Penn State Journal of Law & International Affairs

Volume 7
Issue 3 *Symposium Issue*

April 2020

Autonomous Systems & the Ethics of Conflict

Micah Clark

Claire Finkelstein

Oren Gross

ISSN: 2168-7951

Penn State
Journal of Law & International Affairs

2020 SYMPOSIUM ISSUE

AUTONOMOUS SYSTEMS & THE ETHICS OF CONFLICT

Moderator: Ben Jones
Panelists: Micah Clark, Claire Finkelstein, and Oren Gross

Ben Jones: Welcome everyone to the second panel today, Autonomous Systems and the Ethics of Conflict. My name is Ben Jones. I'm the Assistant Director of the Rock Ethics Institute here at Penn State. I've been tasked with moderating today's panel. We have a distinguished group of panelists who I'll introduce: Micah Clark, here closest to me, is the Senior Scientist for Autonomy, Artificial Intelligence and Cognitive Science at the Applied Research Laboratory here at Penn State. Previously, he was a program officer at the US Navy Office of Naval Research and he holds a PhD in Cognitive Science from Rensselaer Polytechnic Institute.

Next is Oren Gross, he's the Irving Younger Professor of Law at the University of Minnesota Law School. He's an internationally recognized expert on international law and national security law. In addition to his academic work, he previously served as a senior legal advisory officer in the international law branch of the Israeli Defense Forces' Judge Advocate General's Corps.

Ben Jones: And then furthest from me is Professor Claire Finkelstein. She is the Algernon Biddle Professor of Law and Professor of Philosophy at the University of Pennsylvania. She founded and is the faculty director of the Center for Ethics and the Rule of Law at the University of Pennsylvania Law School. She's widely published on the laws of war, issues of national security and legal theory.

We're going to be focused on the ethics of using autonomous weapons systems within conflict zones. This is an area where there's not a lot of consensus yet among ethicists, engineers and others working on this issue. When you look at the literature on this particular topic, you'll find some making a very strong case that there should be an international ban on autonomous weapons systems, and that we need to be putting our efforts into making sure that ban is put in place before these systems are unleashed.

On the other hand, you have some that make the case that there's an ethical obligation to use these systems, that when you look at the history of conflict, you find atrocities, you find war crimes, and that in fact if we would transition to these systems, we would end up in a much better place than we are now. So, the goal today is to tackle some of these issues, look at how these systems are being used, whether or not they can be compatible with international humanitarian law, and with our conceptions of ethics. I'm going to keep my remarks to a minimum because I want to give time to our panelists. I also hope that we have some time at the end for Q&A with the

audience. So, with that, I will hand it over to Micah.

Micah Clark: Thank you, so I'm Micah Clark, and I want to say that I'm here speaking as a technologist-I'm not an ethicist, a lawyer or a policy maker, and I'm speaking only for myself, not for anyone else, certainly not anyone within the government. The genesis for today's discussion and others like it has stemmed from the concerns about the undesirable impermissible and potentially unethical uses of technology. Technology that is becoming ever more accessible, not just to world militaries but to non-state actors.

A lot of these discussions have played out in the public under the rubric of killer robots, and in terms of existential threats posed by artificial super intelligences. To be clear, I am much less concerned with the prospect of actual artificial intelligence than with the manifest abundance of artificial and natural ignorance. Let me also be clear that even without US involvement, kinetic and non-kinetic autonomous weapon systems are coming, and some are already here and no treaty will change that. The advantages militarily or from a terroristic perspective are simply too great.

So in the time remaining, I just want to pull on a couple of threads. In particular, I want to look at lethal autonomy just as a straw man in some sense, and the desire for real accountability to frame some of the challenges that face policy makers and technologists. But before jumping into that, let me try to summarize what autonomous systems are from the operator or user's perspective. So the use

of autonomy essentially consists of constrained, context sensitive authorities that the humans are temporarily seeding to the system. So if we want to get more specific, autonomous systems don't have libertarian free will, so you're delegating to them and to the system the authority to decide and act within a given space of possible decisions and actions at a certain level of granularity over a given time window using predefined certain types and sources of information, relative to certain environmental situational conditions and context, so as to satisfy some set of explicit responsibilities. They are subject to you, the user's, expressed intent and the enforced behavioral constraints within the system that reflect policy and law.

Micah Clark: I know that, that is a mouthful and probably an ear full, but in the end the import is that what you are doing is you are separating the authority to decide and act from accountability for those decisions and actions. That is the point of autonomy; you are giving over to systems the authority to decide and act on your behalf, so that in some sense you don't have to. But there is a desire for accountability within our systems. In fact, I would say that, and much I think of international law is based on this, that there is a presumptive compact affirming the sanctity and intrinsic value of human life. That the decision to intentionally take human life in the use lethal force is of such consequence that we as society, as a species, desire to be able to hold decision makers accountable for the choice. But, of course, you can't hold a toaster accountable, you can unplug it, you can't punish it.

