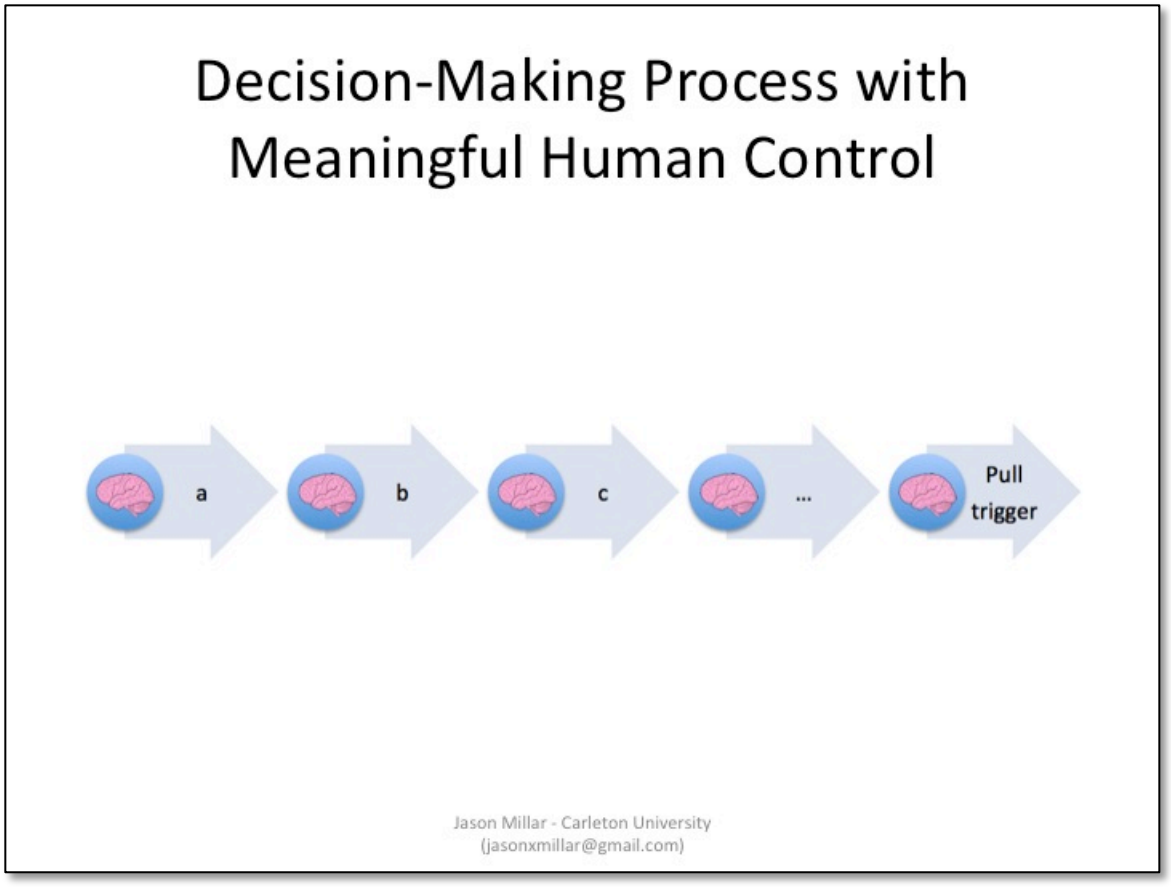
Expert Testimony provided by Jason Millar to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWs), held within the framework of the UN Convention on Certain Conventional Weapons (CCW), Geneva, Switzerland, April 15, 2015.

Thank you mister ambassador. First of all I'd like to extend my sincere thanks to the meeting chair for inviting me to speak here today. I'd also like to thank all of the delegates and representatives from civil society for attending this meeting, and for showing a genuine interest in what is clearly an important global ethical issue.

I am an engineer, and in my early career I worked on the design and manufacture of aerospace electronics for use in military and commercial applications. I was embedded in the world of dual-use technologies for some time. For the past several years, however, my attention has turned to researching the ethics of automation technologies. I've been asked to bring some of that knowledge to bear on the issue of lethal autonomous weapons.

