Forthcoming in:
Battaglia F & Weidenfeld N (eds.), *Roboethics in Film*. RoboLaw Series. Pisa: Pisa University Press.

Who's Afraid of Robots? Fear of Automation and the Ideal of Direct Control

Ezio Di Nucci & Filippo Santoni de Sio

Abstract We argue that lack of direct and conscious control is not, in principle, a reason to be afraid of machines in general and robots in particular: in order to articulate the ethical and political risks of increasing automation one must, therefore, tackle the difficult task of precisely delineating the theoretical and practical limits of sustainable delegation to robots.

1. The good HAL and the bad HAL

Movies provide a good exemplification of a deep-rooted ambivalence in western culture towards task delegation to machines and robots. On the one hand, we recognize the opportunities opened by robotics technology and are fascinated by the idea of automation and delegation to robots; on the other hand we also fear the idea of delegation to machines. Movies have often presented both utopian scenarios with robots becoming the best partners of mankind and dystopian scenarios in which robots rebel, take over, or become in other ways a dreadful threat to humanity. One outstanding cinematographical exemplification of this ambivalence is Stanley Kubrick's *2001: A Space Odyssey*. In his adaptation of Arthur C. Clarke's novel *The Sentinel* Kubrick tells of a space (and time) adventure. One of the main characters is a futuristic computer: HAL 9000. Clarke's and Kubrick's HAL 9000 is a perfect exemplification of the above mentioned ambivalent attitude towards technological delegation. In the

firs part of the movie HAL embodies all the features of the perfect work partner. He performs all the required tasks much more quickly and efficiently than his human counterparts, so relieving humans from a huge amount of work. He is absolutely reliable and is even programmed to have complex and pleasurable verbal interactions with his masters. However, in the second part of the movie HAL becomes his masters' worst nightmare. Not only he starts to malfunction, but when the human crew decide to deactivate him, he deploys all his cognitive powers to resist their plan, he engages in a struggle with the human members of the crew for the control of the spaceship, and he finally tries to kill them all in order to continue the mission according to his own vision and goals.

This ambivalence of feelings towards robots is not only to be found in literary and cinematopgraphical fiction, but is also present in the present-day philosophical debate on the ethics of robotics, and in particular in the debate on war robots. Whilst some argue that a delegation of (part of) war operations to robots may be ethically beneficial, as it may enhance the efficiency of the operations while at the same time reducing the presence of human soldiers on the battleground and possibly reducing the risks of unwanted harm to civilians; other have expressed strong negative reactions towards the idea of delegating killing in war to robots. Even though this ambivalent attitude is in general justified by the fact that automation and delegation to machines – especially in delicate contexts like war – may indeed have undesired negative outcomes and involve unacceptable risks, in the first part of this chapter we argue that absence of direct and conscious control is not, in principle, a reason to be afraid of machines in general and robots in particular. Therefore, in order to articulate the ethical risks of increasing automation one must go through the bother of precisely delineating the theoretical and practical limits of sustainable delegation to robots – which is what we (start to) do in the latter half of this piece.

We start from general considerations from the philosophy of action and responsibility and we try to apply these considerations to the special case of action *through* robots. In the first part of the chapter we claim that automation and delegation are the norm in human action. We often delegate very important tasks to subpersonal systems that are not under the direct control of our conscious self – and we have very

good reasons to do so. In the second part we recognize that automation and delegation is not *always* beneficial, and direct conscious control is sometimes required to avoid unwanted risks. In particular, we argue that direct conscious control is preferable when either of two risks is present, namely: we have reasons to think that the scripts guiding our less than fully conscious behaviour may be in contrast with our general goals; or we have reason to fear that while not in conscious control, other external agents may take control over our behavior and direct it against our best interest. With due caution, these considerations may be applied to the case of delegation to (war) robots.

2. Control delegation and human automaticity

Delegating conscious and direct control is an effective way of freeing up cognitive resources: it is not just that there is nothing wrong in doing it; rather, there is something wrong in not doing it. We could say that one way of learning and mastering a practice is to achieve a level of confidence and expertise such that we can afford to delegate conscious and direct control: we can stop, for example, paying constant attention to the performance.[1]

Particularly interesting for our purposes is that this mundane practice of control delegation as a form of cognitive efficiency is independent from whom or what we delegate to: traditionally, we delegate to unconscious or sub-personal mechanisms within our body, as when through increased coordination we no longer have to look at the racket in order to hit the ball back over the net.

