1. BACKGROUND, SCOPE AND OUTLINE

Robots are emerging in many areas. We read about them in the newspapers almost every day; about drone strikes in war, about robots in nursing homes keeping the elderly company or about cars with the ability to park or even drive without human supervision.[1] When we make a phone call to a company and get to talk to a computer, it seems like the world is becoming more and more automated. And the robots are evolving; research is being made on producing more sociable robots, which are able to read facial expressions and hold conversations.[2]

The development seems to go from industrial robots, confined in certain areas, to machines and systems with increasing levels of autonomy, with the singularity – machines outsmarting humans – and science fiction movies like *Terminator* or *I, Robot* at the far, more speculative end of the spectrum. What happens when the property of autonomy is possessed by a machine or technical system? What philosophical questions are evoked? This thesis will look at some of these questions.

There is no universally accepted definition of autonomy. A basic idea is the ability to be off on one's own, making decisions of one's own, without the influence of someone else, but definitions vary between contexts. In philosophical debates on free will, the criteria for autonomy can be very demanding, whereas in other areas – such as that of robots – a vacuum cleaner or a car can be considered autonomous. Autonomy also seems to come in degrees. For instance, Volvo will introduce autonomous cars – driving on their own for up to 50 kilometers per hour in "traffic jams" – in 2014.[3] Driving "off on its own" does not, however, mean that the car is

[1] Examples of areas where robots are used (summary from Lin et al. 2012, pp. 5–6): Labour and services – half of the over seven million service robots are vacuum cleaners; Military and security; Research and education – laboratory experiments and in the field as well as in classrooms; Entertainment – such as toys; Medical and healthcare – like surgical robots, intelligent prosthetics helping stroke patients to walk, decision support systems; Personal care and companions; Environment – cleaning up after disasters, handling polluted areas.

[2] There is research on making robots lifelike and able to make facial expressions as well as reading facial expressions, and hold a conversation (Wrenn 2012). At the University of Tsukuba there is a robot who, with 97 % accuracy, can interpret smiles and frowns – something that is used in order to teach the robot how to behave (Barribeau 2011). There is also Charles the GPS-car who can read the driver's facial expressions (Chung 2011).

[3] Robotnyheter 2012 (in Swedish) <http://www.robotnyheter.se/2012/10/24/sjalvkorande-bilar-fran-volvo-redan-2014/>

driving to pick up the kids all by itself while the owner is at home preparing dinner. But the car possesses some degree of autonomy, and one could easily suspect that this is only the beginning. There is a development towards increasing autonomy, particularly in the military, where robots are even lethal (Arkin 2009). We are reassured that there will always be a human "in the loop", but this is not as clear-cut as it may seem. There are, for instance, indications that humans tend to trust systems perhaps too much, as indicated by the Aegis-case in 1988 – where a passenger plane was shot down.[4] Some argue that "killer robots" should be banned, something we will return to in section two.

Autonomous system is used as a covering term for different terms in this thesis, like robot, artificial agent and UAV (unmanned aerial vehicle). It is the autonomy which is the relevant feature for the philosophical inquiries in the thesis, not necessarily the embodiment usually associated with robots. An autonomous system embedded into something we would not call a robot is just as relevant. The term robot is most commonly used, however, as it is in the literature.

There does not seem to be a real consensus among roboticists on how to define the term robot (Bekey 2012, p. 18), but autonomy or independence is a common denominator. If one were to define a robot as, for instance, "a computer with sensors and actuators that allow it to interact with the external world", then a computer connected to a printer would be a robot (Lin et al. 2012, p. 17). Something more is required – some level of "thinking" or "acting" separating it from automats, like thermostats or the computer connected to a printer. Arkin uses