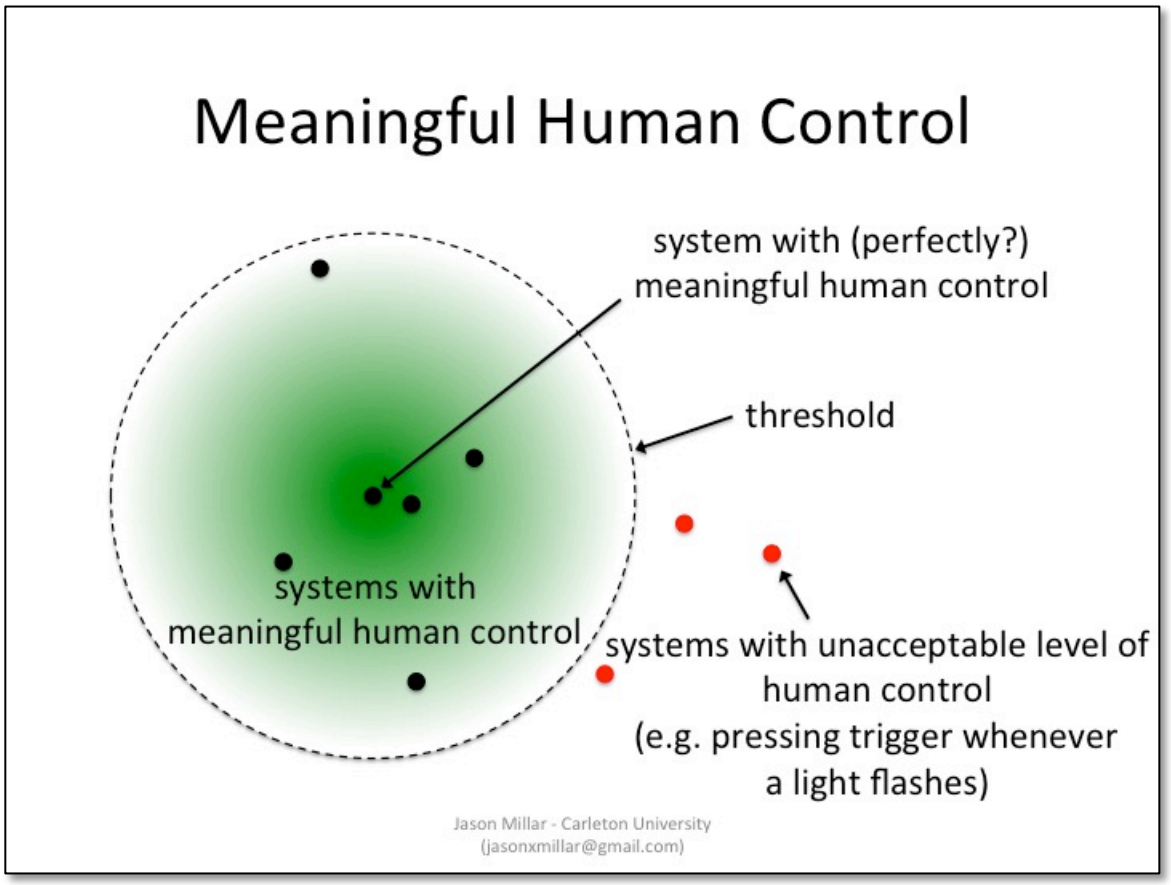
Today I've chosen to focus my remarks on one of the central ethical issues we are confronting with lethal autonomous weapons, and that's the question: To what extent must we keep humans in the loop in order to maintain whatever is ethically required of us when using lethal force? What I see half way through this meeting, is that there is an emerging sense that we must maintain meaningful human control over automated military technologies. And I agree with that. I also agree that meaningful human control is a concept that can work in a policy context. It captures an important ethical aspect of automated weapons technologies by focusing our attention on the humans surrounding them. My goal today is to provide a brief overview of what I think is one of the complexities involved in designing technology for meaningful human control and then suggesting a way forward.



The debate over meaningful human control is partly about the ethics of automating the decision-making process that leads to and includes the decision to use lethal force. It's about whether or not we can take the human out of the process, and still maintain whatever ethical features are necessary for that decision to be ethically justifiable. A requirement to maintain meaningful human control is based on the premise that if we significantly automate the decision-making process, and largely removing the human from it, we lose whatever is necessary for the decision to be ethical.