Micah Clark: Regardless of how we label our technologies or as Dave Atkinson mentioned how much we anthropomorphize technologies, they are artifacts, they are not moral agents, whether under the law or under theology or under philosophy of mind. We read those properties into them, but they aren't real. And this is actually, I think the reasonable view taken in both the DoD Directive on autonomy and weapons systems and in the revised DoD laws of war. We can talk about whether that is a viable position 10 years from now later, but this lack of accountability for artifacts motivates us to reserve for humans the authority to choose lethal actions to keep humans in the loop and to ensure meaningful human control over autonomous systems.

The problem is it is not clear that accountability can be maintained. This is true for present systems and certainly true for future system. So, to briefly explain the problem: autonomous systems are imperfect, they are limited in their scope, their abilities, their understanding. Their decisions and actions are dependent on unforeseeable environmental and situational conditions, that are in situ to where the system is, not where you the operator are at, and certainly not where you, the decision maker, were when deciding to employ this system.

Their internal operations are often unintuitive and exceedingly difficult for humans to internalize accurately. That is okay because it is not like we sufficiently train operators on any of these things. Operators lack the tools and the situational awareness necessary to anticipate both system behavior and what the consequences of that behavior will be in the

situation as it unfolds. In fact, from the military perspective, we lack the tools and even the basic methodology to assess the appropriateness, sufficiency, proficiency and risks of using an autonomous system relative to some given unfolding situation. We do it based off of rough estimates and gut feel and what the vendor says the system can do, but there isn't a methodology there that says these are the trades, which is the kind of thing you would want if you're going to make well-reasoned, well justified decisions on what kinds of systems you're going to employ in the moment.

Micah Clark: So in the end there is this huge chasm between what human operators believe and what is actually true. Yet as these systems become more capable, we will delegate more and broader authorities to them. This will increase not just the system's independence, but the causal distance between the decision to delegate, and the consequences of that decision, and certainly well beyond the causal distance that any human operator could be reasonably expected to foresee. In addition, autonomous systems will be and are being tasked with achieving objectives for multiple simultaneous missions, for multiple simultaneous users. These systems are expected to satisfy between them to choose exactly which of the goals to pursue and how.

The net result of this is near impossibility of actually tracing and determining responsibility for action. So taken as a whole, it is not clear that we could rightly hold any one person accountable, that there would be someone with the sufficient cognizance and fore knowledge that we could point at and say you are ethically

and legally responsible for what this system did.

Micah Clark: So far we have really been talking the context of what might be described as monolithic autonomous systems. If we were to expand out to include swarms, psychological operations, cyber, then the law of war principles of distinction and proportionality become even more problematic if not unsurmountable, which I think is the case for much of cyber. So what is the impact of this? Well, for policy, there are limits to human accountability in the use of autonomous systems, at least for what you could rightly hold someone accountable for.

As a technologist, I need policy to deal with practical issues not just the abstract. So take for example, an operator working with an automatic threat detection and targeting system. The kind of systems that exist, and are widely fielded it around the world today. What matters to me in designing how the human machine team ought to work for that, and what the role and responsibilities are for the operator boils down to relatively simple questions, but much closer to metal than policy. What constitutes sufficient evidence of a threat? What information is relevant to their determination? And what of that is available to the operator? How is the threat determined? What are the relevant thresholds for competence and certainty? What are the uncertainties associated with targeting and force application? What constitutes acceptable risks within that trade space, if the operators would make a decision? How much decision time is going to be available to the operator and

is that sufficient given the legal jeopardy that they are in making it? How much time is there between the operator's decision to act and the actual application of force? In that window of time, a lot of unexpected, unpredictable events can occur that will fundamentally change what is appropriate.

Micah Clark: What is the role of assistive technology? Is trust and reliance, and say, a threat estimation system, is that a valid justification for an operator's decision to act? If so, is the operator required to know when such trust is appropriate, and how would they know that? If it is not sufficient justification, what additional information do they need? And what decision process is required? And are such things available either today or in our future systems? So these and many other questions like it are essential to justifying an operator's decision to act or not.

Now, autonomous systems will and do play a critical role in the use of force. Yes, apart from reactive defensive systems, we can preserve the use of force decision as the prerogative of the human. But that is one small piece of it. If we look at detection and targeting, threat risks, course of action assessment, mission prosecution and so forth, it is technology and increasingly autonomy that is performing those functions. Fire forget may work for ammunition that has a total mission lifetime of thirty seconds, five minutes, ten minutes, if we're talking about a system that is going to do a loiter and interdiction over a three day, three month, three year period, then the decision made at one moment in time that yes, we're

going to use it. That person can't be held responsible for all that might follow.

