Why the Future needs us today

Moral Responsibility and Engineering Autonomous Weapon Systems

JEROEN VAN DEN HOVEN[1]

PHIL ROBICHAUD

FILIPPO SANTONI DE SIO

Delft University of Technology

Dear mr/madam chair, your excellencies,

I would like to express my gratitude to be able to speak on an extremely important topic that will stay with us for the remainder of the 21st century as our increasingly complex and fundamental technologies develop and converge.

The recent history of the ethics of technology has taught us that ethics needs to be pro-active or be too late to make a difference. So we may now be just in time in talking about autonomous weapon systems. Another lesson is that for ethics to make a difference in the real world of technology it needs to take a design stance and use the moral considerations and values that we share as requirements for the design of our technologies. We must prevent a situation where there is a disconnect between abstract moral discussions and the real world of engineering. This proactive design approach to ethics is referred to as *Value Sensitive Design* or *Design for Values*. In *Value Sensitive Design* ethical constraints and aims become the very shapers of innovations at the relevant point in time and a place, where they still can make a difference, instead of fuelling political and academic discussions after the fact. By proceeding in this design oriented and proactive way, value conflicts may even be overcome and reconciled by means of design. It is often argued that this is precisely what genuine innovation is about: the introduction of new functionality that accomodates conflicting or competing values that we find difficult or impossible to satisfy jointly.

Value Sensitive Design is a relatively new concept but will have to become our bread and butter in the 21st century throughout society. We already have good experience with design for privacy and design for sustainability, and are looking at design for inclusion and many more[2]. Design for X for short, where X ranges over the set of moral values we deem important. In the 21st century we will have to be able routinely to articulate values and moral principles and implement them; audit and

[1] We would like to thank Maneesh Verma and Pascal Gemke for their assistance

[2] See for a comprehensive overview, Van den Hoven, Vermaas and Van de Poel (eds.), *Handbook of Ethics, Values and Technological Design*, Springer, 2015.

verify that we have implemented those values; demonstrate that critical technical functions compound to deliver the overall functionality conducive to the morally legitimate goals of the system.

In this sense, *Responsibility is also a Design Challenge*. When it comes to complex, collective, time extended, technologically and institutionally complex endeavours we will not be able to get a satisfactory answer to the question as to who *was* responsible or accountable for untoward outcomes, deaths and human suffering **unless** we have designed responsibility *into* the system right from the start. My conclusion is that responsibility is a non-functional requirement that can and should be designed for.

1. Autonomous Technology

To which types of systems should this novel approach in ethics of technology be applied with priority?

I think it is not realistic to assume that technology will have any time soon in the forseeable future the capabilities that will allow it to be engaged in moral perception, and make the moral distinctions and assesments that humans routinely make. Neither will systems of the near future take the social, legal, psychological contextual information into account, and arrive at moral judgements with the sophistication commonly seen in highly experienced and trained human beings. Therefore no systems in the forseeable future will have the properties that will allow them to satify the moral requirements of IHL, since their application presupposes all of these capabilities.

However, in the next two decades it will be possible to create systems of systems compounding high tech and quasi- autonomous devices that are equipped with advanced analytical tools, machine learning, pattern recognition, reasoning and incredible computational power. These systems can execute tasks – including ones that involve the application of lethal weapons - that are provided to them in large time windows and increasing spatial scopes; they may even dream up new tasks that are relevantly similar to the ones provided by humans and initiate causally efficacious processes.

These are the systems on which we should focus our attention with priority.

2. Meaningful Human Control and moral responsibility

Leaving Fully Autonomous Systems aside for the moment, it should be observed that we already have trouble maintaining a sufficient degree of control over the design, production, implementation and use of large IT systems. One of the fathers of Computer Science, Joe Weizenbaum, already expressed such worries back in the early 1970s in his book *Computer Power and Human Reason*. He quoted admiral Moorer who looked back upon a certain episode of the Vietnam War and said: "It is unfortunate that we had to become the slaves of those damned computers". This remark draws attention to the problematic knowledge dependence relation of human operators working with advanced computer systems and the corresponding loss of autonomy. I have argued elsewhere that once human operators have chosen to work with these IT systems they can become their slaves

because they cannot but comply with the system, or else they take a moral risk they cannot justify *in situ*. This knowledge dependence is a common feature of those system environments of which military systems are a paradigm example. In any case, ever since Admiral Moorer made his prescient, if provocative, statement, many accidents - in armed conflicts and in civilian domains - with many casualties have resulted from problems with human control.