The crucial element, here, is that as a result of mastering a practice there are less cognitive tasks that need absolving in order to bring some performance to successful completion. Once we drop 'cognitive', the same can be said about our relationship to machines: the lights in the

[1] One of us has written extensively on this and related topics: see, for example, Di Nucci E. 2008. *Mind Out of Action*. Saarbrücken, Germany: VDM Verlag; Di Nucci E. 2011. Automatic Actions: Challenging Causalism. *Rationality Markets and Morals* 2 (1): 179-200; Di Nucci E. 2011. Frankfurt versus Frankfurt: a new anti-causalist dawn. *Philosophical Explorations* 14 (1): 117-131; Di Nucci E. 2012. Priming Effects and Free Will. *International Journal of Philosophical Studies* 20 (5): 725-734; Di Nucci E. 2013. Habits, Nudges, and Consent. *American Journal of Bioethics* 13 (6): 27-29; and Di Nucci E. 2013. *Mindlessness*. Newcastle-upon-Tyne, UK: Cambridge Scholars Publishing.

office corridor go on automatically when we walk through the door: that is another example of delegating control to an (external) system so as to minimize the number of tasks which need absolving in order to successfully get to our desk.

One important question is obviously if there are any important differences which result from delegating conscious and direct control to an external system as opposed to an internal system. We will deal with this important issue later in the paper; but, first of all, let us focus on the similarities between control delegation in the two cases: we have already mentioned a first crucial common feature, namely that in both cases control delegation is a way of making the practice more effective.

Interestingly, increased effectiveness is not only the result of saving resources (in one case not having to pay attention, in the other case not having to flip a switch); increased effectiveness is, primarily, the result of the fact that control delegation signals an increased competence in the task: I no longer need to pay attention to my swing or steps (or to look at the racket; or to think about my pin number before typing it in – each of us can insert here our preferred examples from daily life) because I am now better at it; similarly, it is at least to be supposed that in normal circumstances the system to which we delegate control of the office lights is better at turning them on and off than we are ourselves; otherwise it is not at all obvious why we would delegate control of our office lights to an external system.[2]

This point about competence is important because it illustrates the superficially paradoxical nature of delegation: we give up direct and conscious control of a particular task when (and because) we have improved our control over the whole practice: more control, then, results in less controlling. This point is, as we anticipated, only superficially paradoxical once we get clear about the dispositional nature of control. Control is definitely not an activity and possibly not even a state: control is, rather, a capacity. There is the activity of

[2] This point does need to be qualified in at least one important respect which will also be important in the latter part of this paper: namely, the reasons for delegating to an external system in modern society may be more complicated than effectiveness and reliability: maybe there are, say, market pressures to commercialize the management of office lights, so that we delegate to an external system because of such pressure and not because it is more effective or reliable.

controlling which may turn out to be sufficient for control but it is definitely not necessary for control.

Indeed, I think it is easy to show that the activity of controlling is also not sufficient for control, as we can easily generate cases where someone is controlling something over which they actually have no control: imagine an unfortunate security guard who spends the whole night guarding the entrance to a bank to then find out in the morning that the bank has been robbed from underneath during her shift.

One can argue about intuitions here, but it is plausible to describe this scenario by saying that the security guard was controlling without having real control. The same point can be made about actions: I can look at and concentrate onto the incoming ball and the movement of my arm as intensively and carefully as I like, but as I cannot play tennis I won't really be able to control my shot. The idea is, again, that the capacity – rather than any activity – is crucial for control.

Anyway, the sufficiency of the activity of controlling for control is, for our purposes, less interesting than the claim that the activity of controlling is not necessary for control: as in, we may have control without having to do anything, let alone having to do any controlling. Normally, when I act, I don't need to perform any extra activity of controlling: the bare fact that I master the practice of walking constitutes my control over my steps – I don't need to do anything else. As Gilbert Ryle famously pointed out, by looking at exceptions we can easily confirm this simple point:

> When we describe someone as doing something by pure or blind habit, we mean that he does it automatically and without having to mind what he is doing. He does not exercise care, vigilance, or criticism. After the toddling-age we walk on pavements without minding our steps. But a mountaineer walking over ice-covered rocks in a high wind in the dark does not move his limbs by blind habit; he thinks what he is doing, he is ready for emergencies, he economizes his effort, he makes tests and experiments; in short he walks with some degree of skill and judgement (Ryle 1949, p. 42).[3]

[3] Ryle G. 1949. *The Concept of Mind*. Penguin.