[4] Aegis is an integrated naval weapons system with different levels of autonomy, introduced in the 1980s to help defend navy ships against air and missile attacks. The system had four modes: *semiautomatic*, in which the humans interfaced with the system to judge when and what to shoot; *automatic special,* in which the human controllers set the priorities, such as telling the system to destroy bombers before fighter jets, but the computer then decided how to do it; *automatic*, in which data went to human operators in command, but the system worked without them; and *casualty*, where the system just did what it thought best to keep the ship from being hit. Humans could override the Aegis computer in any of its modes. In 1988 it spotted Iran Air Flight 655, an Airbus passenger jet, who was on a consistent course and speed and was broadcasting a radar and radio signal that showed it to be civilian. The automated Aegis had, however, been designed for managing battles against attacking Soviet bombers in the open North Atlantic, not for dealing with civilian-filled skies in the crowded Gulf. Aegis registered the passenger plane with an icon on the computer screen and made it seem to be an Iranian F-14 fighter (half the size) and hence assumed that it was an enemy. Even though the hard data told the crew that the plane was not a fighter jet, they trusted the computer more. Aegis was on semi automatic mode, but not one of the 18 soldiers and officers was willing to challenge the computer. They authorized it to fire, killing 290 passengers and crew, including 66 children (Singer 2009, p. 125). This is sometimes called "automation bias" which is a tendency to trust an automated system "in spite of evidence that the system is unreliable or wrong in a particular case" (Asaro 2009).

separate definitions for *robot* and *autonomous robot*. A *robot* is "an automated machine or vehicle, capable of independent perception, reasoning, and action". An *autonomous robot* is "a robot that does not require direct human involvement except for high-level mission tasking; such a robot can make its own decisions consistent with its mission without requiring direct human authorization, including decisions regarding the use of lethal force" (Arkin 2009, pp. 50–51). The distinction seems to hinge on authorization rather that autonomy, however, since even the "simple" robot can perceive, reason and act.

Thinking – or reasoning – is a loaded term when applied to machines, dating back to the "debate" between Alan Turing, asking the question "Can Machines Think?" (Turing 1950) and John Searle's thought experiment "The Chinese Room" (1980), which will be described further in section 4, along with discussions on whether autonomy – in a stricter sense – is possible in machines. But Arkin's definition of robot works rather well for the purposes of this thesis. Even though it involves terms which can be considered vague, a loose definition involving independence, reasoning and embodiment will serve the purpose for a principled discussion on robots connected to issues like ethics and agency.

The philosophically interesting questions surrounding autonomous systems can be divided into different categories.

One category of questions concerns the ethics of using autonomous systems. There are for instance issues of rights, privacy, safety and dignity. Should we replace humans with robots? What about unemployment? If soldiers are replaced by robots, will the laws of war be affected? This thesis has – regarding this category of questions – a focus on the ethics of using robots in war. A lot of money is being spent on military robots – robots that are even lethal. It is argued that the use of autonomous robots in war may affect the laws of war. This is discussed in papers I: "Is it morally right to use UAVs in war" and II: "Autonomous Robots[5] in War – Undermining the Ethical Justification for Killing?"

A second category of questions is how we can make sure that robots behave the way we want – or, behave *ethically* – when they are "off on their own". Dodig-Crnkovic and Çürüklü (2011) point out that it is not clear today how engineers should design

[5] It should, however, be mentioned that paper II talks of "autonomous robots" – even though this introduction states that "robot" involve autonomy – in order to emphasize the possibility of replacing entire armies.

the software so that its decisions would be effective reaching given goals, as well as ethically sound. The answer to that problem is sought within machine ethics/machine morality (e.g. Allen et al 2005, 2006; Wallach and Allen 2009) or roboethics (Lin et al 2012, chapter 22).[6] There are two main approaches: the top-down, where rules or an ethical theory is programmed into the robot, and the bottom-up, where the robot learns about morality much like a child does. This type of question is discussed in paper III, "Robots and the ethics of care", where one of the normative moral theories – the ethics of care – is discussed and applied to health care robots.

It has been argued that that the development of cognitive machines with "built-in" ethics can help us understand ourselves and the mechanisms of ethical behaviour; to learn about ethics by building artificial moral agents (e.g. Allet et al 2006; Coeckelbergh 2012; Noorman 2012). "Just as AI has stimulated new lines of enquiry in the philosophy of mind, machine morality has the potential to stimulate new lines of enquiry in ethics. Robotics and AI laboratories could become experimental centers for testing theories of moral decision making in artificial systems" (Wallach & Allen 2009, p. 9). Would we like robots to be utilitarians or deontologists, for instance? Ensuring the ethical behaviour of robots by programming an ethical theory into them, might force us to "agree" on what type of morality we would like to see in robots or to scrutinize moral theories in a different light.

