Chapter 12
Autonomous Weapons: Terminator-Esque Software Design

Seumas Miller

12.1 Introduction

Autonomous robots can perform many tasks for more efficiently than humans, for instance, tasks performed in factory assembly lines, auto-pilots, driverless cars; moreover, they can perform tasks dangerous for humans to perform, say, defuse bombs. However, autonomous robots can also be weaponized, and in a manner, such that the robots control their targets (and, possibly, the selection of their weapons). As Sarah Connor, the character in the 1984 Hollywood movie, *The Terminator*, discovered, autonomous weapons are utterly fearless; they don't have emotions, and care nothing for life over death. Further, by virtue of developments in artificial intelligence programming, robots will have superior calculative and memory capacity—this is not just a fantasy limited to the fictitious company, Cyberdyne Systems.

New and emerging (so-called) autonomous robotic weapons can replace some military roles performed by humans and enhance others.[1] Consider, for example, the Samsung stationary robot that functions as a sentry in the demilitarized zone between North and South Korea. Once programmed and activated, it has the capability to identify, track and fire its machine guns at human targets without the further intervention of a human operator. Although, not humanoid, they are "Terminators."

[1]For an earlier account of these issues, see Seumas Miller "Robopocolypse?: Autonomous Weapons, Military Necessity and Collective Moral Responsibility" in (ed.) Jai Galliott and M Lotze, *Super Soldiers: The Ethical, Legal and Social Implications* (Ashgate 2015), pp. 153–166.

S. Miller (✉)
Australian Graduate School of Policing and Security, Charles Sturt University, Canberra, Australia

Delft University of Technology and the University of Oxford, Delft, The Netherlands
e-mail: semiller@csu.edu.au

H. Prunckun (ed.), *Cyber Weaponry*, Advanced Sciences and Technologies for Security Applications, https://doi.org/10.1007/978-3-319-74107-9_12

Fig. 12.1 An MQ-1 Predator on its return to Bagram Air Base, Afghanistan, after an Operation Enduring Freedom mission. Predators have been used for armed interdiction, as well as intelligence, surveillance, and reconnaissance (Photograph by U.S. Air Force Master Sgt. Demetrius Lester and courtesy of the U.S. Air Force)

Predator drones are used in Afghanistan and the tribal areas of Pakistan to kill suspected terrorists. While the ones currently in use are not autonomous weapons they could be given this capability in which case, once programmed and activated, they could track, identify and destroy human and other targets without the further intervention of a human operator. Additionally, more advanced autonomous weapons systems, including robotic devices, are being planned.

In this chapter, the moral implications of autonomous robotic weapons are explored. This will be done by addressing several questions. Firstly, in what sense are such weapons really autonomous? Secondly, do such weapons necessarily compromise the moral responsibility of their human designers, computer programmers and/or operators and, if so, in what manner and to what extent? Finally, should certain forms of autonomous weapons be prohibited? (Fig. 12.1)

12.2 Autonomous Weapons

Autonomous weapons are weapons system which, once programed and activated by a human operator, can—and, if used, do in fact—identify, track and deliver lethal force without further intervention by a human operator. By *programmed* I mean, at least, that the individual target, or type of target, has been selected and coded into the weapon system. By *activated* it is meant, at least, that the process culminating in the already programmed weapon delivering lethal force has been initiated. This weaponry includes weapons used in non-targeted killing, such as autonomous anti-aircraft weapons systems used against multiple attacking aircraft or, more futuristically,

against swarm technology (for example, multiple lethal miniature attack drones operating as a swarm so as to inhibit effective defensive measures); and ones used or, at least, capable of being used in targeted killing (for example, a predator drone with face-recognition technology and no human operator to confirm a match).

We need to distinguish between what are termed the *human in-the-loop*, *human on-the-loop,* and *human out-of-the-loop* weaponry. It is only human out-of-the-loop weapons that are autonomous in the required sense. In the case of human-in-the-loop weapons the final delivery of lethal force (for example, by a Predator drone), cannot be done without the decision to do so by the human operator. In the case of human on-the-loop weapons, the final delivery of lethal force can be done without the decision to do so by the human operator; however, the human operator can override the weapon system's triggering mechanism. In the case of human out-of-the-loop weapons, the human operator cannot override the weapon system's triggering mechanism; so, once the weapon system is programmed and activated there is, and cannot be, any further human intervention.