Now, it seems to be the case that when we talk about meaningful human control, we're not merely worried about being able to satisfy the requirements of international humanitarian law, though that is certainly necessary, nor are we merely interested in having a human pull the trigger when prompted. Meaningful human control involves something more. As I understand it, it's about maintaining something "uniquely human"

in the decision-making process; something that lethal autonomous weapons would simply lack.



Let's imagine we adopt the idea of maintaining meaningful human control as a requirement in the design of semi-autonomous weapons to see what work the term can do for us. So I'm asking you to imagine a world in which we produce semi-autonomous weapons, but when doing so we have decided that we must ensure that the humans operating those weapons systems are in meaningful control of them.

Now, in practice, it would be unlikely that we would be able to design systems such that every user of every semi-autonomous system had the same level of control over the system. Different design choices will lead to different kinds of interactions between the operators and the technology. So we can imagine meaningful human control looking something like it's depicted in this slide. For some imaginary system we might achieve

"perfect meaningful human control", whatever that might be, as indicated by the point at the centre of the dotted circle. But we shouldn't expect to. More likely, assuming we get the design right, we'll have systems where humans are in varying levels of meaningful control, represented by the black points inside of the dotted circle. The green is meant to indicate that, according to this model, the closer you get to the centre, the more meaningful that control is, ethically speaking. Then we'll have systems where there are unacceptably low levels of human control, say, a system in which the human operator simply pulls a trigger when a light flashes, indicated by the red points outside of the circle.

In this picture of meaningful human control, there is a threshold separating two kinds of systems: those that are ethically permissible, and those that are not. And those two kinds of systems are differentiated by their operators' relative ability to maintain some pre-established level of meaningful human control.

I think we have good reasons to accept this picture. And if we do, then an important question emerges from it. What kinds of design factors are going to move us closer to the center of the circle, and what kinds are going to move us farther away? After all, if we're going to use meaningful human control as a gauge, we want to be as close to the centre as possible, so we need to figure out how to design systems to sit there.

There are many design factors that would affect where a semi-autonomous system might land in the picture I've sketched. Clearly, automating all of the critical functions contributing to the decision-making process, including the decision to use lethal force, would put us outside of the circle. That seems relatively straightforward.

But I'd like to focus on a body of evidence emerging in the field of moral psychology to make the case that our efforts to maintain meaningful human control will be slightly more complex than simply keeping the human in the loop. I'd like to suggest that meaningful human control could be undermined quite unexpectedly by automation technologies, even if we keep humans in the loop in very non-trivial ways. I'll repeat that,

just to make sure the point I'm about to make is clear: the evidence I wish to describe suggests that we could keep humans well in the decision-making loop, and still drift towards the edge of the circle of meaningful human control.

Researchers in moral psychology have, for the past ten or fifteen years, been uncovering some surprising facts about our ethical decision-making capacities. They have produced robust findings demonstrating that seemingly unimportant situational factors can significantly impact our ability to make consistent ethical decisions, the kinds of consistent decisions that would seem essential for maintaining meaningful human control while operating semi-autonomous weapons systems.