An example of the possibility for a deeper understanding of the mechanisms of ethical behaviour would be discussions on the importance of emotions in moral agents. Autonomous war robots with synthetic emotions such as guilt are being developed in order to learn from their mistakes based on that synthetic guilt: "Our laboratory has considerable experience in the maintenance and integration of emotion into autonomous system architectures […] Guilt is said to result in proactive, constructive change [Tagney et al. 07]. In this manner, guilt can produce underlying change in the control system for the autonomous agent" (Arkin 2009, pp. 141–142). Research is also being done on attempts to model artificial mirror

[6] Roboethics according to Lin et al is human centered; not the ethics of robots but the ethics of robotics researchers (Lin et al 2012, p. 348), whereas machine ethics/machine morality can incorporate the morality of the machines themselves. These topics are also investigated under the headings of computer ethics and philosophy of technology.

neurons in robots in order to make robots at least behave *as if* they were compassionate (Lin et al 2012, p. 75).

A third category of questions concerns agency and responsibility, indicating a possible or potential uniqueness of autonomous systems compared to other technologies. The questions in this category take us into a rather controversial and sometimes speculative area, with discussions and suggestions which may be hard to swallow at times. Can there be such a thing as an artificial moral agent, for instance – responsible for its own actions? Or is this a "non-question"; something that can be negated instantly, without further investigation, since there is always a human programming the artificial "*agent*"?

The moral agency in and the ascription of responsibility to autonomous systems has been discussed by several authors (e.g. Dodig-Crnkovic and Çürüklü 2011; Floridi and Sanders 2004; Johnson 2006; Matthias 2004; Stahl 2004; Wallach & Allen 2009). Dodig-Crnkovic and Çürüklü believe that the ethical aspects of autonomous AI have been insufficiently researched until now, partly based just on what they believe to be the misconception that intelligent artefacts just do what they have been programmed to do (e.g. Dodig-Crnkovic and Çürüklü 2011; Lin et al 2008).

Agency and responsibility in artificial agents are discussed in paper IV and V in this thesis: "The Functional Morality of Robots" and "The Pragmatic Robotic Agent".

Section five in the introduction will be about connecting today with the future. How do we make the right decisions today, in order to deal with something extremely uncertain? How do we navigate between science fiction and what seems possible today?

According to Singer (2009), technology prognostications of nonscientists often fail because they do not pay close attention to what is technically feasible and not:

> […] Scientists' predictions tend to overstate the positive, especially when it comes to war. Franklin, Edison, Nobel and Einstein, for example, all thought that their inventions would end war. They knew the science, but not the social science. Both groups tend to disregard how social change can intervene and intertwine with technology, yielding not one definite future, but rather many possible futures (Singer 2009, p. 14).

This quote indicates that interdisciplinary work is important for identifying and assessing the future threats and risks of potentially so called disruptive technologies.[7] This line of thinking was picked up by the CopeTech project[8] in which this doctoral thesis had its starting point. The purpose of the CopeTech project was to develop methods for dealing with future, potentially disruptive technologies. This was done by incorporating participatory and co-evolutionary elements. The basis for a method was developed and presented in the article "Assessing Socially Disruptive Technological Change" – paper VI in this thesis – and further developed and tested in a workshop where the developing method was applied to autonomous systems, a potentially disruptive technology which the project chose to focus on.

Many argue that the area of intelligent robots is one of the most promising future emerging technologies (e.g. Gates 2007; Warwick 2009). It has also been argued that the development of intelligent machines with the ability to assess the effects of their actions on others and adjust their behavior accordingly "may ultimately be the most important task faced by the designers of artificially intelligent automata" (Allen et al. 2000, p. 251).

Another reason for CopeTech's choice to focus on autonomous systems was that it had not received much attention from research on risk and threats in the future. This made it an interesting case for developing a method involving co-evolution and participatory elements; to systemize discussions between people from different areas, looking into the future. The work of CopeTech is developed in paper VII in this thesis: "A co-evolutionary creative prototyping approach for the future of autonomous robot systems and potential implications on the civil protection sector", describing the ground work for a method from paper VI, contrasting it with other co-evolutionary methods and applying it to the case of autonomous systems. This will be described further in section five.