The lethal use of a human-in-the-loop weapon is a standard case of killing by a human combatant and, as such, is presumably, at least in principle, morally permissible. Moreover, other things being equal, the combatant is morally responsible for the killing. The lethal use of a human-on-the-loop weapon is also in principle morally permissible. Also, the human operator is, perhaps jointly with others (such as his or her commander—see discussion below on collective responsibility as joint responsibility), morally responsible, at least in principle, for the use of lethal force and its foreseeable consequences. However, these two propositions concerning human on-the-loop weaponry rely on the following assumptions:

(1) The weapon system is programmed and activated by its human operator and either;
(2) (a) On any and all occasions of use, the delivery of lethal force can be overridden by the human operator; and, (b) this operator has sufficient time and sufficient information to make a morally informed, reasonably reliable judgement whether or not to deliver lethal force or;
(3) (a) On any one occasion of use, but not all, occasions of use, the final delivery of lethal force can be overridden by the human operator; and, (b) there is no moral requirement for a morally informed, reasonably reliable judgement on each and every occasion of the final delivery of force.

A scenario illustrating (3)(b) might be an anti-aircraft weapons system being used on a naval vessel under attack from a squadron of manned aircraft in a theatre of war at sea in which there are no civilians present.

There are various other possible such scenarios. Consider a scenario in which there is a single attacker on a single occasion in which there is insufficient time for a reasonably reliable, morally informed judgment. Such scenarios might include ones involving a kamikaze pilot or suicide bomber. If autonomous weapons were to be morally permissible the following conditions at least would need to be met: (i) prior clear-cut criteria for identification/delivery of lethal force to be designed-into the

weapon and used only in narrowly circumscribed circumstances; (ii) prior morally informed judgment regarding criteria and circumstances, and; (iii) ability of operator to override system. Here, there is also the implicit assumption that the weapon system can be switched off, as is not the case with, for instance, biological agents released by a bioweapon.

What of human out-of-the-loop weapons, i.e. autonomous weapons? These are weapons systems that once programmed and activated can identify, track and deliver lethal force without further intervention by human operator. They might be used for non-targeted killing in which case there is no uniquely identified individual target such as in the above described cases of incoming aircraft and swarm technology. Alternatively, they might be used for targeted killing. An example of this would be a Predator drone with face-recognition technology and no human operator to confirm match. However, the crucial point to be made here is that there is no human on-the-loop to intervene once the weapons system has been programmed and activated. Three questions now arise. Firstly, are these weapons systems autonomous in the full-blooded sense of moral autonomy in common use in relation to many, if not most, freely performed, morally informed human actions? (They are "morally informed" because taking someone's life is a morally significant action and, therefore, the person taking this life ought to be making a morally informed decision.) Secondly, are humans fully morally responsible for the killings done by autonomous weapons or is there a so-called responsibility gap? Thirdly, should such weapons be prohibited? Let us begin by getting a better understanding of the notion of moral autonomy.

12.3 Moral Autonomy

In respect of the notion of an autonomous agent, whether human, Martian or otherwise, two sets of distinctions need to be kept in mind.[2] The first is between rationality and morality. An autonomous agent is a rational agent. However, arguably, being rational is not a sufficient condition for autonomy. Rather an autonomous agent needs also to be a moral agent.

The second distinction pertains to sources of potential domination. An autonomous agent is one whose decisions are not externally imposed; he or she is not dominated by external forces or other persons. Nevertheless, an autonomous person is also possessed of self-mastery; he or she is not dominated by internal forces, for instance, addictions.