Micah Clark: So we need policymakers and technologists to work together to ensure that the results, in choreography of mission teams makes sense. For human operators, they only have the resources we give them, whether that is time, information, assessments, predictions and so forth. The question is what do they need if there is to be real accountability? And is that something we can give them? For autonomous systems themselves, look, they need a codified calculus for right decision making, whatever that might be. Policies that are based on reasonable persons and other such legal fictions aren't much help. Autonomous systems are not reasonable persons. They are compliant automatons that do exactly what they were designed and told to do. They have no common sense. They have no understanding of the world. They are going to do what they were designed and told to do, even if it is completely wrong.

So switching to technology, despite what you might expect, I actually don't think the key challenge for technology is ensuring ethically correct behavior. That has certainly been looked at. There are various proposals for how that might be done either through verification or ethical governors, there are approaches to doing that. I think the hard challenge on the technology side is characterizing the proficiency and performance of these systems. So as technologists we need to be up front about what information is used and what is ignored, how confidence and certainty are treated, how decision options within the

system are conceived, evaluated and selected. What competence is assumed within the planning process and what performance is actually achieved.

We need to understand in reality the performance envelope on these systems relative to system's belief and as well as our beliefs and actuality in the world. Well, that sounds straight forward, it is exceptionally hard, especially for active learning technologies. The performance envelopes are highly nonlinear, they are always perverse corner cases that no one has thought of, and these failures of imagination are the historical Achilles heel both in the intelligence community and in the engineering world, and will continue to be.

Micah Clark: So I don't think we will ever know the behavior of our systems across the totality of the problem space, that is acceptable versus unacceptable performance relative to a variety of acceptability criteria including ethical behavior relative to either all possible situations or a given under a specified situation that we're trying to evaluate. I don't think we will ever know that. I don't think it is theoretically possible to know that. But we might be able to at least characterize the performance boundary between known acceptable and simply unknown. Does not mean the system is going to do something crazy, does not mean it is going to go wild and go off the reservation. We just don't know, we have not tested that, we have not looked at that yet, but we know we can accept its performance within these operating conditions and everything outside of that, simply unsure. If we know that, that is a

key critical enabler for having appropriate delegation trust and reliance in autonomous systems.

Ultimately, however we parcel out responsibility and authority between the human team members and the systems, we need that piece. We need to understand not just what the systems will do, but that is how they operate and how they behave appropriately for the kind of situation we're sending them into. And with that I'll yield the, probably zero, time I have left.

Ben Jones: Thank you Micah. Oren.

Oren Gross: Thank you to the Center and for the Journal for organizing, and to Penn State for hosting. It has really has been a pleasure to listen to the previous speakers. I like the framework that Brian introduced in the first panel, i.e. that of storytelling and thinking about the story. Of course, there's also the question of the timeline or the time frame of the story. Are we thinking about next year, next five years, next ten, fifty or more? So when you are talking about AI, for the non-technologists in the room, there is always the question of what is the future that you envision? Do you envision the Terminator? Is the first image that just comes up? Or is the first image that comes to mind that of R2D2? And of course, even with Terminator, which movie are you thinking about? After all, the Terminator actually turned good at some point.

So let me start off by putting forward my completely non provocative, and I'm sure generally accepted, claim: Humans out of the loop is not necessarily a bad thing. Human

input may not only be at some point unnecessary but dangerous to our soldiers, our civilians and to "their" civilians. In developing that claim, I will note that the question of accountability that we heard quite a lot about is a critical question, but it is not the *only* question. Let me develop the claim.

Oren Gross: Throughout human history, technological advancements have been a paramount factor in creating, maintaining, shifting and destroying military advantage and dominance. Despite that the basic component of the military has remained and will remain, at least for the foreseeable future, human soldiers. At the same time carbon-based human soldiers are increasingly the major limiting factor for operational dominance in conflict. Simply put, human beings are becoming the weakest link in armed forces. When you think about speed of decision making, about big data, how much information is coming in, about the pressures of combat, about our physical capacities and limitations, about our cognitive capacities and about the cognitive burden of decision making, and about emotions, both good and bad, you realize that all of these somehow limit our capacities in making good and timely decisions.

The relationship between humans and weapons has been changing and shifting since the appearance of humans on the face of the earth. Yet, underlying all these changes is the perception that human beings exercise direct, albeit not necessarily full, control over weapons. From the first time that a human hurled a stone at an enemy, he (at least first more likely than she) did not have full control over the weapon. He had direct control, but

not full control. Still we had human decision-making on issues of life and death.