[7] A disruptive technology or innovation is something that overtakes, substantially changes or creates a new market with a different set of values (Christensen 1997). The term "disruptive technology" has been used as a synonym of "disruptive innovation", but according to Christensen (2003) the latter is now preferred, because market disruption has been found to be a function usually not of technology itself but rather of its changing application. In this thesis (as in the CopeTech project) the term technology will be used since increasing autonomy in robots is more a development than an innovation, at least so far.

[8] The CopeTech project was a collaboration between KTH Royal Institute of Technology and FOI (The Swedish Defence Research Agency).

2. ETHICAL ISSUES CONCERNING THE USE OF ROBOTS IN WAR

There are many ethical issues surrounding the increasing use of robots, like unemployment, conflicting rights, loss of status, integrity and safety. In this thesis, there is a focus on ethical issues surrounding military robots; how the use of such robots may affect the laws of war. It can be argued that military robots have a more significant impact ethically compared to other robots today, since at least some of them – what is commonly known as drones (unmanned aerial vehicles – UAVs), for instance – are involved in killing people.[9] Voices have been raised to restrict or even prohibit the development of "killer robots", since lethal robots with complete autonomy would – according to The Human Rights Watch, for instance – be incapable of meeting international humanistarian law (HRW 2012, p. 3).[10]

The main question in paper I is whether it is morally right to use UAVs in war. Ethical issues surrounding the use of UAVs are systematized, focusing on UAVs today. The term autonomy in the military comes in different degrees when connected to robots, ranging from *remote controlled*, via *semi-autonomous* (systems with a degree of autonomy but remaining supervised) to – it is claimed – *fully autonomous* (Quintana 2008). Autonomy is not the most important feature today, however – UAVs today are mainly semi-autonomous; controlled remotely, often by operators situated on the other side of the globe, with some autonomy in navigation, for instance (Lin et al. 2008, 2012). Most UAVs are used for surveillance, reconnaissance and target destination, but some deploy weapons

[9] Even though service robots may be developing rapidly, they represent only a fraction of the expenditures on military robots (Lin et al 2012, p. 109).

[10] The Human Rights Watch report 2012 (here abbreviated HRW 2012): "Losing Humanity – The Case against Killer Robots". The Human Rights Watch and Human rights program at Harvard Law School recommend prohibition of "the development, production, and use of fully autonomous weapons through an international legally binding instrument", advocate adoption of "national laws to prohibit the development, production and use of fully autonomous weapons", and to "commence reviews of technologies and components that could lead to fully autonomous weapons". (HRW 2012, p. 5.) This is partly due to the "Marten's Clause" (1899, 1907 Hague Convention, codified in Art 1(2)): "In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from dictates of public conscience" (HRW 2012, p. 25). See also Roberts and Guelef (2010), p. 70.

(Quintana 2008). The UAVs today do not make decisions to kill, and it is often pointed out that there is always a human in the decision-loop – but there is a development towards increasing autonomy. The US Department of Defense "envisions unmanned systems seamlessly operating with manned systems while gradually reducing the degree of control and decision making required for the unmanned portion of the force structure."[11] Unmanned robots can actually be divided into three categories regarding human involvement: (i) human-in-the-loop: select targets and deliver force only with a human command, (ii) human-on-the-loop: select targets and deliver force under the oversight of a human operator who can override the robot's actions, and (iii) human-out-of-the-loop: capable of selecting targets and delivering force without human input or interaction (HRW 2012, p. 2).

UAVs have advantages such as being able to move close to targets without exposing human crews, so the number of casualties can be significantly lower with the use of UAVs. Another advantage is that there are fewer requirements for sophisticated and expensive long-range missiles, since the UAV can move in close to a target. UAVs can move faster than airplanes with humans on board, and can stay in the air for days or even weeks. The ability to function autonomously without the assistance or control of an operator – category (iii) above – enables a lower probability of being detected by enemy sensors, since there is no signal between the operator and the UAV.[12]

Arguments against the use of autonomous robots in war are, for instance, that the issue of responsibility would be unclear (Sparrow 2007) and that the threshold of entering war would be lower, if the risk of losing lives would be decreased (Asaro 2008). Also, it has been argued that robots would not be able to comply with the laws of war (HRW 2012, Sparrow 2009). According to Asaro, the use of autonomous technologies is neither completely morally acceptable, nor completely morally unacceptable under Walzer's interpretation of just war theory.