Evidently, an autonomous person is both rational and moral. So, what is it to be a rational person? Evidently a rational person is possessed of a continuing, rationally integrated structure of mental attitudes, such as intentions, beliefs and desires. Moreover,

[2]Stanley Benn, *A Theory of Freedom* (Cambridge: Cambridge University Press, 1988) and Seumas Miller "Individual Autonomy and Sociality" in (ed.) F Schmitt *Socialising Metaphysics: Nature of Social Reality* (Lanham: Rowman & Littlefield, 2003), pp.269–300.

the attitudes in question, notably beliefs, are evidence-based. In short, the mental attitudes of a rational person are both rationally coherent and based on evidence.

Second, a rational person's actions and dispositions to action are based on such coherent and evidence-based attitudes. So their actions are rational in the light of their mental attitudes (which are themselves rational).

Third, for a person's attitudes and actions to be rational in this sense the person must surely engage in both practical (action-oriented) and theoretical (knowledge-oriented) reasoning that makes use of objectively valid procedures; such as deriving valid conclusions from evidence and selecting means on the basis of their efficacy in respect of relevant ends, and so on.

Fourth, the concept of a rational person or being needs to be relativized to empirical circumstances; including inherent properties of the kind of rational beings in question. And, it is possible that there are rational persons who are not human beings, for example Martians or creatures from a far-flung and yet undiscovered planet. If so, then such non-human rational beings might not have all the inherent properties that human beings have. For example, human beings, but not necessarily other rational beings, have emotions, are highly social, live for a finite number of years, and so on. Naturally, a rational being will act rationally in the light of such additional inherent properties (as well as contingent external features of their environment).

Fifth, a rational being can engage in rational scrutiny of their extant higher order attitudes, such as beliefs about their own beliefs. If, for example, a rational person is engaged in self-deception (and, as a consequence, has false beliefs about one's own motives) then, at least in principle, such a person can come to recognize and eliminate this self-deception.

Sixth, evidently, rationality in the sense in question admits of degrees; some people, for example, are better than others at drawing true conclusions from the evidence presented to them.

Seventh, rational beings or, at least, fully rational beings are able to choose their ultimate ends, i.e. those ends that are not simply the means to further ends. Perhaps one's own personal happiness is an ultimate end chosen by many in individualistic social groups, although human beings can choose different ultimate ends, e.g. high social status, great political power, justice for the poor and downtrodden, the survival of future generations threatened by climate change etc. If it is argued that the ability to choose ultimate ends is not a necessary condition for rationality it can be replied that it is certainly a necessary condition for autonomy. For if a creature did not choose its ultimate ends then those ends must surely have been brought about either by the intervention of some other creature, or by some inanimate causal process. Either way, the autonomy of the creature in question is compromised.

Even if it is held that robots could, at least in principle, be possessed of the first six features of rational agency, it is not the case that they could be possessed of the seventh feature; robots cannot choose their ultimate ends since these are programmed, or otherwise designed, into them.

So, much for rationality; what of morality? Someone can be rational, up to a point, without necessarily being moral. Consider, for example, a highly intelligent

psychopath. Such a person may well pursue their goals efficiently and effectively and make sophisticated evidence-based judgments in doing so. Accordingly, psychopaths can be highly rational. However, psychopaths do not care about other people and are happy to do them great harm if it suits. Likewise, psychopaths, even if they recognise in some sense the constraints of morality and pay lip-service to them, do not feel the moral force of moral principles and ends. In short, psychopaths can be rational and yet are not moral agents. Therefore, rationality and morality seem to be different, albeit related, concepts.

To return to an earlier point, perhaps rationality is relativised to inherent properties. If so, since psychopaths lack some of the inherent properties of other human beings; for instance, concern for the welfare of others, a moral sense, their rationality is more restrictive and, to this extent, they are less rational than their fellow human moral agents. If this is correct then arguably psychopaths are not simply non-moral or less than fully moral, they are also less than fully rational. At any rate, roughly speaking, a human moral agent is a rational agent who is disposed to make true judgments and valid inferences in relation to the moral worth of human actions, principles, ends, and so on, and to act on those judgments and inferences where appropriate. More generally, moral agents are rational agents who are sensitive to moral properties in the sense that they recognise moral properties as such and respond appropriately to them. While robots are sensitive to physical properties, e.g. heat and light, they are not sensitive to moral properties. Accordingly, robots are not moral agents.