Other contexts in which the loss of meaningful control raises important questions include the actions of pilots, surgeons, fire fighters, soldiers, managers, and operators (users). They have obligations respectively to transport persons, save lives, extinguish fires, rescue innocent citizens from the hands of the enemy, create jobs and prevent explosions. In carrying out these obligations, these actors are dependent on the designers and engineers who have supplied them with the tools they need to do what they ought to do. If these systems are designed not to accomodate them as responsible human agents, they are bound to fail in this respect. However, this dependence on supplied technology should not function as an excuse for users and operators in case they fail or are implicated in untoward outcomes. Users and operators are arguably here in a position similar to that of a drunk driver who causes a car accident while out of his mind, but who could have avoided the accident by, say, deciding to take public transport. There is a shared and collective responsibility of a higher order (a so-called *meta-task responsibility*) for creating the circumstances in which others - or our future selves - can do what they ought to do. A responsibility for our responsibilities.

Many theories of responsibility put control centre stage. According to so-called compatibilist theories of responsibility, agents are responsible for their actions only if they control them, and in order to be in control of an action an agent must be responsive to sufficient moral and practical reasons to act. This entails that in the presence of strong reasons to act, the agent must have the capacities to (a) recognize these reasons and (b) bring himself to perform that action in a sufficiently broad range of circumstances[3]. These capacities must, moreover, be an integral part of who the agent is.

In line with this, we could thus construe MHC as follows:

A has Meaningful Human Control over X if

(a) A has taken ownership of a decisional mechanism (the mechanism of practical and moral reasoning) through which he implements his decisions about acting upon X
(b) the decisional mechanism is –known to be – responsive to the moral and practical reasons of A

An important aspect of Meaningful Human Control can thus be put in terms of a tracking relation between an agent's moral reasons and his causal interventions in the world. States in the world vary with specific mental states of human agents that we usually refer to as moral reasons and processes of moral reasoning. Ethical appraisal and evaluation of persons and their behaviour is always about this relation. Assuming that we know what the right-making properties are, we always prefer a situation where people act *because* of these properties to a situation where they merely happen to act in mere accordance with them. *This* notion of Meaningful Human Control, when fully developed, will give us a touchstone for ensuring that we can design systems that preserve the kind of control that is central to responsibility. Everything which transpires in such a system should demonstrably

[3] Fischer, J. M. & M. Ravizza 1998. Responsibility and Control. A Theory of Moral Responsibility. CUP.

and verifiably stand in the right relation to human moral reasons and reasoning– no matter how many system levels, models, computation or devices of whatever nature separate a human being from the ultimate effects in the world – some of which may be lethal.

People who will develop and design tools that are unimaginable to us now should be concerned with preserving meaningful human control along these lines. And even if these tools exhibit high levels of autonomy, they would not have been around *but for* us. Any fully autonomous agent or system exhibiting high level autonomy and capable of largely unsupervised task performance must be embedded in systems of systems that have been Value Sensitively Designed for accountability and moral responsibility.

3. Value Sensitive Design for Humanity (IHL and IHR)

So the way forward is that we will have to learn how to design for responsibility and accountability in weapon systems to make them demonstrably compliant with the fundamental moral principles underlying IHL and Human Rights Law. This will be difficult because we will be confronted more and more by computational technologies which obfuscate their inner workings or which are by their very nature intransparent (neural nets, quantum computers).

A possible solution to these and other responsibility problems is to start to think about and work on *responsibility sensitive design,* that is thinking about responsibility as a matter of (a) *design* (b) *in advance* (c) in a fairly *fine grained* way in order to bring about the conditions under which human beings can be held responsible at all. This is a daunting task, scientifically, legally, institutionally and morally.

In the meanwhile -while we still have some time - the UN should undertake action and encourage the study of evidence based global governance, and foster mechanisms of confidence and trust building between states regarding emerging and converging technologies.

The moral qualities of the world in the second part of this century depend on efforts in careful and meticulous design for values today.