About Trust (in Autonomous Systems)

Patrick Van Hoeserlande

What if to need to find a lost person in a defined area, you have access to three different systems: a person, a dog, and a machine (in this article, I use the word machine for an autonomous system that can perform a complicated task). Who or what would you trust to tell you when they find that lost person? Would you trust a person, a dog, a machine? After one, 100, 1000 successful training sessions? A search and rescue certified person, dog, or machine? Which system would you trust if it came back empty-handed 'telling' you that there is no lost person in the area? What system would you trust knowing that the dog has a superior smell and the machine multiple sensors? Which system would you engage when that lost person is a family member?

These hard to answer questions show trust is a complex concept with hard and soft components. Although advantageous, hard data alone will not be enough to convince most humans. You earn trust; most of the time through long relationships wherein somebody or something proves to deliver what is expected. However, sometimes a single act or word of mouth may be sufficient to earn somebody's trust.

Although there may be circumstances wherein people, leaders or operators, may gain trust in autonomous systems through quick fixes, these instances will be rather the exceptions and certainly not an element for planning. In the case of autonomous systems operating out of the reach of the operators (examples underwater drones, over-the-horizon systems), the trust challenge is even greater. With positive experiences with autonomous systems as the best road towards trust, we can use a wide spectre of possibilities to bridge the gap between what these systems can and what they are allowed to do.

Before diving into these possibilities, we need to point out that the interaction between autonomous systems and human operators can be based on under- and over-reliance (resulting from under- or overtrust). Under-reliance means that the human does not take full advantage of the machine's capabilities possibly leading to safety issues, as humans might become overloaded causing erroneous behaviour. An extreme form of under-reliance is a situation in which the user does not accept the system at all. Overreliance means that the human allows the system to act autonomously on a task, although the system is not capable of doing that task. Overtrust is a dangerous condition to be avoided for critical systems.

One thing is certain, without extra efforts, autonomous systems will very slowly gain our trust leaving the opportunity open to more risk-taking adversaries to employ them in ways we do not feel comfortable. ISIS has famously deployed armed drones in many

of its attacks and Russian soldiers have deployed ground robots to Syria. There's no indication that any of these machines were fully autonomous, but the manner in which their operators used them suggests that was more of a technical - rather than ethical – barrier.[1] A low percentage of chance to finish the job may be made good by employing a high quantity of systems and still create an effect at a relatively low cost. How must trust do you need to have in one system if the mission success rate can be made good by quantity. Joseph Stalin's citation that quantity has a quality all its own stays valid with modern technology.

DESIGN FEATURES

Although current technological developments look very promising in providing a high degree of autonomy, trust may be maintained through the requirement for the so-called "human in/on the loop" (e.g., weapon release criteria, legal obligations related to the safety of navigation). Even if this might limit the autonomy of the system at the start, this requirement could be relaxed as operators and leaders learn to trust the decisions of the system. Humans are bad at detecting rare faults, meaning that a high-quality autonomous decision-making process endangers the effectiveness of the human-in-the-loop setup.

Sometimes 'halo' and 'horn' effects bias our judgements about people and their actions. These effects have also their impact on human-machine trust. We could exploit these effects by making sure that autonomous systems are good-looking.

Another effect that supports increasing trust in autonomy is anthropomorphization. By giving names, we consider non-human subject looks like it has a face, we would like to be friends with it, or we cannot explain its unpredictable behaviour. Whatever the reason, it creates a bond and with it comes a level of trust. To exploit this effect, we should stimulate operators to give their systems names.

Similarly, the ability of a machine to communicate and interact with operators in a human-like way by 'facial' expressions and voice plays on the soft factor to augment trust.

EDUCATION AND TRAINING FOR OPERATORS AND LEADERS

It is important to educate operators and leaders in the interaction with new technology and to enhance the experience through training to build the right level of human trust in employed systems.

The principle "Train as you fight – Fight as you train" should be incorporated as much as possible. Increased processing capacity, alongside the adoption of open, networked, architectures will enable the use of simulators for the vast majority of training needs,

---

[1] Soldiers Don't Trust Robot Battle Buddies. Can Virtual Training Fix That? - Defense One, 02 Dec 20.

increasing frequency and efficiency. However, for as long as operators need to operate close to in-theatre autonomous systems, there will still be a need to conduct live training to acclimatise personnel. There will also be an enduring need to validate simulator modelling with live training and operational assessments.