Here it is worth noting the distinction between non-rational and irrational agents, and between non-moral and immoral agents. A non-rational agent *cannot* make judgments or inferences. An irrational agent has the capacity to make such judgments and inferences, but has some significant deficit in their rationality, and thus makes a significant number of false judgments and/or invalid inferences, or often fails to act on the results of their practical reasoning. Similarly, a non-moral agent lacks the capacity to make moral judgments and act on them; an immoral agent, by contrast, is merely (significantly) deficient in their moral judgment-making, or often fails to act on their correct moral judgments. That said, sometimes it is not clear whether we should think of a person as non-moral (or non-rational) or as immoral (or irrational).

Given that a human life involves sensitivity to moral properties and, relatedly, responsiveness to moral reasons, an autonomous human being will be both rational and moral. Understood in the way outlined here, rationality and morality imply independence and self-mastery. Someone who is dominated by the overriding desire to please an authority figure, and who only acts in accordance with that aim, will not count as autonomous. Similarly, the autonomous human being must be able not only to make good judgments about what to believe and how to act, they must be capable of acting in conformity with those judgments. A drug addict, for example, may know perfectly well that it is unwise to keep injecting drugs, but find themselves unable to act on that knowledge; their lack of self-mastery in respect of their desire for the drug means that they lack autonomy, at least in this area of their life.

To say that an autonomous human being is independent and possesses self-mastery, does not, of course, imply that autonomy is incompatible with all forms of constraint. The autonomous person cannot infringe the laws of physics or the laws of logic. The fact that a human agent cannot hope to fly when they jump off a tall building, or cannot both walk and not walk at the same time, does not undermine their autonomy. Further, an autonomous person can choose to comply with the law without comprising their autonomy. Specifically, when human beings choose to comply with laws because these laws embody their moral beliefs, principles and ends, then they may well be acting autonomously; the laws in question being in effect self-imposed.

In light of the above we can now see that autonomous human beings are ones who decide for themselves what is important and valuable to them, and possess the capacity to make reason-based choices on the basis of recognising, assessing and responding to relevant considerations, including non-moral facts, moral principles and ultimate moral ends. When we call an act autonomous, we mean that it is something done by such a person, on the basis of such a response. As we have seen, robots are not autonomous beings in this sense.

Moreover we can also now see that autonomy admits of degree and is in part constituted by various moral features, including freedom of thought and action None of us, presumably, is completely autonomous, since we all fall short of full rationality, perfect morality, absolute self-mastery and so on. And since these things vary from person to person, some people are more autonomous than others. Moreover, someone might be autonomous in one area of their life, but not another. Nevertheless, we achieve the status of an autonomous human being – someone who is entitled to decide for themselves how they wish to live – when we are sufficiently autonomous. Further, autonomy can be undermined if one or more of its constitutive moral features are compromised, e.g. if a person is imprisoned.

There is a presumption that all human adults, at least, have achieved that status. This presumption is defeasible. We may be able to show that a person is so deficient in various conditions of autonomy, such as rationality or self-mastery, that they should not be counted as autonomous, and that others might be justified in making decisions on their behalf. But, absent such defeat, we all possess the status of autonomous human beings. By contrast, there is no such presumption to be defeated in the case of robots.

12.4 Moral Responsibility and Autonomous Weapons

Let us now return to *so-called* autonomous weapons—that is human out-of-the-loop weapons.[3] We have seen that autonomous robots and, therefore, autonomous weapons are not autonomous in the sense in which human beings are since they do not choose their ultimate ends and are not sensitive to moral properties. However,

[3]This section is an abridged version of Seumas Miller Chap. 10 in his *Shooting to Kill: The Ethics of Military and Police Use of Military Force* (Oxford University Press, 2016).

the question that now arises concerns the moral responsibility for killings done by autonomous weapons. Specifically, do they involve a responsibility gap such that their human programmers and operators are not morally responsible or, at least, not fully morally responsible for the killings done by the use of these weapons?[4]