Oren Gross: That did change at some point in time to a sort of partnership model of relationship between human beings and weapons, some sort of complementarity between us and our weapons systems. There are things that computers do better and there are things that humans do better. Computers can do number crunching or deal with big data better. Their speed of response is much faster than ours. They may have total recall capabilities (keeping with the Schwarzenegger theme). They have disembodied intelligence and instant transfer learning so that they can transfer data and knowledge even if one system is destroyed. Yet, there are things that humans do better than machines (at least for now), such as our capacity to engage in ethical decision-making, to adapt to new circumstances, and to show emotions and feelings such as empathy. In addition, juxtaposed with the silicon-based machine's number crunching power, we are endowed with evolutionary cleverness, allowing us to be prune decision-making trees.

But now we speak of the next stage in our relationship with weapons, a stage that some have called de-humanized war, or what I would rather call the post-human war represented by a move to fully autonomous weapons systems that would reduce or eliminate altogether human control. General Allen refers to this as 'hyperwar.' I should note that the term 'hyperwar' has already been used to descrive World War II. But what we talk about now involves machine learning algorithms, artificial intelligence powered autonomous decision

making, advanced sensors, miniaturized high powered computing capacities, high speed networks, cyber capabilities, and things such as autonomous swarms.

Oren Gross: One major effect or result of all of these capabilities and capacities is the minimization of human involvement in decision making. So we might increasingly see as, as General Allen suggested, humans providing broad high level inputs while machines do the planning, executing and adapting to the reality of the mission, and take on the burdens of thousands of individual decisions with no additional human input.

So is that a good thing or a bad thing? Well, it depends on your perspective and attitude. We already spoke about the "responsibility gap," i.e. the offloading of responsibility to what is, in essence, not a moral agent. The offloading of responsibility to what is not a moral agent. Humans should bear the moral responsibility. We want humans to make judgment calls such as decisions about proportionality and be ultimately responsible to life and death decisions. And so we seem to need some sort of human control over the machine.

Oren Gross: Thus, we come to ask to what extent do humans, could humans and should humans maintain control over sophisticated weapons systems? Many of you are familiar with the Observe, Orient, Decide and Act or OODA loop as was developed by the military strategist John Boyd. In the context of human-machine relationship there can be three options. First, a human can be in the loop. Here we speak of machines that are capable of targeting and

striking solely as a result of a human directive. Second, a human can be on the loop where the relevant weapon system is capable of independently targeting and delivering force while under the supervision of a human who retains an override capacity. Finally, a human can also be completely out of the loop, i.e., when a weapon can target and deliver force without any human input or interaction.

When we speak of meaningful human control, we usually focus on level one or two, i.e., human in the loop or human on the loop. But consider the following questions and challenges. Meaningful human control over what exactly? What is it that we need to regulate? Part of the problem here is that we are not entirely sure what the technology is going to look like and as a result, we currently have definitions of autonomy, and autonomous weapon systems that vary greatly. We do not have any widely accepted definition of what an autonomous weapon system is. We also do not have an accepted conception of the exact stage in which meaningful human control ought to be exercised. Is it the stage of developing, programming, designing, or training of the weapons systems? Is it at the stage of developing machine-specific rules of engagement, or the stage of the decision to deploy autonomous weapons systems in specific combat operations?

Oren Gross: Nor are we even sure about what "meaningful" human control actually requires in order to be meaningful. We have some essential elements that we thing should be evaluated when we are talking about meaningful human control, such as informed decisions, sufficient information,

or effective control over the use of the weapons system, but even those are quite general and nonspecific.

But let us assume that we got over all of these obstacles and are able to agree on what "meaningful human control" means. There still remain significant challenges to the very concept of meaningful human control that cast grave doubts as to its usefulness. Before I turn to these challenges at the end of my presentation allow me to remind you all that, as I noted earlier, since the beginning of time, humans have been employing weapons that lack perfect real time situational awareness of the target area. The essence of projectile weapons is that we do not have full control over their trajectory, nor can we suspend or abort the attack after launching them.

We have been discussing the question of responsibility and accountability. It is, undoubtedly, a critical question but it is not the only one. Another important issue is how to minimize the harms of armed conflict to civilians, civilian objects and even to soldiers. What means and methods should we use in order to minimize such harms? And to me the question is if we have means and methods that actually allow us to minimize harm, does it matter whether those are human controlled or whether those are eventually going to be autonomous? If we believe that there are now or that there would be means or methods of warfare that protect humanity better than other means or methods, and that those means and methods are still within the lawful bounds of the laws of armed conflict, then irrespective of whether the means to that end are human or

machine or some sort of a combination of both, we need to clearly think about those.