[11] US Dept of Defense (2011) "Unmanned systems Integrated Roadmap FY 2011-2036; "Reference Number 11-S-3613, 2011, p. 13 <http://www.defenseinnovationmarketplace.mil/resources/UnmannedSystemsIntegratedRoadmapFY2011.pdf> Accessed May 2013.

[12] Thanks to Magnus Bengtsson at the Swedish National Defence College, for explaining this to me.

It might be argued that ethical evaluations of weapons used in war – such as UAVs, irrespective of their level of autonomy – are meaningless since war is unethical in itself. The ethical evaluation in paper I is made against the backdrop of the laws of war (LOW), as codified in, for instance, the Geneva and Hague conventions. The rules of *jus ad bellum* specify what criteria must be fulfilled in order to start a war, where "just cause" is the most important one. The rules of *jus in bello* establish criteria for ethical means of fighting once at war.

The rules of *jus ad bellum* and *jus in bello* are summed up below[13]:

Jus ad bellum:

- *Just cause*: The reason for going to war needs to be just and cannot therefore be solely for recapturing things taken or punishing people who have done wrong; innocent life must be in imminent danger and intervention must be to protect life. Examples: self defense from external attack, punishment for a severe wrongdoing which remains uncorrected. This is the first and most important rule.
- *Right intention*: The state must intend to fight the war only for the sake of its just cause. Force may be used only in a truly just cause and solely for that purpose – correcting a suffered wrong is considered a right intention, while material gain or maintaining economies are not.
- *Last resort*: All peaceful and viable alternatives have been seriously tried and exhausted or are clearly not practical.
- *Legitimate authority:* War is only between states.
- *Reasonable chance of success*: The idea is that a state's resort to war must be considered to have a measurable impact on the situation.
- *Proportionality*: The anticipated benefits of waging war must be proportionate to its expected evils or harms. (Also known as the principle of macroproportionality to separate it from the *jus in bello* principle of proportionality).

Jus in bello:

- *Proportionality/Excess*: An attack cannot be launched on a military objective if the civilian damage would be *excessive* in relation to the military advantage – the value of an attack must be in proportion to what is gained.
- *Discrimination*: Only military targets and enemy combatants can be attacked.

[13] Summed up from Orend (2008) and the conventions, printed in Roberts and Guelef (2010).

- *Necessity*: The attack must be necessary (just war should be governed by the principle of minimum force). This principle is meant to limit excessive and unnecessary death and destruction.
- *Weapons*: All international laws on weapons prohibitions must be obeyed, such as the ban on chemical and biological weapons. Nuclear weapons are considered taboo.

There is also *jus post bellum* which concerns regulations connected to war termination; to ease the transition from war to peace (Orend 2008)[14]. The effects of robots on *jus post bellum* are not treated in this thesis.

It might be argued that it would be sufficient to look solely at *jus in bello*, since the UAV is something that is used once *in* war. In paper I it is argued that the possession of UAVs might affect the interpretation of *jus ad bellum* as well, since UAVs might increase the inclination to start a war. The reason for this, it is argued in paper I, is that UAVs have advantages in terms of reducing casualties for the UAV possessor, and may make war seem more like a risk-free enterprise – in extreme cases even like a computer game – thereby lowering the threshold for starting a war. The possession of UAVs may – more than other weapons – also affect the interpretation of the LOW, for it may determine which normative moral theory the interpretation of the LOW will be based on.

When looking at weapons used today, it might be argued that there is no morally relevant difference between UAVs and other weapons. According to Asaro, it is important to note that "even if robots did make it easier for a nation to go to war, this in itself does not decide whether that war is just or unjust" (Asaro 2008, p. 48). In paper I, it is argued that there are relevant differences between UAVs and other weapons. First of all: compared to other weapons that might give one country the option to win a war without losing any lives of its own soldiers – like chemical, biological or nuclear weapons – UAVs are permitted according to the LOW. Nuclear weapons are not formally prohibited (which chemical and biological weapons are), but are considered taboo and have not been used in war since World War II. A complete ban of nuclear weapons is being considered by the United Nations. Among permitted weapons today, UAVs may, more than other weapons, provide the owner with a severely increased inclination to start a war against a