Consider the following scenario, which, I contend, is analogous to the use of human out-of-the-loop weaponry. There is a villain who has trained his dogs to kill on his command and an innocent victim on the run from the villain. The villain gives the scent of the victim to the killer-dogs by way of an item of the victim's clothing and then commands the dogs to kill. The killer-dogs pursue the victim deep into the forest and now the villain is unable to intervene. The killer-dogs kill the victim. The villain is legally and morally responsible for murder. However, the killer-dogs are not, albeit they may need to be destroyed on the grounds of the risk they pose to human life. So, the villain is morally responsible for murdering the victim, notwithstanding the indirect nature of the causal chain from the villain to the dead victim; the chain is indirect since it crucially depends on the killer-dogs doing the actual physical killing. Moreover, the villain would also have been legally and morally responsible for the killing if the 'scent' was generic and, therefore, carried by a whole class of potential victims, and if the dogs had killed one of these. In this second version of the scenario, the villain does not intend to kill a uniquely identifiable individual,[5] but rather one (or perhaps multiple) members of a class of individuals.[6]

By analogy, human out-of-the-loop weapons—*killer-robots*—are not morally responsible for any killings they cause.[7] Consider the case of a human in-the-loop or human-on-the-loop weapon. Assume that the programmer/activator of the weapon and the operator of the weapon at the point of delivery are two different human agents. If so, then other things being equal they are jointly (that is, collectively) morally responsible for the killing done by the weapon (whether it be of a uniquely identified individual or an individual qua member of a class).[8] No-one thinks the weapon is morally or other than causally responsible for the killing. Now assume this weapon is converted to a human out-of-the-loop weapon by the human programmer-activator. Surely this human programmer-activator now has full individual moral responsibility for the killing, as the villain does in (both versions of) our

[4]Ronald Arkin ("The Case for Ethical Autonomy in Unmanned Systems" *Journal of Military Ethics* vol. 9 2010 pp. 332–341) has argued in favour of the use of such weapons.

[5]It is not a targeted killing.

[6]Further, the villain is legally and morally responsible for foreseeable but unintended killing done by the killer-dogs in the forest, if they had happened upon one of the birdwatchers well-known to frequent the forest and mistakenly killed him instead of the intended victim. (Perhaps the birdwatcher carried the scent of birds often attacked by the killer-dogs.)

[7]See R. Sparrow "Killer Robots" *Journal of Applied Philosophy* vol. 24 2007 pp. 63–77. For criticisms see Uwe Steinhoff "Killing them safely: Extreme asymmetry and its discontents" in B. J. Strawser (ed.) *Killing by Remote Control: The Ethics of an Unmanned Military* (Oxford: Oxford University Press, 2013).

[8]Each is fully morally responsible; not all cases of collective moral responsibility involve a distribution of the quantum (so-to-speak) of responsibility.

killer-dog scenario. To be sure there is no human intervention in the causal process after programming-activation. But the weapon has not been magically transformed from an entity only with causal responsibility to one which now has moral or other than causal responsibility for the killing.

It might be argued that the analogy does not work because killer-dogs are unlike killer-robots in the relevant respects. Certainly, dogs are minded creatures whereas computers are not; dogs have some degree of consciousness and can experience, for example, pain. However, this difference would not favor ascribing moral responsibility to computers rather than dogs; rather, if anything, the reverse is true.

Clearly, computers do not have consciousness, cannot experience pain or pleasure, do not care about anyone or anything (including themselves) and, as we saw above, do not choose their ultimate ends and, more specifically, cannot recognize moral properties, such as courage, moral innocence, moral responsibility, sympathy or justice. Therefore, they cannot act for the sake of principles or ends understood as moral in character, such as the principle of discrimination.

Given the apparent non-reducibility of moral concepts and properties to non-moral ones and, specifically, physical ones,[9] at best computers can be programmed to comply with some non-moral proxy for moral requirements.[10] For example, "Do not intentionally kill morally innocent human beings" might be rendered as "Do not fire at bipeds if they are not carrying a weapon or they are not wearing a uniform of the following description." However, here as elsewhere, the problem for such non-moral proxies for moral properties is that when they diverge from moral properties, as they inevitably will in some circumstances, the wrong person will be killed or not killed (as the case may be)—as an example, the innocent civilian wearing camouflage clothing to escape detection by combatants on either side and carrying a weapon for personal protection is killed while the female terrorist concealing a bomb under her dress is not.