3. Control and controlling

It is particularly important to distinguish between the activity of controlling and control as a capacity when thinking about control delegation to internal and external systems. This allows us to clarify that control delegation to external systems must not be conceived as losing control but rather as out-sourcing the activity of controlling. The fact that members of the department no longer have to flip a switch in order to turn on the light in our office corridor does not mean that we no longer have control over those lights: it only means that we have effectively out-sourced the activity of monitoring the turning on and turning off of our corridor lights.

Again, the example shows that out-sourcing monitoring may result in more rather than less control: we now have a solution against both laziness and forgetfulness, for example – which is effective against wasting energy. Because we have developed a system which prevents mistakes that have to do with laziness and forgetfulness we can actually argue that through this external system we now have more rather than less control over our energy consumption: again, through the out-sourcing of the activity of controlling we have improved control because we have perfected our capacity through the deployment of technological aids.

Let us then distinguish between a stronger and weaker claim here: according to the stronger claim, out-sourcing the activity of controlling may result in more control; according to the weaker claim, out-sourcing the activity of controlling does not need to result in any less control. For the purposes of this paper, we don't really need to decide between those two claims because they both serve our purpose of arguing against skepticism against technology which is motivated by supposed loss of control.

As long as we maintain control over the whole system, there is then nothing to be afraid of: control delegation is, it turns out, a form of empowerment. And we have shown, importantly, that this empowerment is independent of whether we delegate control to an internal system such as our body or to an external system such as, say, a computer program.

What is crucial, then, is that we in turn maintain overall control over the system to which we have out-sourced the activity. This overall control may take many different forms which will also vary depending on the social and technological context[4]: in the case of our office's automatic lights, it seems for example that what is crucial to whether or not we have control is that there is a reversible decision-making mechanism of which we are part; say we have decided in a department meeting to introduce automatic lighting and we could reverse that decision in a future meeting or even call a meeting for that particular purpose.

If such a mechanism is in place, then we can talk about control; if there is no such decision mechanism or we are not part of it – say university administrators have decided to introduce automatic lighting and we have no influence on that decision one way or another – then we cannot claim to have control over it, independently of whether we like automatic lighting or not or whether it serves our purposes or not.

Here one may object that the criteria for control attribution in this case should be more demanding: say, we can only claim to have control over our office lighting if – even though the lighting has been made automatic – there is a switch off button somewhere which can reverse the system to a manual one. Exact criteria for the attribution of control are difficult to establish and it is an interesting question whether or not one would need a general shut-down button to claim control in this case.

Imagine that a guest to the department would object to the system on some reasonable ground (maybe she thinks that we should move to an alternative system which is even more energy-efficient): now it seems that we could not reasonably claim that the current system is out of our hands just because there is no shut-down button; if there is a decision-making mechanism to which we have access, then it seems that we would have to admit some form of control and offer to bring

[4] This may mean that, in the end, some form of contextualism will be true for the concept of control; this may be more or less good news, but that will not be our concern here: we do not mean to put forward a full-blown account of control here but only to point to some important features of control for what concerns a particular debate in the ethics of technology.

up the issue in the next department meeting if indeed we accept the reasonable grounds which our guest brings against the current system.[5]

4. How much control do we need?

The fact that a shut-down button may not be necessary for control when it comes to office lighting does not mean that a shut-down button is never necessary, though. Take the case of out-sourcing our war effort to so-called autonomous drones and other military robots.[6] Now it seems that a political leader with access to the decision-making mechanism which approved outsourcing military tasks to autonomous drones or robots may be made responsible even if she does not have access to a shut-down button; but they may not on the other hand be said to control the drones (not anymore, anyway) if there is no shut-down button even though they may have approved their deployment.