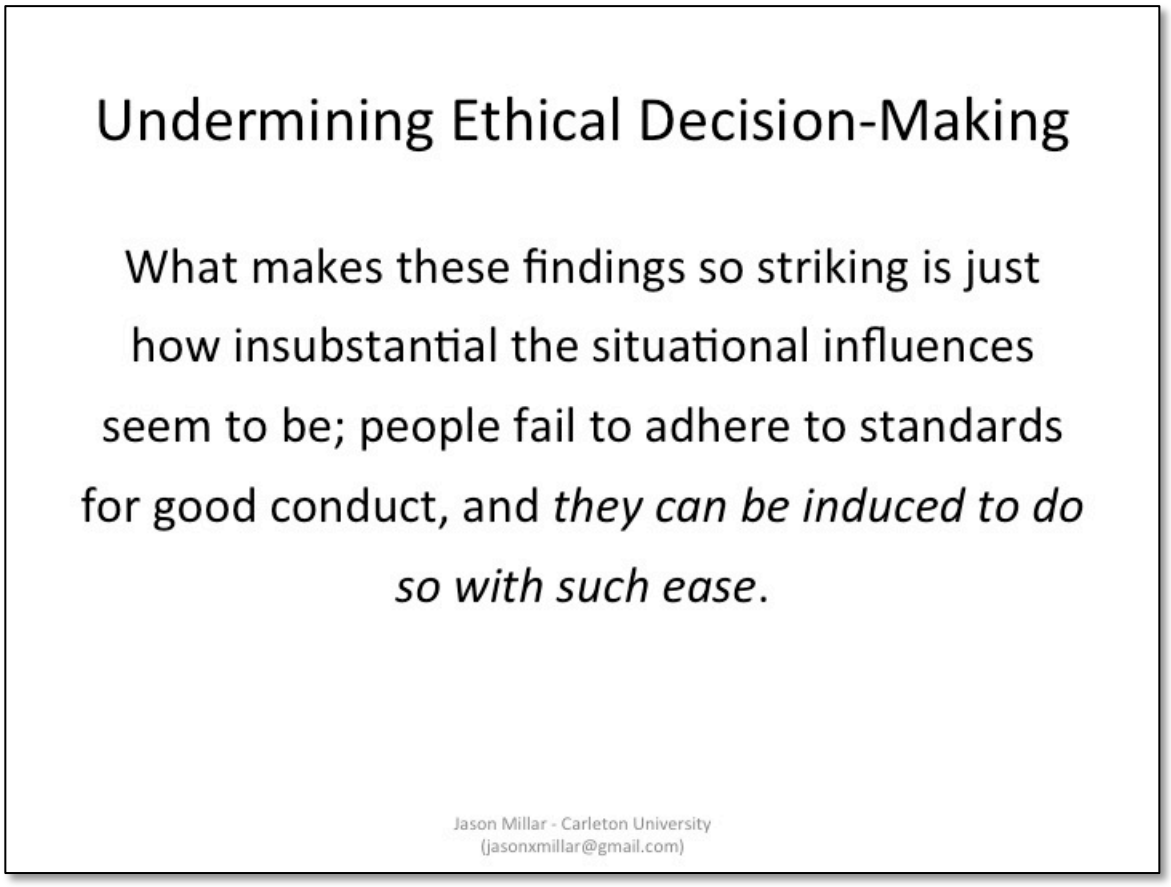
Let me describe a few examples of this research.

In one study, participants were shown a video depicting an ethically questionable behaviour, and they were asked to judge how "bad" the behaviour was. However, just prior to watching the video, half of the participants were first shown a short funny clip from a popular late night comedy show. Those participants who were "primed" with the funny video clip, tended to judge the second video, the one depicting the questionable behaviour, less harshly than the other participants. Put another way, the participants who did not see a funny video, judged the behaviour more harshly. The first video, one would think, should have little or no effect on a person's ability to make subsequent ethical judgments. Yet it does. We can fairly describe this, I would argue, as an erosion of human control over the participants' ethical decision-making capacity.

In another study, researchers found that 80% of participants stopped to help an injured man, so long as the ambient noise levels were low. That sounds like good news. But when a power lawnmower was running nearby, that number dropped to 15%. In this case, ambient noise levels affected the participants' ethical decision-making. This too could be interpreted as an erosion of human control.

In yet another study where participants were shown a video and asked to rate the "badness" of its ethically questionable content, participants sitting at a dirty desk judged

