The "Bright Green Line" of Responsibility

In the near future . . . . New York SWAT teams are equipped with "smart" rifles that prevent the shooting of unarmed targets. Hostage shootings by SWAT personnel immediately drop dramatically followed by steady increase in successful outcomes and a minor rebound in hostage shootings by SWAT personnel. Studies reveal that the most successful SWAT personnel have adopted a strategy of "shoot everything that moves and let the gun sort it out" and that it would take a better than ten-fold increase in the error rate of the rifles before this wasn't the best strategy in terms of outcome. The "smart" rifle has become the arbiter of who lives and who dies.

In Los Angeles, SWAT teams begin to take advantage of "armed telepresence" using modified DARPA disaster-relief robots. Particularly popular/effective are the "pre-targeting" and "aim-correction" functionalities which provide inhuman speed and accuracy to even the rawest recruits. Unfortunately, hostage shootings by LA SWAT personnel rise as the new assisted-human speed outpaces unassisted human judgment. Using the "smart" rifles would solve that problem but effectively take the human entirely out of the loop. The "killer robot" will have arrived.

The robotization of warfare is generally regarded and treated as a slippery slope. Thus, calls have recently been made for international principles/executive orders that machines should not be making decisions that are harmful to humans. Interpreted precisely as written, this should be inarguable for some years to come. Unfortunately, however, the interpretation of the term "making decisions" has been dangerously imprecise – an abstraction that "leaks" horribly.

Is the "smart" rifle "making a decision" or is it merely executing an algorithm that a responsible human being has approved? Are "pre-targeting" and "aim-correction" functionalities decisions or merely implementations of previous human orders? Indeed, is the "killer robot" (blindly following orders with reasonably predictable results and equipped with a "kill switch" and a virtual guarantee that it will save far more lives than it will end) somehow more "responsible" than an autonomous car?

The critical distinction lacking in DoD Directive 3000.09 and the Human Rights Watch's "Losing Humanity: The Case against Killer Robots" is the distinction between "operational" autonomy and what could be called intentional or volitional or "command" autonomy – between "tools" with no intent and responsibility and autopoietic "entities" with intent or goals. James Moor speaks of a "bright line" between machine ethics and the full ethical agency of the average adult human but arguably there are two lines.

There clearly is a bright red terminator crossed when a tool is able to learn deeply or self-program and is not completely transparent. Such "tools" are no longer predictable nor can they be reliably manipulated – and, if complex and advanced enough, probably best regarded as wild animals or young children. Eventually, possibly, far enough in the future, there will be a bright green finish line where machines can be given responsibility because they ***will*** reliably be responsible ethical entities – but that is an argument for another day.