Control and responsibility are importantly related, but it seems as though we do not control all and only the things for which we are responsible: we may well be responsible for events that are truly beyond our control as long as some other condition applies – for example that we had some degree of control over things that could have been reasonably expected to result in the current state of affairs (more on this issue later). Also, at least in the sense of culpable we may say that we are not responsible for all the things that we can control, as in all the cases in which we are justified.

Here one could object that we have come to the limits of our analogy between delegating control to internal unconscious and sub-personal biological systems and delegating control to external technological systems: what do autonomous drones have to do with learning to walk or automatically type in your pin number? We have said so far that the common ground is constituted by the fact that in both cases we out-source and delegate conscious and direct controlling

[5] I think this intuition does at least in part depend on the fact that – at least in normal cases – we take it that the system does serve the purposes of all those involved; so it is our access to the decision-making mechanism plus the fact that we use the system and are happy with it which together motivate the intuition that in this case we have control even though we don't have access to the shut-down button; having analyzed the intuition as above, I think nothing bad follows from it for our argument.

[6] See our forthcoming book *Drones & Responsibility* on this topic.

of some task while at the same time maintaining overall control over the practice of which the particular task is a part. One could object that nothing can go wrong with an internal unconscious or sub-personal biological system in the way in which a drone may short-circuit and go AWOL; and that, anyway, the epistemic authority we have over our own body and self cannot be compared to any relationship we have to external technology.

Let us deal with each of these points in turn, starting from the last one: a principled distinction between the own body and self and external technological systems is a dubious one. Without even worrying about fancy future developments, just think of Neil Harbisson (Else 2012) [7]: whether or not he should count as a cyborg, it's going to be extremely difficult to distinguish between his internal and external systems and even more difficult to draw a principled distinction between his control over his internal systems and his control over his external systems.

Let us move on to the point about things going badly wrong with drones at a level at which they cannot go wrong within our own body: it is not just that our body can malfunction more or less predictably just as much as any technological system (indeed, one could even argue that malfunctioning in technological systems is more predictable than biological malfunctioning because we have designed the former in a way in which we have not designed the latter); more importantly, we can be ignorant about own our body, self and psychology just as much as we can be ignorant about external systems.

Two basic examples of the way in which control over our own psychology and agency is not special in any ontological or epistemological way are the phenomena who are interesting to psychoanalysts (without getting into complicated or controversial stuff, just think of simple Freudian slips) and the priming literature: plenty of experiments over the last few decades have shown that our behaviour can be influenced without our realizing it and also without our losing the perception of control; simplifying, we do what we have been primed towards without realizing that we have been primed and also

[7] Else L. 2012. A cyborg makes art using seventh sense. *New Scientist* 215 (2877): 50. In case you don't know who that is, you may start from this Wikipedia entry: http://en.wikipedia.org/wiki/Neil_Harbisson.

without having the impression of reduced control or diminished agency (for a classic example of this, see Bargh's original work: Bargh & Chartrand 1999; Bargh, Chen and Burrows 1996).[8]

The point we are making by mentioning Freudian slips and priming isn't a controversial one, as we are not claiming that those things are not within agential control: the point is, rather, that the very fact that these phenomena are at least indirectly within agential control shows that control over internal biological systems is not different in kind from control over external technological systems and that it is subject to influences that the agent is not necessarily immediately aware of. To mention just one example: evidence that agents can be made to walk slower by being primed with words which remind them of the elderly isn't – in epistemic terms – different from a user manual for a drone or the technology behind automatic lighting; agents (can) have the same kind of access to both.

Another obvious objection may be lurking here: office lights which turn on and off automatically are an incredibly basic and simple example of control delegation when compared to robots: so that what we say about automatic office lights may not apply to robots and, more in general, delegating control to robots and other complex technology is a much more complicated and less reassuring practice than automatizing office lights, which raises many further difficult questions. Let us then look at whether something important changes – in terms of control - when we apply our argument to extremely complicated systems such as present and future robots.