Notwithstanding the above, some have insisted that robots are minded agents; after all, it is argued, they can detect and respond to features of their environment and in many cases they have impressive storage/retrieval and calculative capacities. However, this argument relies essentially on two moves that should be resisted and are, in any case, highly controversial. Firstly, rational human thought, notably rational decisions and judgments, are down-graded to the status of mere causally connected states or causal roles, for example via functionalist theories of mental states. Secondly, and simultaneously, the workings of computers are upgraded to the status of mental states, for example via the same functionalist theories of mental states. For reasons of space I cannot here pursue this issue further. Rather I simply note that this simultaneous down-grade/up-grade faces prodigious problems when it comes to the ascription of autonomous agency. For one thing, autonomous agency

[9]The physical properties in question would not only be detectable in the environment but also be able to be subjected to various formal processes of quantification and so on.

[10]See, for instance, Arkin "The Case for Ethical Autonomy in Unmanned Systems," op. cit. and the reply in Miller *Shooting to Kill,* Chap. 10, op. cit.

involves the capacity for non-algorithmic inferential thinking, for example the generation of novel ideas. For another, to reiterate, computers do not choose their own ultimate ends. At best, they can select between different means to the ends programmed into them. Accordingly, they are not autonomous agents, even non-moral ones. So, while killer robots are morally problematic this is not for the reason that they are autonomous agents in their own right but this brings us to our third and final question.

12.5 Prohibition of Autonomous Weapons

Our final question concerns the prohibition of autonomous weapons in the sense of human out-of-the-loop weapons. This question should be seen in the light of our conclusions that such weapons are not morally sensitive agents and their use does not involve a responsibility gap. Rather there are multiple human actors implicated in the use of autonomous weapons: there is collective moral responsibility in the sense of joint individual moral responsibility.[11] The members of the design team are collectively—that is, jointly—morally responsible for providing the means to harm (the weapon). The political and military leaders and those who follow their orders are collectively (i.e. jointly), responsible for these weapons being used against a certain group/individual. Take, for example, intelligence personnel who are responsible for providing the means to identify targets, and the operators who are responsible for its use on a given occasion since they programmed/activated the weapons system. Moreover, all the above individuals are collectively—in the sense of jointly—morally responsible for the deaths resulting from the use of the weapon, but they are responsible to varying degrees and in different ways; for instance, some provided the means (designed the weapon), others gave the order to kill a given individual, still others pulled the trigger, etc. These varying degrees and varying ways are reflected in the different but overlapping collective end content of their cooperative or joint activity. Thus, a designer has the collective end to kill some combatants in some war (this being the purpose of his design-work); a military leader has the collective end (in issuing orders to subordinates) to kill enemy combatants in this theatre of war; and an operator the collective end to kill enemy combatants A, B & C, here and now.

It is important to note that each contributor to such a joint lethal action is individually morally responsible for his/her own individual action contribution—an individual weapons operator who chose to deliver lethal force on some occasion or perhaps, in the case of an on-the-loop weapon, not to override the delivery of lethal force by the weapon on this occasion. This is consistent with there being collective, that is, joint, moral responsibility for the outcome—the death of an enemy combatant, the death of innocent civilians.

[11] Seumas Miller "Collective Moral Responsibility: An Individualist Account," in Peter A. French (ed), *Midwest Studies in Philosophy*, vol. XXX, 2006, pp.176–193

It is also important to note the problem of accountability that arises for morally unacceptable outcomes involving "many hands", that is, joint action, and indirect causal chains. Consider, for example, an out-of-the-loop weapon system that kills an innocent civilian rather than a terrorist because of mistaken identity and the absence of an override function when the mistaken identity is discovered at the last minute. The response to this accountability problem should be to design-in institutional accountability. Thus, in our example the weapons designers ought to be held jointly institutionally and, therefore, jointly morally responsible for failing to design-in an override function, that is, for failing to ensure the safety of the weapon system; likewise, the intelligence personnel ought to be held jointly institutionally and, therefore, jointly morally responsible for the mistaken identity. Analogous points can be made with respect to the political and military leaders and the operators.