Oren Gross: To be sure there are going to be failures. We are not going to get down to zero mistakes. But the real question is what is the standard by which we judge autonomous weapons systems? Are we looking for systems that would have an unrealistic zero risk of failure or do we want systems that are at least as good as, and most likely better than, humans in upholding the laws of armed conflict? We know that when we deploy soldiers, they are going to make mistakes. If you want no mistakes then do not deploy soldiers at all.

Meaningful human control is also a wishful thinking to some extent. Consider what it means to have a meaningful human control over a driverless car. Ultimately, when such cars are available to us, we will want to sleep or read as the car is driving itself. What, then, is our meaningful control over such vehicles? There is also the question of time. With the massive amount of incoming data and the required, speed of decision making, it is unclear how much control you actually can, or should, have over autonomous weapons systems, especially when under fire. In fact, I would suggest that soldiers will not have the luxury to slow things down.

Oren Gross: There are other considerations that we need to take into account that may prevent us from meaningfully controlling those machines in the long run. Consider, for example, the automation bias, i.e., the fact that we put greater degree of trust in computer-generated information than in other sources of

information. Then there is the phenomenon of automation complacency. If I am on the battlefield, and there is a lot of data coming in and the computer tells me X, do I have the time or capacity to second guess it? Do I have the willingness to try to reconfigure and rethink this? Or do I just take X as a given ignoring, for the most part, the possibility of malfunctions or machine errors. This is often coupled with the challenge of machine explainability which makes sophisticated systems practically "immune" to human analysis.

And so, as my time is up, let me conclude by suggesting that meaningful human control may, in fact, be not only unnecessary but actually dangerous to both soldiers and civilians.

Claire Finkelstein: Alright. So this will be, at first, a radical change of topic. One of the most gripping books I have ever read, is a book called The Mascot. It's the story of a little boy named Alex Kurzem during the Second World War. Alex was a five-year-old Jewish boy who watched his family, his mother and his siblings murdered in Ukraine. He escaped and found himself in the woods in Lafayette and was eventually captured by a Latvian SS unit, lined up along our church wall with other Jewish prisoners. He was about to be shot by a firing squad at that moment. For some reason, he reached out and said, can I have a piece of bread? And suddenly the commander told his squad to lower their rifles. He took Alex into the church, pulled down his pants, saw that he was uncircumcised, raised his pants up again and said, "Don't ever let anyone do that to you again."

Claire Finkelstein: He took Alex and made him a miniature SS uniform and adopted him into the unit. Alex survived the war inside a Latvian SS unit, hence the name the Mascot as he was their mascot. So human moral reasoning. The most challenging question raised about autonomous systems is whether we should embrace the idea of non-human actors engaged in self-determining action when lethality is at stake. Do we want machines to have the ability to make decisions with life and death consequences without humans in the loop? Whether self-driving cars, autonomous weapons systems, medical diagnostic programs, and many other applications that are in the works. We are not just playing chess anymore and the stakes are extremely high. Although academics and scientists in the artificial intelligence communities have written about machine intelligence for many years, the question of whether computers can engage in moral reasoning rises to prominence now in this debate with particular urgency. The critical nature of the current moment may be obscured by the fact that we have had semi-autonomous weapons around for a long time.

Claire Finkelstein: The drones that were used heavily by the Obama administration to fight Al Qaeda and AQAP in northern Iraq, Afghanistan, Syria, Yemen, Sudan, and elsewhere were not fully autonomous weapons systems. According to Department of Defense Directive 3000.09, our weapon systems are those that can select and engage targets without further intervention from a human operator. Unlike semi-autonomous weapon systems which leave target selection, the hands of humans and reserve computer activated agency for

implementation of preselected aims. Autonomous Weapons Systems require computer systems to exercise judgment. They must exercise perceptual judgments, spacial awareness, judgment about context and changing conditions. They must be able to capture the best of what we all know as common sense, which is especially hard to capture, and as one roboticist at a conference I held once quipped to me, especially by roboticists who were rather lacking often in common sense themselves. Most controversially, autonomous systems must be able to exercise a certain special kind of judgment, which is ethical judgment.

Now, I believe that roboticists, engineers and computer scientists vastly underestimate the difficulty of this latter task. Whether that is just to help keep us philosophers and lawyers employed, I do not know, might be my own cognitive bias. But what they fail to realize, in my view, is just how complex moral reasoning actually is and how little we know about what it involves. Note how much more difficult this challenge is than other challenges AI has faced. In other areas we have clear criteria for success. If you're trying to model human reasoning around spatial awareness in the way that human drivers do, we know if we have succeeded, if the autonomous vehicles are successful in getting passengers safely to their destination. If we want to know whether a medical diagnosis program successfully models, physician reasoning, we need only look at whether the intelligent diagnostic programs get it right.