5. The curious case of Kenneth Parks

Losing direct conscious control on one's actions may indeed sometimes lead to very bad consequences. This is true both in the case

[8] Bargh J.A. & Chartrand T.L. 1999. The Unbearable Automaticity of Being. *American Psychologist* 54: 462–79; Bargh J. A., Chen M. & Burrows L. 1996. Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action. *Journal of Personality and Social Psychology* 71: 230–44. We are aware of the controversy surrounding replications of priming experiments and we are happy, for that reason, to give a conditional structure to this part of our argument: if the priming literature is sound, then… Nothing much turns on this particular point for our overall argument anyway.

of delegation to external devices and in the case of delegation to subconscious parts of oneself. Consider the curious case of Ken Parks. In a night of May 1987, the 23 year-old Canadian man Kenneth Parks rose from the couch on which he was lying in front of the TV, put on his shoes and jacket, walked to his car and drove 23 kilometres to the home of his parents-in-law. He entered the house, killed his mother-in-law by repeatedly stabbing her, and seriously injured his father-in-law. He then left the house and drove to the police station where he told police that he thought he had killed some people[9]. Parks pleaded not guilty, on the basis that he had been sleepwalking the whole time. His memory of the events of that night was confused and fragmented. He seemed not able to recall anything about his violent actions; apparently, he realized that he had a knife in his hands only after he had left his parents-in-law's house. He had no previous history of violence, and on examination he appeared horrified by what he had done. Indeed, Parks did horrifying things and he did them *because he was less than fully conscious*.

Whatever the best psychological explanation of Parks' violent behaviour[10] – whether his behaviour has to be explained in terms of the behaviour of networks in subcortical and brainstem areas responsible for the generation of innate and archaic emotional and motor behaviours that we all have; or rather in terms of Parks' individual psychological conditions of stress and anxiety connected to his personal situation and to a delicate meeting with his in-laws that he would have had the day after that tragic night – whatever the best psychological explanation we may confidently say that *Parks would have never killed his mother-in-law if he had been awake*.

Moreover, we do not need to refer to such a unique case to realize that when it comes to morally relevant actions, conscious deliberation may often prevent us to do inappropriate things. It is, for example, advisable to not take important decisions or perform actions with relevant consequences while in a condition of stress, sleep deprivation, drunkenness, sexual arousal, hunger and other circumstances that reduce our ability to rationally focus and engage in a proper rational

[9] Parks was found not guilty, and his acquittal was subsequently upheld by the Canadian Supreme Court, in *R. v. Parks*, (1992) 2 S.C.R. 871.

[10] See N. Levy, *Consciousnees and Moral Responsibility* (Oxford: Oxford University Press, 2014), ch 4 for a discussion of the topic.

deliberation over the moral implications of our actions. Moreover, it is certainly true that it is often not a good idea to talk and act while in a state of great anger or psychological frustration.

However, this does not mean that, whilst automation and delegation to subpersonal mechanisms may be good to perform morally irrelevant actions like playing golf and switching on and off our office's light, automation and delegation are *always* to be avoided when it comes to complex morally relevant actions. And thus that, for instance, machine and robots should *never* be employed in activities like healthcare and war operations, where the well-being, if not the life of people, is at stake. It only means that *under certain conditions* automation and delegation must be avoided. Some of these conditions may be identified through a closer examination of the abovementioned examples of delegation to less-than-fully-conscious parts of ourselves.

According to different possible explanations (and circumstances), somnambulistic violent behaviour like Parks' may be seen either as *completely random*, or as responsive to "scripts" *that do not embed any appropriate rules of behaviour* – be them archaic and innate simple neurological structures or individual and acquired complex emotional reactions to a difficult situation[11]. Similarly, what is undesirable about acting under the influence of mind-altering substances like drugs or alcohol or "visceral factors" like hunger, sleep, or sex drives is the fact that these behaviours follow either no rational pattern at all, or only very simple and short-sighted desires and cravings[12]. Finally, our emotional reactions are undesirable only insofar as they are expressive of attitudes *that are not rationally justifiable*.

In all these cases of less than fully consciously deliberated behaviours, it is not simply their not being the product of conscious deliberation that causes them to be inappropriate; it's their being controlled by irrational, inappropriate, too simple or not flexible enough scripts. Indeed, other less than fully consciously deliberated behaviours, for instance some emotionally triggered reactions to a moral injustice, are often not only clearly appropriate, but also, in some circumstances, the only or most appropriate response. The reason is

---

[11] *Ibidem*.

