People do. Soldiers do.

Soldiers are presented with a choice when facing a combatant on the battlefield. They can either line up the shot with their rifle and pull the trigger, or not. This decision is a simple one–if that combatant is actively trying to kill you, you shoot. If they surrender, you do not.

But how do you define surrender in that case? Is it the moment the combatant puts down their arms? Should they do that movement slowly? Or should they do that movement fast? What happens when a combatant falsely surrenders by putting down their rifle, only to quickly pull out a pistol to continue fighting?

Surrender, in this case, is not simple.

But it gets more complicated than that because surrenders don’t just come from individuals. Take, for instance, the *Nicholas* Incident during the Gulf War (1990-1991). Iraqi soldiers were seen occupying an oil platform, which possessed a state of battle readiness best described by fortified anti-aircraft positions with strategically placed ammunition stocks residing next to said positions.1 The helicopters observing this also saw two people on two separate platforms waving white flags in surrender. Because this was not the whole contingent of Iraqis surrendering, and considering the battle-readiness of the oil-rig, the decision was made to assault and neutralize the threat the Iraqi soldiers posed via attack helicopters. Five of the Iraqis were killed and 23 were taken as prisoners. Of those prisoners, only *one* stated that the personnel on the oil platform had wanted to surrender.

Some of the Iraqi soldiers might’ve legitimately wanted to surrender, but that wouldn’t have rendered their force combat ineffective. Any attempt to make good on those surrenders would’ve meant nothing, because the Navy would not have been able to take those surrendering soldiers prisoner while under fire from the Iraqis that had *not* surrendered.

If a few soldiers–but not all within a given group–surrender, then it is hard to determine if those individual surrenders should be honored or not, especially if honoring those surrenders would put your own troops in danger.

Surrender is a complex beast in active combat on both the individual and group level. It is never a simple yes or no.

This brings us to an important question: how can autonomous weapons systems effectively make the complex decision of whether to accept a surrender or not? Autonomous weapons systems are, in this instance, systems powered by algorithms trained on real world data that possess the capability to identify targets and eliminate them using weapons without human intervention.

The first potentially mass produced autonomous weapons platform is the XQ-58A Valkyrie drone, which recently had its first completely autonomous AI driven test on July 25 of this year.2 It is a relatively cheap, “attritable” (meaning expendable compared to, say, an F-35) design that the U.S military hopes to mass produce,3 along with other vehicles that are either on the drawing board or are being experimented with currently.4

As regards to the algorithms running the Valkyrie, the air force puts it best: “The algorithms were developed by AFRL’s Autonomous Air Combat Operations team. The algorithms matured during millions of hours in high fidelity simulation events, sorties on the X-62 VISTA, Hardware-in-the-Loop events with the XQ-58A and ground test operations.”2 AFRL is the Air Force Research Lab, simulation events are where the algorithms are trained entirely on fictional digital simulations, and the X-62 VISTA is a testing aircraft that is being run by similar algorithms and is collecting as well as training on valuable real world data. Finally, Hardware-in-the-Loop refers to training the algorithms on digital simulations that are fed real-time real world data. Overall, the project is very ambitious and seeks to have the Valkyrie ready for combat in just a few short years.3

With the understanding that AI-driven autonomous weapons systems are coming within the next few years, let us return to the concept of surrender.

Let us say, for example, that the Valkyrie is flying a combat flight with an F-35 on an air-support mission. The adversary is neer-peer (a politically correct term that refers to the rivals of the United States, such as China and Russia) and their troops are holding a position on a mountainside which is being assaulted by U.S troops. Those very same troops have called in the airstrike that the F-35 and the Valkyrie are a part of.

But then the neer-peer troops wave the white flag to the U.S forces. This initiates good-faith surrender negotiations; however, the F-35 and the Valkyrie have no idea that this is happening in this scenario because their communications are going to be jammed (neer-peer adversaries *will* have this capability; it's just a powerful radio transmitter that broadcasts across all frequencies). Therefore, the F-35 and the Valkyrie cannot communicate with each other and cannot be told that the position they were ordered to airstrike is already surrendering.

Via the extensive sensor suite of the F-35, the pilot makes the decision to not airstrike his target, given that he can see the surrender ongoing on the ground. The pilot understands what he sees, but the Valkyrie doesn’t. Its last orders were to airstrike the target, and within the communications jamming that’s occurring, neither the F-35 nor troops on the ground can tell it to break off and return to base.

The Valkyrie bombs the surrendering soldiers.

This is not an optimal outcome.

“But,” you might ask, “Why not hardcode the Valkyrie to return to base upon cessation of communications via jamming? Wouldn’t that prevent the incident above?” Yes, yes it would. But then what would be the point of operating such an aircraft if the second their comms are jammed they stop functioning as warplanes? The military would never accept such a limitation.

The only way to avoid the above scenario is to train the AI on machines like the Valkyrie to recognize and understand the nuances of human surrender, potentially without auditory data and with primarily visual data from sensors and cameras. There would be numerous challenges in designing such a system; however, with enough time and funding, they could be overcome, but not in a way that could account for the complexity that such a system would encounter on actual battlefields.

What happens, for instance, when you have enemy soldiers who fake surrender–and abuse the new surrender detecting algorithms we have trained autonomous weapons platforms on–in order to take down said platform? We encounter the same problem with ordinary soldiers and fake surrenders, but as with the *Nicholas* Incident mentioned prior, the judgment of those soldiers is usually sufficient to detect any lies most of the time.

So then, how exactly do you determine, from the standpoint of visual and audio data, what a proper surrender is and what a fake surrender is? Is it by an arbitrary number of soldiers in a group laying down their arms? What if 99% of a company lays down their weapons, but 1% continues fighting? Do you accept the surrender of the 99% as being proper, or is it invalid because of the 1% that kept fighting? What happens when the autonomous weapons platform has to make a decision based upon limited information, such as seeing all visible soldiers surrendering but knowing that there are more soldiers inside the building that all the visible ones are standing on. Are those two groups considered the same in the eyes of the machine’s AI? Does it then accept the surrender of the visible soldiers and assume the soldiers inside haven’t surrendered, or does it accept and assume the surrender of all soldiers visible and inside the building?

When it comes to the future of AI powered war-machines, these are questions that *must* be asked. You can train algorithms to eliminate the enemy via autonomous weapons platforms without any humans in the loop. But can you train algorithms to recognize surrender? That’s something that’s by far more complex than an algorithm simply recognizing an enemy combatant and neutralizing them.

The future is almost here for autonomous weapons platforms, and we cannot overlook the ethical issues they raise for the concept of surrender in armed conflict.

**References**

1Robertson, H. B. (1993). The Obligation to Accept Surrender. *Naval War College Review*, *46*(2), 103–115. http://www.jstor.org/stable/44642452

2Air Force Research Laboratory Public Affairs. (2023, August 3). AFRL AI agents successfully pilot XQ-58A Valkyrie uncrewed Jet Aircraft. Air Force. https://www.af.mil/News/Article-Display/Article/3481081/afrl-ai-agents-successfully-pilot-xq-58a-valkyrie-uncrewed-jet-aircraft/

3Garamone, J. (2023, September 7). Hicks discusses Replicator Initiative. U.S. Department of Defense. https://www.defense.gov/News/News-Stories/Article/Article/3518827/hicks-discusses-replicator-initiative/

4U.S. Naval Forces Central Command Public Affairs. (2023, November 2). Exercise Digital Talon advances unmanned lethality at sea. United States Navy. https://www.navy.mil/Press-Office/News-Stories/Article/3576850/exercise-digital-talon-advances-unmanned-lethality-at-sea/