As we have seen, human-*out*-of-the-loop weapons can be designed to have an override function and an on/off switch controlled by a human operator. Moreover, in the light of our above example and like cases, in general autonomous weapons ought to have an override function and on/off switch. Indeed, to fail to do so would be tantamount to an abnegation of moral responsibility. However, against this it might be argued that there are *some* situations in which there ought not to be a human on-the-loop (or in-the-loop).

Let us consider some candidate situations involving human *out*-of-the-loop weapons that might be thought not to require a human in or on the loop.

(1) Situations in which the selection of targets and delivery of force cannot in practice be overridden on all occasions and in which there is no requirement for a context dependent, morally informed judgement on all occasions e.g. there is insufficient time to make the decision to repulse an imminent attack from incoming manned aircraft and there is no need to do so since the aircraft in a theatre of war are clearly identifiable as enemy aircraft.
(2) Situations in which there is a need only for a computer-based mechanical application of a clear-cut procedure (e.g. deliver lethal force), under precisely specified input conditions (e.g. identified as an enemy submarine by virtue of its design etc.) in which there is no prospect of collateral damage (e.g. in open seas in the Arctic).

However, even in these cases it is difficult to see why there would be an objection to having a human on the loop (as distinct from in the loop) especially since there might still be a need for a human on the loop to accommodate the problems arising from false information or unusual contingencies. For instance, the 'enemy' aircraft or submarines in question might turn out to be ones captured and operated by members of one's own forces. Alternatively, one's own aircraft and submarines might now be under the control of the enemy (e.g. via a sophisticated process of computer hacking) and, therefore, should be fired upon.

A further argument in favour of autonomous weapons concerns human emotion. It is argued that machines in conditions of war are superior to humans by virtue of not having emotions since stress/emotions lead to error. Against this it can be pointed out that human emotions inform moral judgment and moral judgment is

called for in war. For instance, the duty of care with respect to innocent civilians relies on the emotion of caring; a property not possessed by robots. Moreover, human stress/emotions can be controlled to a considerable extent, e.g. combatants should not be combatants if not appropriately selected/trained, and the influence of stressors can be reduced, e.g. by requiring some decisions to be made by personnel at some distance from the action.

The upshot of this discussion is that human out-of-the-loop weapons are neither necessary nor desirable. Rather autonomous weapons should always have a human on-the-loop (if not in-the-loop). Furthermore, not to do so would be an abnegation of responsibility. Accordingly, autonomous weapons in the sense of human out-of-the-loop weapons should be prohibited.

12.6 Summary

In this chapter, certain aspects of the morality of autonomous weapons has been discussed. Specifically, autonomous weapons have been described and the sense in which such weapons are autonomous specified. Autonomy was defined as it applies to human beings and it has been argued that autonomous weapons are not autonomous in this sense. The claim that there is a responsibility gap in the use of autonomous weapons has been discussed and it has been concluded that in fact there is no such gap. Human beings are fully morally responsible for the killings involving the use of autonomous weapons. Finally, it was suggested that human out-of-the-loop weapons are not desirable, indeed they are inherently morally problematic; as such, they should be prohibited.

Principal Concepts

The principal concepts associated with this chapter are listed below. Demonstrate your understanding of each by writing a short definition or explanation in one or two sentences:

- Autonomous weapons;
- Moral autonomy;
- Human in-the-loop;
- Human on-the-loop;
- Human out-of-the-loop; and
- The responsibility gap.

Study Questions

1. Explain what is meant by a rational being.
2. Describe the seven principles of rational agency.
3. Explain what is meant by moral autonomy.

Learning Activity

In some scholarly circles, it has been argued that human out-of-the-loop autonomous weapons should be prohibited. Reflecting on the discussion in this chapter, debate the reasoning for-and-against such a proposition. In particular, is a blanket band on such weapons warranted? Could ethical safeguards for out-of-the-loop weapons be developed; or is the issue so problematic that it is not worth pursuing from a policy point of view?