[12] Cfr. J. Elster and O.J. Skog (eds.), *Getting Hooked: Rationality and Addiction* (Cambridge MA: Cambridge University Press, 1999).

that emotions may embed and express deep-rooted appropriate moral and social reactions and attitudes in a way that rational deliberation cannot always do. It is certainly bad to scream "I hate you! I do not want to see you anymore" and to kick a beloved person out of one's home under the push of a momentary burst of rage triggered by an episode of irrational jealousy. The same rage and the same behaviour may be completely appropriate if triggered in a long-time abused woman by one last provocation from a violent partner. Indeed, it is an old idea in moral psychology – one that goes back at least to Aristotle – that a good moral education partly consists in the acquisition of automatic patterns of reaction, and that it is part of the psychological structure of the virtuous person that, at least in some circumstances, she does not have to think in order to be able to do the right thing: in a slogan, *to think is to hesitate… and the virtuous person does not hesitate*.

Admittedly, it may difficult to make a case for the possibility of performing complex good moral actions while in a state of somnambulism, under the effect of mind-altering substances like alcohol and drugs, or under the pressure of basic visceral factors like strong hunger, thirst, sex drives, or psychological stress. However, for the sake of our present point, we do not need to make this case. We have not claimed that automation or delegation is always morally appropriate, or that *any* kind of automation or delegation may be morally appropriate under *any* circumstance. What we are claiming is that even in the case of complex, delicate, and morally relevant actions, some kinds of automation and delegation may be morally appropriate. Or, in other words, that delegating a complex, delicate, and morally relevant action to a script or a program is not *necessarily* bad. It depends on whether the content, the complexity, the flexibility and the reliability of the script or the programme fits with the goals that it is meant to realize.

6. The Cabinet of Dr Caligari

A related though different concern with automation and delegation is that by losing direct conscious control on morally relevant actions humans may lose or restrict human autonomy. In order to make sense

of this concern and to assess its reasonableness in relation to delegation to technological devices, we will start again from looking at more paradigmatic cases of loss of autonomy, and then we will use them as a tool to draw the limits of reasonable concerns in relation to the human interaction with robots and machines.

Consider another – this time fictional – story of somnambulistic crime: the case of Cesare in *Das Cabinet des Dr. Caligari* by Robert Wiene. The evil Dr Caligari has found a way to hypnotize his victim, Cesare, in such a way that Cesare commits a series of crimes devised by his "master" while sleepwalking. As in the real case of Ken Parks discussed above, Cesare is not responsible for his actions. However, unlike Parks who acted by realizing some unconscious scripts causally connected to his own neurobiology or his own personal history, Cesare is unconsciously acting according to the scripts deliberately prepared by his master, Dr Caligari. Cesare's case thus concerns us in a different way than Parks – it is not just a case of loss of control, it is a paradigmatic case of loss of autonomy, as it entails the intentional manipulation by another agent; and being in someone else's power is "the paradigm of unfreedom"[13].

It is important to note that in cases of manipulation, it is not the loss of control on his thoughts and actions alone to make Cesare lose his autonomy. It is the combined presence of loss of control plus the following three additional elements: (a) there is another (human) agent (b) who has some access to the first agent's thoughts and thus controls his actions according to the controller's own plans, and (c) the goals and plans of the controlling agents are in contrast with the plans of the one that is controlled. This analysis helps us to draw important distinctions between those delegations of control that involve loss of autonomy and those that do not. Compare the case of the evil psychiatrist Dr Caligari with a case of voluntary psychological treatment of addiction. If through a series of sessions a skilled therapist manages to understand the patient's psychological structure in a way that the patient herself cannot, we may say that the therapist has exclusive access to (a part) of her patient's mind; and if it is the case that the

[13] B. Williams 'How free does the will need to be?', in B. Williams, *Making sense of humanity and other philosophical papers 1982-1993* (Cambridge: Cambridge University Press, 1995) 4.

skilled therapist also knows how to treat her patient's addiction, then we may say that the therapist controls her patient's addictive behaviour (or at least that they share control of it, given that an effective therapy will require the patient's active collaboration). However, if thanks to her therapist's treatment, the patient gets rid of her addiction, we would not say that the therapist has manipulated her behaviour, or has reduced her autonomy by intervening on her mind. If anything, we would say that the therapist has helped the patient – on her request - realize her goals, and has therefore helped her to exercise or even improve her autonomy.