Claire Finkelstein: And by the way, my understanding is this has been particularly unsuccessful though my information here may be outdated. In short, moral reasoning about life and death is different when we're examining the ethical side than the more outcome based reasoning we're trying to capture when we're trying to model navigation, spatial awareness and diagnostic programs. With moral reasoning we do not know what counts as getting it right. What most deeply characterizes moral reasoning is not necessarily the outcome as much as the process. So here I wished to highlight two critical questions in this area when we talk about modeling ethical reasoning. The first is what exactly are human beings doing when they engage in ethical reasoning? And the second is whatever that is, whatever they're doing, is this something that we really want autonomous systems to do? We tend to assume that, but actually once we see the way humans reason morally, to me it's an open question: whether or not that thing is exactly what we want to be modeling.

Claire Finkelstein: So let me turn first to the first question. There is very little agreement in the philosophical and psychological literature about the nature of human moral decision making. Two basic views on this question have persisted over the ages. The first regards moral reasoning as the specific application of general abstract moral principles. These are highly abstract, moral norms such as the second formulation of constant categorical imperative that instructs us never to use human beings as a means, but only as ends in themselves. Moral reasoning on this view consists in the application of general abstract principles to particular situations. So

let's call that the top-down view of moral reasoning.

Now, a second view sees moral reasoning as the integration of highly fact specific elements calling for an overall weighing of morally salient features and analogical reasoning, based on a comparison with similar situations with similar features. A person reasoning in this way might notice, for example, that there are four morally relevant aspects of a situation, and we call similar situations in which these same elements were present here. She might then implicitly assign weights to these different elements, and consider how the result in this case based on such assignments might correspond to the results in other cases. On this view, moral reasoning would be more particularistic and context sensitive. It is also on this view analogical neighborly based on drawing analogies between the current situation and other situations involving similar features. So I'll call this view of moral reasoning bottom-up.

Claire Finkelstein: It is relatively easy to imagine how a computer might be able to reason morally if reasoning is top-down. Computers excel at applying general rules or principles to particular instances. One model of this was provided by attempts to build a machine that could make difficult ethical decisions in medical cases. Ethicists considered four principles essential to decisions in this area: autonomy, justice, beneficence and non-maleficence.

The ethical decisions ethicists had to make sure all were involved. They thought of some application of these four principles to the

particular medical situations that arose. Traditionally, bioethicists have maintained that all four principles must be satisfied if a course of action is to be endorsed as ethical. There is no weighting of these principles necessary. They thought all constituted necessary and jointly sufficient conditions. As long as it is possible to teach a computer when these principles are satisfied and when they are not in any given situation. There is no reason to suppose that this kind of reasoning could not be modeled by an intelligence system, but is this what ethical decision making is really like?

Consider a different example. A sentencing jury has been paneled to consider whether a convicted killer should receive the death penalty in what we now refer to as a bifurcated trial. The jury has not itself pronounced the defendant guilty. It has functioned as to say whether the state's request that the defendant receive the death penalty should be granted under current constitutional doctrine. The state's death penalty statute sets out a list of aggravating factors, and the jury must be instructed to find that at least one aggravating factor exists. Mitigating factors, however, are treated differently. Current death penalty jurisprudence insists that mitigating factors be non-enumerated, meaning that the defense may present anything to the jury that it believes speaks in favor of mitigation. There was no restriction on the type of evidence that may count in this regard. Moreover, the only type of death penalty statute that is currently accepted under Supreme Court jurisprudence is one that has a defined list of aggravating factors and a completely undefined or open treatment of mitigating factors. And somehow

the jury is supposed to identify an item from the list of aggravating factors, and then do what with the mitigating factors in combination with aggravating factors, we actually don't know.

Claire Finkelstein: What the Supreme Court rejected in this instance over the course of many years was what looks like a very top-down process. Namely, mandatory assignment of the death penalty based on a rigid list of aggravating factors, and treating mitigating factors according to that list. On the other hand, it also rejected the other extreme, which is completely unguided discretion, which would have been entirely particularistic and context dependent. In other words, the only scheme that the Supreme Court decided is constitutional in this area seems to be an odd and ill-defined mix of top-down moral reasoning and bottom up. Okay. So how should an ethical juror consider mitigating evidence under this sort of death penalty scheme? We have no algorithm for that. Whether or not this captures anything of what ordinary moral reasoning is about in this highly artificial and legalistically bounced circumstance, I don't know, but are reasons why the Supreme Court came to this. That may ring somewhat true in our sense of the reliability of decision making.