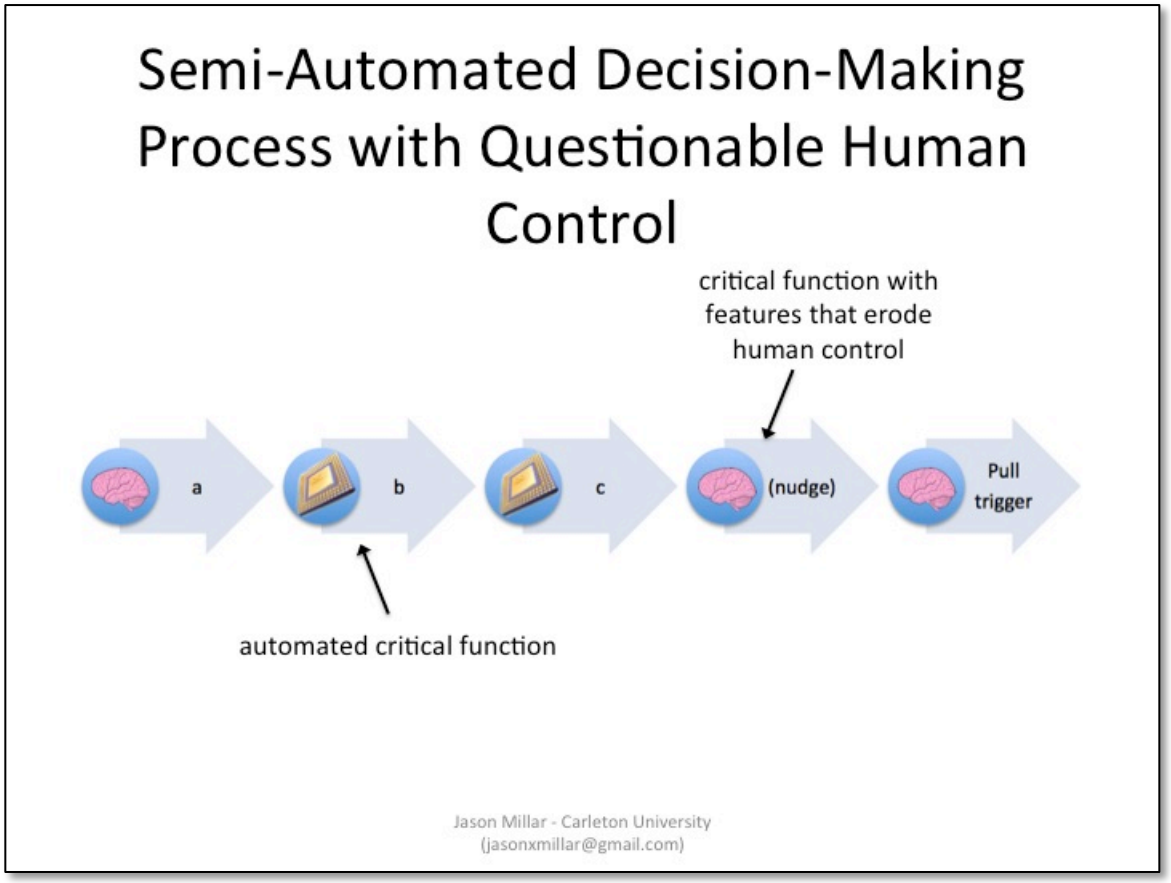
the depicted behaviour more harshly than those sitting at a clean desk. Again, seemingly irrelevant situational factors affected the consistency of participants' ethical decision-making. And again, we might consider this an erosion of human control.



There is no shortage of these kinds of studies. The findings are repeatable, and consistent across cultures. And if you aren't familiar with them, they might surprise you. These findings challenge the common assumption that ethical decision-making is entirely under the control of the individual. To paraphrase a prominent researcher in the field, "What makes these findings so striking is just how insubstantial the situational influences seem to be; people fail to adhere to standards for good conduct, and they can be induced to do so with such ease."

