and the initial trust value. The initial trust value is the trust the human firefighters have in the robotic firefighters at the beginning of the simulation. The trust threshold is the value below which the human firefighters will not call on the robotic firefighters. The trust value can increase (or decrease) based on the performance of the robotic firefighters when they are called upon. The trust value decreases more rapidly than it increases to reflect the human bias against autonomous systems. The trust threshold will lower if the workload gets too high for the human firefighters (there are too many fires for them to extinguish) and as the human firefighters get fatigued from fighting too many fires consecutively. The interface has graphs that display the number of fires extinguished by the human and robotic firefighters and an indication of the current trust level.

Simulation Results

The results from the initial version of the simulation are not especially surprising. As can be seen in the graph in figure 6, the robot firefighters are not called upon initially. At this point, the human trust level is still below the threshold level. However, once the situation starts to get out of hand, the human firefighters start to call upon the robotic firefighters to assist in extinguishing the fires. Once the situation is under control, the trust in the robotic firefighters starts to drop a little bit as the human firefighters’ biases come into play. Finally, as the human firefighters grow fatigued toward the end of the simulation, they call upon the robotic firefighters again to help put out the last of the fires.

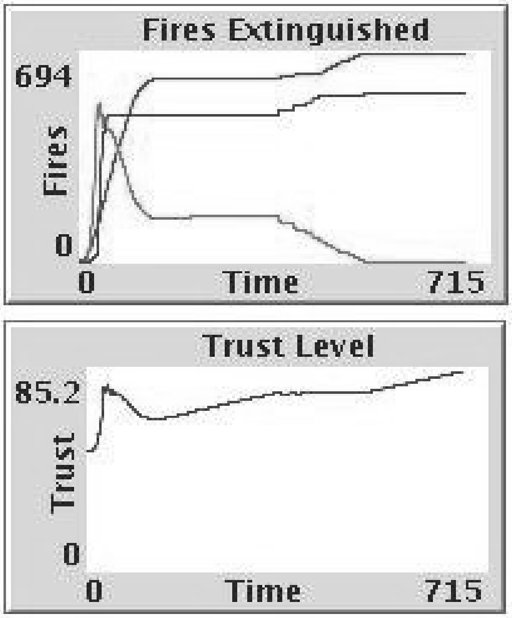


Figure 6. Graphs illustrating the results of a typical simulation run. The top graph shows the number of fires extinguished (with separate traces for the human and robotic firefighters) and the bottom graph shows the changing level of trust the human firefighters have in the robotic firefighters over time.

The results of the simulation do validate the expectations for the scenario, but the simulation is currently too simple to provide answers to the really interesting questions about the human-robot trust relationship in a hazardous environment. To answer these questions, the firefighting simulation needs to be made more complex, to reflect the trust factors that have been identified in the literature.

Conclusion and Future Work

The survey of the current literature on trust of autonomous systems should make it apparent that there is no widespread agreement on the factors that contribute to human trust in robots – especially in hazardous environments. Some important factors are experience and reputation. The experience will come from individual humans who take a chance on autonomous systems out of necessity or curiosity. The reputation will come from the prevalence of autonomous systems that reliably answer a need that can not be satisfied by humans, or at least, not without greater risk. At present, there has been some experience with robotics in combat zones that has led to increasing demand for robots that can perform the most dangerous tasks (such as explosives disposal) under human control, but without putting a human at risk (Gates 2008). However, the use of autonomous systems in combat zones or other hazardous environments, continues to be a capability that is far in the future (OSD 2007).

To try to identify the elements of trust that are most critical to human acceptance of robotic assistance in hazardous environments, the firefighting simulation is going to be developed further. Variables that represent some of the elements of trust identified in the literature, such as reliability and predictability, are going to be added to the interface. Also, the simulation is going to be made more realistic by including human trust levels in other humans and heterogeneous robot types. For example, robots that move more slowly than other types, but can carry more fire suppressant or robots that can traverse rough terrain.

In addition to extending the firefighting simulation, a more diverse scenario could also promise to yield interesting insights into human-robot trust issues. The RoboCup Federation has developed a simulation of rescue operations on a city-wide scale (see figure 7). The RoboCup Rescue Simulation currently simulates human police, ambulance, and firefighting agents, under the control of dispatcher agents (RoboCup 2008). The simulation can be readily extended to add in robotic police, ambulance, and firefighting agents. It could also be possible to create a human user interface for the role of dispatcher, to run human trials on human-robot trust. In other words, what would an autonomous system need to do for an experienced dispatcher to decide to dispatch a robot to assist a human agent or in place of a human agent?