We may apply a similar reasoning to the interaction with robots. Delegating (part of) the control on one's actions to machines and robots may indeed sometimes open the way to the agent being manipulated, provided the following circumstances realize: (a) there is at least another (human) agent who (b) because of the task delegation to a machine or a robot by the first agent (c) is able to use the first agents' actions to realize his own plans/goals and (d) these plans/goals are in contrast with those in the light of which the first agent started her activity. Delegation or shared action with trustworthy or controllable agents is not in itself an issue; it is the basis of successful joint enterprises. Being open to manipulation or influence by agents with contrasting plans, intentions or values is the issue. What is really scary about HAL 9000 in the second part of Kubrick's *2001: A Space Odissey* is not HAL's being way more powerful and knowledgeable than his human partners. It is his becoming a full-fledged agent to a point in which he starts having *his own* goals, his own *ambitions* and *values* and starts using his power and knowledge to intentionally contrast the interests of the human members of the crew.

7. Hackers, gamers, lobbyists

Before concluding, we want to show how three prominent concerns about the deployment of robotics technologies in warfare may be understood and assessed through the idea of manipulation as presented in the previous section.

Firstly, there is the concern that war robots may be hacked by enemies and used against those who had put them on the ground. This

is a straightforward case of manipulation. An agent A delegates his work to a robot, and that robot and the information that it contains is then hacked and used by agent B against agent A. Here one can envisage a clear analogy with the case of psychiatry mentioned above. Imagine a case in which an evil colleague of your therapist steals your therapist's notes about your mental conditions and somehow uses these information to control your behaviour against your interest. Even though the possibility of using your own plans against you is certainly not a risk *uniquely* present in the case of war robots – spying has always been part of war – the concerns about the risks of manipulation in the use of war robots are certainly reasonable.

A second concern is that by replacing traditional weapon systems with high-tech robotics technology like armed remotely controlled drones traditional soldiers will be replaced by "cubical warriors", that is drone operators sitting, operating and killing from a station in their home country or anyway far away from the battlefield. This replacement is seen as a potential risk among other things because it is thought that these new soldiers may bring a sort of "Playstation" mindset into war; for example by not fully realizing the life-and-death character of their job, drone operators may end up being more reckless, or even – in a very pessimistic scenario – killing for fun. Also this concern may be framed in terms of manipulation. In fact, the idea here is that drone operators may end up bringing their own values and goals into war, and thus they may turn war operations into something different than originally planned by politicians and military commanders, i.e. a sort of war *game*. Whether this concern is grounded or not seems to depend crucially on who drone operators are and what their training is. The more they are trained as traditional soldiers the less grounded the concern seems to be.

Finally, some think that huge *economic investments* in new robotics technologies by states may let politicians and military commanders be manipulated in a different way. Here the idea is that even if it is indeed in principle possible to draw ethical limitations in the use of war robots, once a huge economic investment in these technologies is made and the economic advantage in their use becomes relevant (both for private companies providing technologies and for the state itself) economic considerations will take the upper hand over the ethical and political agenda of the state; and war robots will be in the end used also beyond

the reasonable limits suggested by ethical considerations. Since this may be the topic for a different piece, we will restrict ourselves to two general considerations. First, the possibility of *this* kind of manipulation of political power by economic interests is ubiquitous and does not seem to be a specific argument against the use of *robotics technologies* in war, anymore that it is an argument against investment in military equipment in general.[14] Secondly, whereas it is in general true that once a given technology has been produced, economic interests may unduly influence the choices of political power, in the case of emerging technologies, the state may at least keep the power to direct its own investment in research toward the design and productions of weapons that embed its own desired values and goals, according to the ideas of socially responsible innovation and value-sensitive design[15].

Authors

Ezio Di Nucci is Assistant Professor of Philosophy at Universität Duisburg-Essen, Germany.

Filippo Santoni de Sio is a Post-Doctoral Fellow at Delft University of Technology, The Netherlands.

[14] It may be true that the more expensive a technology is, the more pressing is the above risk. But, again, that would not be a specific point about drones.

[15] J. van den Hoven, 'Value Sensitive Design and Responsible Innovation', in R. Owen, J. Bessant and M. Heintz (eds.) *Responsible Innovation - Managing the Responsible Emergence of Science and Innovation in Society* (West Sussex: John Wiley, 2013) 75-84.