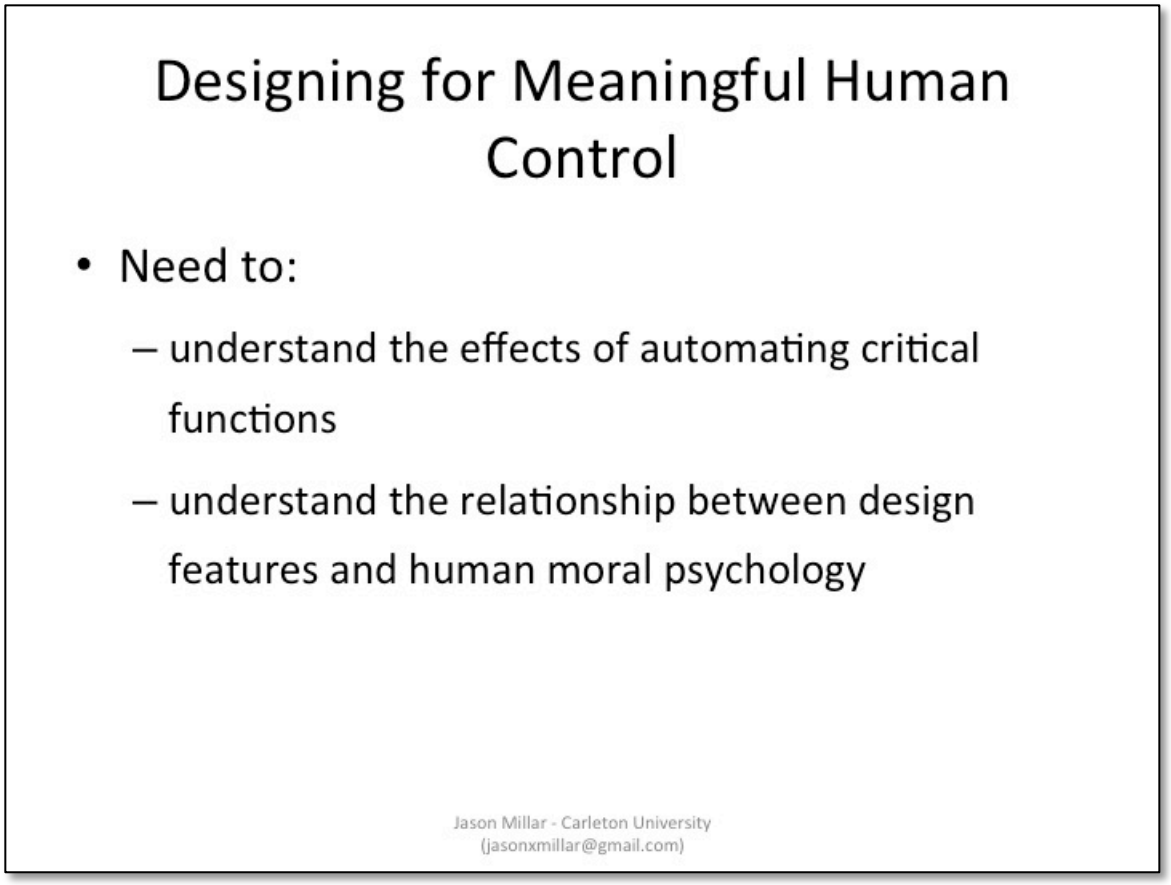
Equally striking are the implications for meaningful human control in the context of semi-autonomous weapons. If we automate critical functions in the decision-making

process leading up to the use of lethal force, the particular design features of those automated functions—that is, their user interfaces with all their noises, graphical representations and controls that go with them, their physical packages, and the physical space within which they are used—all of these things could induce psychological effects like those in the studies just described. That could lead to an erosion of meaningful control. Now, as one of my fellow experts explained on Tuesday morning, the military solution to meaningful human control could take the form of rigorous training and strict adherence to standards of conduct. However, it is not clear whether, or to what extent, designers and engineers are accounting for these kinds of moral psychological findings when they design and test semi-autonomous weapons. Those effects could act to undermine an operator's training. To be fair, it's also not clear that these effects will have a major impact, but without the evidence, we should add them to our list of concerns.



If seemingly insubstantial situational factors, such as lawnmower noises, priming events, framings, dirty desks, and the like, can erode human control over ethical decision-making, we have reason to believe that Human Computer Interface design choices might also erode human control in the lethal force decision-making process. For example, a semi-autonomous weapon that nudges an operator to slightly increase, or decrease, the value determination of a particular target, perhaps in a way that affects necessity and proportionality calculations, creates a bias that could carry through to the final decision on lethal force. Combine these little erosions with enough automated functions in the critical path, and you could find the system drifting toward, even crossing over, to the outside of the circle. If there's a takeaway here, it's that each little nudge erodes meaningful human control in a way that moves us out from the centre of the circle towards the threshold of ethically unacceptable technology.