[14] There is little international law regulating *jus post bellum*, so one must turn to the moral resources of just war theory (Orend 2008).

country that does not possess the same technology. UAVs are also different from long-range missiles in being more flexible. A UAV may go closer to a target without risking the life of the "pilot" – that is, the operator, who is often situated on the other side of the globe. This is another aspect of UAVs, making warfare dangerously similar to a computer game and therefore increasing the inclination to fire. Distance is one of the most important factors when it comes to firing at other human beings, which with UAVs is combined with experiencing no risk for one's personal safety (Grossman 1996).

One problem with the LOW, pointed out in paper I, is that they are too open for interpretation, and that different normative moral theories might provide conflicting results – to the advantage of the UAV possessor. Three terms – "just" ("just cause"), "necessary" and "excessive" – are discussed in paper I, with utilitarianism, deontology and virtue ethics as a backdrop. The conclusion indicates the importance of revising the LOW or adding some rules that focus specifically on UAVs. For instance, if a country that possesses UAVs intends to start a war against a country that does not, then it is particularly important to determine that the cause really is just.

The ethical issues regarding UAVs in war today – which paper I attempts to systemize – concern the implications of remoteness rather than autonomy, whereas paper II, "Autonomous Robots in War: Undermining the Ethical Justification for Killing?" has a more futuristic approach. There it is argued that where large parts of – or even entire – armies would be replaced by robots that are autonomous to a high degree, the justification for killing, as interpreted by just war theory, would be substantially undermined. A necessary criterion, it is argued, is based on reciprocity of risk, something that may be eliminated with autonomous robots.

The main difference between traditional war theory – which the LOW are based on – and the challenging views, is the implicit assumption of moral equality of combatants in the traditional view. In paper II it is argued that reciprocal imposition of risk is a necessary condition for the moral equality of combatants. This is supported by the fact that the LOW has a strict division between *jus ad bellum* and *jus in bello*, and by quotes from traditional war theorists regarding threat and harm. It is also argued that advanced, autonomous robots violate the principle of reciprocal imposition of risk, and thereby substantially undermine the

ethical justification for killing. It is investigated whether autonomous robots create a special type of asymmetry (risk imposition) – aside from asymmetry due to strength or asymmetry regarding goals – but concluded that it is a subcategory to that of strength. If one belligerent uses unmanned robots, the ethical assumptions that the LOW rest on become substantially undermined. And the ethical justification cannot be transferred.

Paradoxically (even though there is a sharp division between *jus ad bellum* and *jus in bello*), the traditional view on justification, with its justification for killing in war based on the moral equality between combatants, for which the reciprocal imposition of risk is a necessary condition, may find it difficult to permit the use of autonomous robots in war. The justification for a robot to kill a human according to the implicit assumptions in the LOW is substantially undermined, as the LOW stand today. The question is more open with the challenging views – something that has to be established case by case, depending on whether the cause is just, and whether the robot-combatants are just or unjust combatants.

To prevent or even prohibit the development of military robotics may not be a viable option, but there may be a need to revise or make additions to the LOW. Another suggestion for dealing with the emergence of robots in war is to consider challenging views regarding the justification for killing in war, which more extensive revisions of the LOW might be based on.

Another aspect connected to the use of robots in war is the fact that even though the UAVs today are not autonomous in the sense of pulling the trigger themselves, they are autonomous to a certain degree, and can assist the operator in different ways. This means that there is a relationship between one autonomous and one at least semi-autonomous agent, which potentially affects the conduct in war. This will not be discussed further in this thesis.

3. THE ETHICAL BEHAVIOUR OF ROBOTS

With increasingly autonomous robots moving about, in future decades part of our daily lives "as soldiers, as companions and as slaves" (Wallach and Allen 2009, p. 47), it is important to make sure that they behave adequately – and, in some situations, even ethically. Dodig-Crnkovic and Çürüklü argue that artificial morality should be seen as a "necessary companion to artifactual intelligence in artificial agents" (2011, p. 69).