There are still many questions to answer, but the answers to these questions could benefit not only the designers of autonomous systems, but also the potential users of these systems, and – ultimately – all of us.

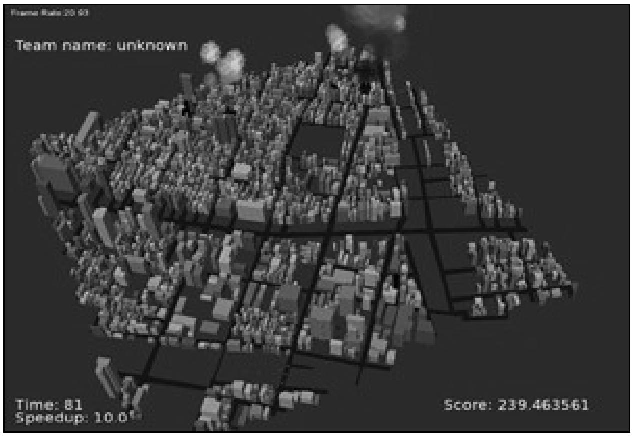


Figure 7. The user interface for the RoboCup Rescue Simulation.

References

Castelfranchi, C. and Falcone, R., 1999. “Principles of Trust for MAS: Cognitive Autonomy, Social Importance, and Quantification,” *Proceedings of ICMAS 1999*. Online: http://aois.org/99/castelfranchi-ICMAS-paper.rtf.

Fullam, K., 2007a. *Agent Reputation and Trust Testbed.* Online: http://www.lips.utexas.edu/art-testbed/.

Fullam, K. and Barber, K., 2007b. “Dynamically learning sources of trust information: experience vs. reputation,” *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007).*

Gambetta, D. ed., 1988. *Trust: Making and Breaking Cooperative Relations,* New York, NY: Basil Blackwell Ltd. Online: http://www.nuffield.ox.ac.uk/users/gambetta/gambetta trust book.pdf.

Gates, R., 2008. Secretary Gates Remarks at Maxwell-Gunter Air Force Base, Montgomery Alabama, *DefenseLink News Transcript*, April 21, 2008. Online: http://www.defenselink.mil/utility/printitem.aspx?print=http://www.defenselink.mil/transcripts/transcript.aspx?transcriptid=4214.

Houghton Mifflin, 2007. “Autonomy,” *American Heritage Dictionary of the English Language.* Boston, Mass.: Houghton Mifflin.

Honda, 2007. *Asimo: The World’s Most Advanced Humanoid Robot.* Online: http://asimo.honda.com/default.aspx.

Lazarski, A., 2001. “Legal Implications of the Uninhabited Combat Aerial Vehicle,” *Air & Space Power Journal*, March 2001.

Murphy, R., 2004. "Rescue Robotics for Homeland Security," *Communications of the ACM*, special issue on Homeland Security, vol. 27, no. 3, March 2004, pp. 66-69.

Office of the Secretary of Defense, 2007. *Unmanned Systems Roadmap, 2007-2032.* Online: http://www.acq.osd.mil/usd/Unmanned%20Systems%20Roadmap.2007-2032.pdf.

Popular Mechanics, 2008. “The Inside Story of the

SWORDS Armed Robot 'Pullout' in Iraq: Update," *Popular Mechanics Online.* Online: http://www.popularmechanics.com/blogs/technology_news/4258963.html

Ramchurn, S., et al, 2004. "Devising a Trust Model for Multi-Agent Interactions using Confidence and Reputation," *Applied Artificial Intelligence*, 18, pp. 833-852. Online: http://users.ecs.soton.ac.uk/nrj/download-files/jaai04.pdf.

RoboCup Rescue, 2008. *Rescue Agents.* Online: http://www.robocuprescue.org/agentsim.html

Wikipedia, 2008. *Autonomous Robot.* Online: http://en.wikipedia.org/wiki/Autonomous_robot.

Wilensky, U., 1999. *NetLogo*, Center for Connected Learning and Computer-Based Modeling. Evanston, IL: Northwestern University. Online: http://ccl.northwestern.edu/netlogo.