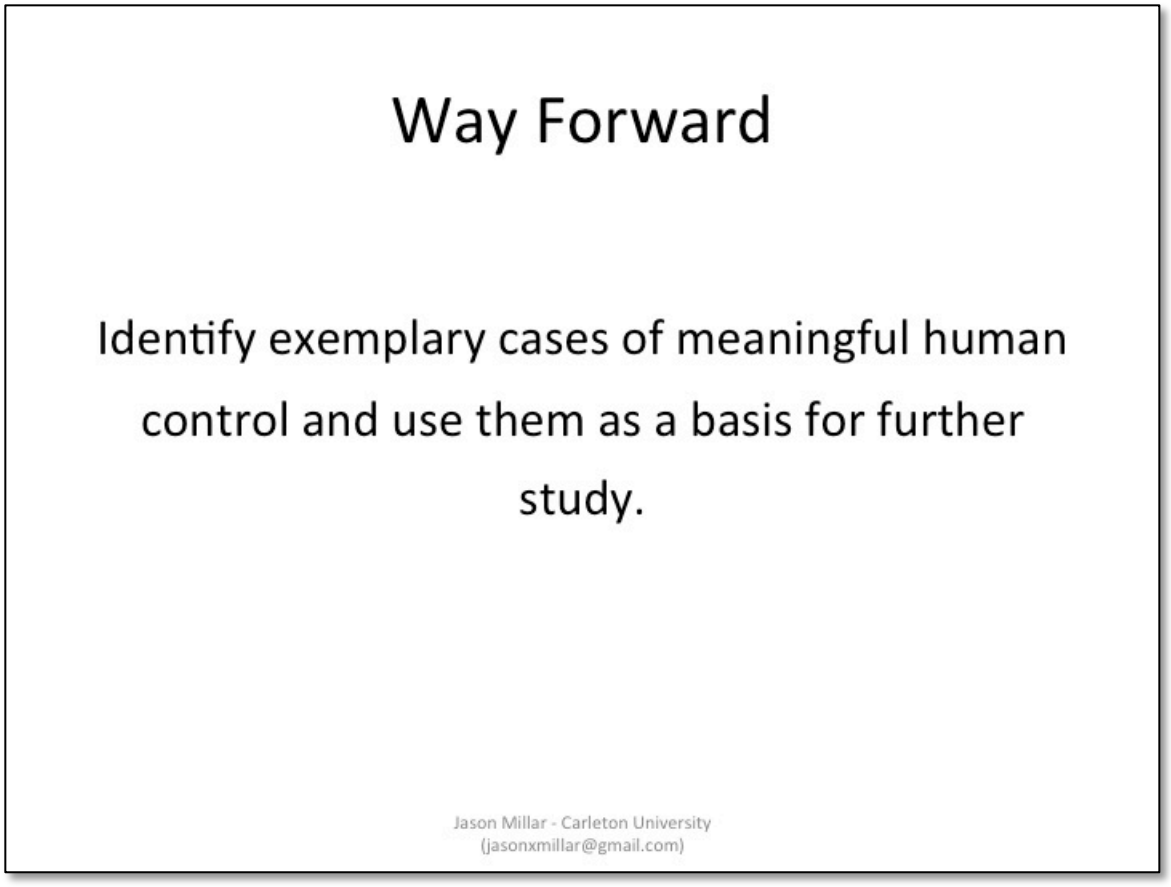
Here's where the dual use comes in. Complex semi-autonomous weapons systems could borrow many dual-use components originally developed for non-lethal systems. Those components could include a number of interfaces, algorithms and other functions in the critical decision-making path. But, non-lethal components will have been developed for use in situations where the ethical stakes are relatively low. Thus, they will likely not have been designed with any consideration having been given to their tendency to nudge operators.



So, if we are going to design for meaningful human control, in addition to understanding the effects of automating critical functions, we are going to have to invest time and effort trying to understand the relationship between design features and human moral psychology, so that we keep systems within the circle.

However, I think it's possible to do this. It is possible to design semi-autonomous weapons, and to establish meaningful human control. Of course, a lot of work will have to be done to determine a kind of baseline for meaningful human control. Also, we'll have to figure out how to set the threshold and how to tell when a system has drifted too far from where it should be.

But I'm an optimist. And because of that I'll end my presentation by modestly proposing how we might start to develop that baseline.



The first task, I think, is to loosen our grip, just a bit, on trying to define meaningful human control. Definitions are elusive, and slippery when you finally think you have one in hand. Rather, I propose we look around and find exemplary cases of semi-autonomous weapons systems where operators clearly have meaningful human control, the kind of control that makes possible the ethical use of force, and clearly respects international humanitarian law. By studying actual systems we should be able to start framing a discussion around design standards, and at least take a step forward. I suspect with some hard work the definitions will eventually start to coalesce.

Thank you.