Arkin (2009) points out that building a weapon that destroys everything that moves is easy for scientists, and that the true challenge is a machine that kills ethically and rethinks strategies. As mentioned in the previous section, the notion of "ethical killing" may sound not only controversial but plain awful, but the idea is – as already mentioned – killing regulated by the LOW. Arkin also points out that "[i]f a human being in the loop is the flashpoint in this debate, the real question is then, *at what level* is the human in the loop?" (Arkin 2009, p. 7, my emphasis). He points out that several military robotic automation systems "already operate at the level where the human is in charge and responsible for the deployment of lethal force, but not in a directly supervisory manner" (Arkin 2009, p. 7). We also have examples of humans trusting computers more than their own judgment (e.g. Singer 2009, pp. 124-125; Wallach and Allen 2009, pp. 40-42).[15]

There is research on how to make UAVs more autonomous and – possibly – being allowed to decide when and where "to pull the trigger". Today humans, not robots, interpret the LOW. But in the future, the robot may interpret these laws, at least the laws concerning conduct once at war, as specified in the rules of *jus in bello*. For instance, it is prohibited to "kill or wound an enemy who, having laid down his arms, or having no longer means of defense, has surrendered at discretion" (Orend 2008[16]). Another issue concerns discrimination – the idea that only combatants and

[15] For other discussions on trust connected to robots, see for instance Coeckelbergh (2012). For example, one may trust a cleaning robot to do what it is supposed to do – cleaning, for instance – but Coeckelbergh argues that robots do not do just what they are made for, and that an instrumentalist view of technology is inadequate. One example – influenced mainly by Heideggerian phenomenology – is the insight that technological artefacts "do" more than is intended by humans: they co-shape how we understand and act in the world (Ihde 1990; Verbeek 2005). See also the co-evolutionary elements in papers VI and VII.

[16] Stanford Encyclopedia of Philosophy; no page number.

military objects are legitimate targets. In order to cope with requirements such as that of discrimination, it is argued that humans can rely on situational and instinctive knowledge, something that is difficult to encode in a robot due to the limits of artificial intelligence (Arkin 2009).

Two methods for implementing ethical behaviour in robots are often discussed: the top-down and the bottom-up. The so called top-down method is about having a set of rules that can be turned into an algorithm and programmed into the robot. Examples are "the Golden Rule, the Ten Commandments, consequentialist or utilitarian ethics" (Wallach and Allen 2009, p. 84). Asimov's three laws of robotics (Asimov 1942) is a classic version of the so called top down-method. There is also an attempt with divine-command ethics (Bringsjord and Taylor 2012).

McLaren (2006), Moor (2006) and Wallach and Allen (2009) have pointed out the difficulties with top-down methods, summed up by Arkin (2009, p. 93):

1. The ethical laws, codes, or principles (i.e. rules) are almost always provided in a highly conceptual, abstract level.
2. Their conditions, premises or clauses are not precise, are subject to interpretation, and may have different meaning in different contexts.
3. The actions or conclusions in the rules are often abstract as well, so even if the rule is known to apply, the ethically appropriate action may be difficult to execute due to its vagueness.
4. These abstract rules often conflict with each other in specific situations. If more than one rule applies, it is not often clear how to resolve the conflict.

Arkin believes that "battlefield ethics are more clear-cut and precise than everyday or professional ethics, ameliorating these difficulties somewhat, but not removing them" (Arkin 2009, p. 94). Work is being done, however, to create an ethics code in order for military robots to understand the LOW. Examples on how to implement ethical rules are described in some detail by Arkin (2009; chapter 8: How to represent ethics in a lethal robot).

The bottom-up method is the idea that robots should learn morality, much like a child does. The idea of mimicking child development was expressed by Turing: "Instead of trying to produce a programme to simulate an adult mind, why not rather try to produce one which simulates a child's? If this were then subjected to an

appropriate course of education one would obtain the adult brain" (Turing 1950; 456).

One example of a bottom-up method would be building synthetic emotions into machines since emotions can be a tool for regulatinge behaviour (e.g. Arkin (1998, 2009); Becker (2006); Dodig-Crnkovic and Çürüklü; Coeckelbergh (2010); Fellous and Arbib (2005); Minsky (2006); Vallverdú and Cascuberta 2009)). One idea for military robots is to compile human behavioral responses towards various scenarios, with the help of previous battle simulations. The aim is to incorporate an "ethical governor" which "suppresses, restricts, or transforms any lethal behavior" and if a lethal behaviour is determined to have been unethical, the system must prevent or reduce the likelihood of that happening again via an "after-action reflective review or through the application of an artificial affective function (e.g. guilt, remorse or grief)" (Arkin 2009, pp. 65–66, 183–187).