In his paper "This Is My Robot. There Are Many Like It But This One Is Mine," Major Yurkovich argues that "inability to (a) understand artificial intelligence (AI) and (b) train daily, will compound to create an atmosphere of mistrust in valuable systems that could otherwise improve the lethality of Infantry Marines." The key to building that trust might be allowing operators to help train the AI-powered machines that serve beside them, as opposed to just handing a soldier, Marine, or airmen a robot and sending the pair off to war together. "Teaching and developing AI agents within a simulated environment by the end user indicate there is the potential for better trust in the AI agent by the end-user when placed as a teammate" within a human-machine team, Yurkovich wrote. This is an approach called interactive machine learning (see also my story 'BLEU BREACH').

M&S WITH BOUNDARY EXPERIMENTS

Countless simulations supported by experimentations and real-life demonstrations of its predicted behaviour in extreme situations of the operational envelope can proof its trustworthiness. The extremer and less certain these situations are, the higher the impact of such demonstrations.

PEER AND THIRD-PARTY VALIDATION

Verification and validation of the individual systems and the collective must not be limited to the initial commissioning but regularly executed. NATO-wide information sharing will offer great benefits with greater access to peer validation and internal data testing. Another option is to create a NATO body for validating autonomous systems.

HUMAN-MACHINE TEAMING

Although very promising, machines are not a panacea, certainly not soon, and complementary approaches are needed. The employment of autonomous systems will be directed on the dirty, dull, dangerous, and difficult (4 Ds), while trust and understanding develop.

Humans are not well suited for these 4Ds jobs, but are more flexible and dexterous, can think beyond algorithms to come up with unique ways of solving problems, are empathetic, have emotional intelligence and more. Whatever the improvements in AI, autonomous systems cannot think beyond algorithms to solve problems creatively, demonstrate empathy and emotion, or invent. Humans are needed to program, repair and teach/train autonomous systems.

Separately, autonomous systems and humans cannot reach beyond their inherent limitations. Together, in concert, machines and humans will improve capabilities beyond the simple sum of the components.

MACHINE FOLLOWING HUMAN AND THE OTHER WAY

A specific type of human-machine teaming is the redundant tasking whereby one does the same as the other in serial. Trust increases as the gap between the measured performance of autonomous and human-controlled systems shrinks or tilts in favour of the first. As with most other approaches, this takes time and may even lead in the beginning to longer operations, but the effect of an autonomous system find for example a naval mine after a manned system cleared the field will be tremendous and worth the investment, even when that was during training or an exercise.

HUMAN AGAINST MACHINE

Taking the former approach a bit further brings us to competitions of human against machines. Comparing the performances of a task that can be repeated by different teams demonstrates the value of autonomous systems. Having a set of tasks to perform has the additional value of showing which system is better in which scenario.

MISSION IMPOSSIBLE

When operators have great faith in the ability of their system, they may be willing to try a mission deemed nearly impossible. A task that others have tried and failed may convince sceptical of the trustworthiness of the machine. These may come in the form of real tasks or as a challenge. A good challenge is a task considered only possible in a not-so-near future.

TRUST BY USE AND IMPROVE

The last possibility is to earn trust by using the systems as early as possible. Every group has early adopters willing to put effort into improving systems that are not fully mature. Through their willingness to test the systems in different situations and giving feedback on possible improvement, the system will mature faster than through conventional validations procedures. This group of operators will also have a much better understanding of the capabilities of the machine. We should involve real operators as early as possible in the spiral development of autonomous systems.

THE REAL CHALLENGE: MACHINE LEARNING

Current machines are not learning on the job, but in preparation of it, the machine, unlike its human counterpart, that leaves is the same one that returns (unless somebody has tampered with it). Their neural network is frozen at the start of the mission and not updated with experience.

The trust question will become more complicated when machines will learn while on a mission from its own experience or even from inter-machine learning. This could be another step in autonomy with incredible benefits but will pose a huge challenge to our relationship with machines.

Looking at the numerous possibilities to enhance trust in autonomous systems, we need to research to determine what approach is best for what combination of scenario, system, operator, and leadership. There is no time to waste. The sooner we start using these systems, even not fully mature, the better.

Note: This article is an elaborated version of my contribution to the November 2020 Innovation Challenge 'Trust in Autonomous Systems'.