Claire Finkelstein: So now back to autonomous moral reasoning machines. The thought I have is that whether or not computers are likely to be effective, moral reasoners depend on the nature of moral reasoning itself. If moral reasoning is top-down, computers have a comparative advantage. But if moral reasoning is more bottom-up, I think computers will find themselves at a relative disadvantage. How do

we decide? Well, I think that is very difficult to say and philosophers have not decided among themselves what moral reasoning looks like. Is it possible that in considering the welter of mitigating factors in our unguided thinking that human beings in fact are much more chaotic, and much more "seat of the pants" than anything a computer could capture.

Now, briefly on the second question then. Suppose despite all of these caveats that I raised about the difficulty of understanding moral reasoning, suppose we did manage to capture the nature of human moral reasoning.

Would we actually want to program robots, computers, autonomous weapon systems to replicate that reasoning? What if human reasoning is so context sensitive and particularistic that, in fact, we wouldn't want any computer algorithm to replicate that even if we could replicate it? Might it be that we would rather have more consistent moral reasoners? Moral reasoners whose processes were more transparent to who were more consistent, who were more reliable? Could, in fact, computers be better, more successful moral reasoners just the way they turned out to be superior at chess? Well, I think that's an essential normative issue that we haven't really dealt with in this literature, and in our efforts. It's not as easy as many of you folks think. Apart from the difficulties of knowing what human moral reasoning is like, the Meta-question here is whether or not that reflects a value that we want to replicate and endorse.

Claire Finkelstein: I think that before we can actually answer the question of whether or not the inconsistency,

hesitation, puzzlement, the agony, the uncertainty, the vacillation, and all of those things that characterize true moral reasoning, is something that we want a computer to replicate. We have to understand better what our values are. Now, here's just one thought I'll close with, which is perhaps in this domain. What we care most about is the process. Or at least if we care about correct moral outcomes, we care about those outcomes in a way that incorporates the path dependency, the messy complicated human process by which we reach those outcomes. That part of how we assess the moral correctness of outcomes is by assessing the process by which they came about. Thank you.

Ben Jones: Well, I would like to thank the three of you for those insightful and provocative remarks. I had prepared some questions but there may be some questions or comments from the audience. So for the time we have left, if there are questions, please come forward.

Audience: I want to follow up on both what you just laid out as a spectrum as possible directions for moral reasoning in artificial machines. I guess the first question I have is why you think that the bottom of what to characterize this bottom-up is more difficult for machines? In general, the different types of AI Algorithms. Some of them will do better with what they characterize as the top-down. The statistics driven ones might be, in fact, very good at the bottom-up, right? And the second question I have is why do you characterize these as potentially up positions or alternatives when it might be in fact the case that people use both of those?

Claire Finkelstein: Yeah. Well, that's a great question, and your first one is one that I anticipated of course. And in this regard, I have to confess to having entered this literature many years ago when people were trying to build connectionist machines, neural nets, which I understand ended up being a spectacular failure. However, I think we learned a lot from that.

Audience: People who did connectionist modeling, actually, even in the 1980s, really came up with very instructive models and, for example, how people acquire information, right? Or how kids acquire information too. I want to caution you in both directions.

Claire Finkelstein: This is something that I would love to learn more about. I think that one of the big challenges that I expect for those trying to model that kind of learning and that kind of horizontal reasoning will be the value judgements that one has to make in identifying the salience of the relevant features that one picks out as the analog. So just as we teach our students how to identify the holding of a case and how challenging that is, given that you recharacterize the holding depending on the context in which you see it. So there is nothing that leaps out at you and says, "Hi, I'm a morally relevant and salient factor. Pay attention to me." Right? So at each stage in the process, they're going to be valued judgments. And, of course, that is the stuff of moral controversy and the stuff of normative reasoning, legal reasoning and so on. So I don't know how, though I'd love to learn more, you're going to capture that in a learning machine.

Audience: So in fact, if you are asking the psychological question right, of all people do that, there is literature on it, and there is work on it, and I pick them off and sell that as well. If you're asking the AI side of it, are people dealing with it? I agree with you. Typically, that's problematic. There's a whole literature on value of that, right? They call it value alignment, even though there is no value in any of this. And it's not clear what's being aligned either. That is a very problematic approach or just entirely sophistically driven. There will be no such reason, as you alluded to, that that system would engage in because all it does is determinate if it is in status to do action. So those are very problematic.

Ben Jones: So I will have the final question, seeing no one at the microphone. And this is to follow up with Oren and also related to Micah's work. Is this a correct characterization of your position? So if we're able to develop autonomous weapon systems that minimize overall harm, but we can't hold anyone accountable, we should go with this system because though we have these problems with accountability still fewer innocent people are getting injured and killed on the battlefield. Or maybe even just generally. And then sort of related to you Micah, how does that mindset fit with your interactions with folks in the military? How do they approach that problem? What's their perspective on it?

Oren Gross: So I'll start. I think it's a close approximation. Obviously, the question of accountability and the fact that we don't have a human being in the doc is disturbing. And the way that international law has moved suddenly since

World War II is towards greater and greater accountability. But we have challenges in the laws that currently stands, as to how far we go. So issues of command responsibility, and those issues are not going to go away. To what extent, I mean, again, this is almost like a balancing issue. To what extent where do you . . . Actually, when I say balancing, it's also, where do you want the cost to lie? So if the price of not having a human in the doc is that you saved 10 lives, 20 lives, a hundred lives, 500 lives, at what point do you stop and you're saying it's worth it, right?

I don't know that we can know it yet, what the answer is. Because again, the technology is moving. We don't know what the technology is going to be, but to me at some point, if the technology is going to be at such a level where we are actually going to be able to save hundreds of lives, and the cost of that will be, that there is no human being in the doc. Even though I understand the potential exploitation of that, it's a price that at least a conversation we should have.

Micah Clark: So with perspective to that view, I would agree that that is the most conceivable technological future, the easiest to achieve. I think keeping any kind of real accountability in the systems is very difficult. I think whether that is desirable from the larger societal and legal perspective is something we have to decide for ourselves. It doesn't fall out of the math, and I don't have an answer for that. It's a trade. Now the view of the technology and minimizing death risks. The humans, certainly in civilians fear that is what we're aiming at. But in the military sphere, it's a double edge sword. Certainly what the

commander wants is something that's lethal in an active shooting war. I'm not necessarily trying to minimize total casualties. I'm trying to minimize casualties to my own troops, but I want to kill the enemy as completely and as fast as possible.

Not necessarily because their death is my goal, but I want to crush their will to fight, and their ability to resist because that is my job in the military. Not that I'm in the military, I'm not. And that is what their role is in preserving the freedom interests. What have you of their side of the conflict. Now there is the issue of both collateral and accidental death. We always want to minimize that. Now in terms of what is the military looking for with respect to these things? The military wants accountability. Part of it is because that is something that they have as a core value within the history of military in general, and certainly the US military on that chain of command that you know who is responsible and someone always is.

Micah Clark: That is ingrained. They would like that. I don't know if that is possible to give them and at the same time give them the kind of mission effectiveness they care about. In terms of the other aspects of the systems minimizing death versus maximizing application of force to your enemy. One of the principle concerns is not so much on the raw ethnic side of it, even though there is a lot of work on ethical reasoning, it's or the computational philosophy side of it, but the predictability. If I'm from military standpoint and willing to buy this system or choose to use this system, I want to be able to have enough predictive, projective, anticipatory capability that I understand that, if

I use it in this situation, I've got very good chance of achieving the effects that I intend, and that if I don't achieve that, I can reasonably anticipate what the contingent situations will be.

If it is a random dice roll, well in my work I might not, it might kind of snake eyes, and we don't know anything. Then just from a risk management perspective, how I'm balancing my force, how I'm planning to prosecute the mission, it's the risk reward doesn't make sense there. So the military wants accountability in the systems that they are not looking for terminator, and they want the ability to predict and anticipate and understand what the mission effectiveness of these systems will be in a contested, uncertain, under-specified kind of situation. And those are two things that the types of technology we use now, especially the active learning and the deep network statistical approaches, they are mad at that. The test and evaluation and V and D processes that we use for other kinds of munitions and frankly the kinds of systems in general, even automotive and planes, they're almost inapplicable for these kinds of active learning systems.

Micah Clark: And so, as the raw technology side, we have an idea that people want ethically constrained systems, and we can argue about, it is because we want them to be better ethical reasons than us. Or it is simply that, "Hey, we need them to operate in the gray areas." And in the gray areas where there is lots of data about what the right answer is. We want them to at least match our expectations, and what we anticipate them to do without the necessarily saying that it is the morally right thing. We want transparent

systems that can produce explanations. We want to understand when systems will fail, before they fail and be able to understand the failures after they've occurred. From the technology standpoint, we can't give you any of that. We're working on it, but we're not there. And that is the challenge that whether it is on the legal side, where the law relies on innate human abilities to ground out the theories, or it is on the military side where we want accountability, we want to understand if this system will work or not in future conflicts. Those are things we can deliver yet.

Ben Jones: Okay. Well, thank you. Thanks to each of you. Could we have a round of applause for our panelists?