According to Wallach and Allen (2009, pp. 113–114) bottom-up methods are easy to build when they are directed at achieving one clear goal, but "when the goals are several or the available information is confusing or incomplete" it is much more difficult to provide a clear course of action. Also, bottom-up approaches might lack "some of the safeguards that systems guided from the top down by ethical theories offer"– that the top-down methods seem "safer" (Wallach and Allen 2009, p. 114). They argue that there need to be a combination of top down and bottom up strategies in order to produce artificial moral agents. [17]

In this thesis there is no attempt to determine which method is better – top-down, bottom-up or a mix – but to explore one top-down method; a particular normative theory as a potential safeguard for robotic behaviour, namely *ethics of care* (EoC). In paper III this theory will be investigated and connected to robots.

EoC has not previously been given much attention in the literature on robots. Aimee van Wynsberghe (2012) may be the first to connect EoC to the design on robots. Her focus regarding robots, as well as the focus in paper III, is on robots in

[17] Wallach and Allen (2009, p. 79) point out that the term "top-down" is used in a different sense by engineers, who approach challenges with a top-down analysis through which they decompose a task into simpler subtasks. They use the term in a way that combine the philosophical (top-down norms, standards and theoretical approaches to moral judgment) and engineering sense of the approach.

health care (care robots), which poses different ethical questions compared to those of robots in war. For instance: "[care robots] require rigorous ethical reflection to ensure their design and introduction do not impede the promotion of values and the dignity of patients at such a vulnerable and sensitive time in their lives… [there] are no universal guidelines or standards for the design of robots outside the factory." (van Wynsberghe, 2012). See also Sharkey and Sharkey (2010) and Vallor (2011). [18]

Van Wynsberghe suggests that an ethics of care, combined with *care centered value- sensitive design*, should determine how care robots should be designed, since such a methodology would pay tribute to the care perspective rather than a pre-packaged ethical theory (van Wynsberghe 2012). She discusses the robots of today or the near future, while the discussion in paper III is intended to be relevant for future robots in possession of more autonomy.

It is not, however, entirely clear how the concept of "care" should be interpreted in a normative setting, for instance, as when being used in an ethical theory such as EoC. In paper III it is argued that one should distinguish between "care" in a wider sense, as the word is used in daily parlance, and "care" as a value for guiding moral action.[19] An important challenge will arise when we need to design robots that are able to make decisions that have ethical implications such as needing to choose which patient to help first; a scenario that might not be too far off in the future. Therefore it is important to be precise when discussing and defining the key concept of care as it is and should be used in a normative theory such as EoC. EoC has different versions, which makes the theory difficult to pin down, and paper III is a contribution to how the term "care" should be understood in an ethical, rather than a natural sense. There is a suggestion that ethical care should focus on the perceived care of the care receiver rather than the giver. The conclusion is that EoC may be a theory that would work well for robots in a health care setting.

[18] Van Wynsberghe points out that value-sensitive design (VSD) has been praised by computer ethicists and designers for its success in incorporating ethics in the overall design process of computer systems or ICT (Van den hoven 2007). According to Cummings (2006) VSD might guide the design process of other technologies as well.

[19] Coeckelbergh articulates the differences among "shallow", "deep" and "good" care, where shallow care refers to routine care that lacks intimate and personal engagement (Coeckelbergh 2010; 183). Plausible interpretations of different kinds of care, and the ethically relevant sense of care, are discussed in paper III.

4. ARTIFICIAL AGENCY AND RESPONSIBILITY

> Regardless of whether artificial morality is genuine morality, artificial agents act in ways that have moral consequences. This is not simply to say that they may cause harm – even falling trees do that. Rather, it is to draw attention to the fact that the harms caused by artificial agents may be monitored and regulated by the agents themselves. (Allen et al 2005, p. 149).

We will now enter a controversial, sometimes speculative area of autonomous systems and philosophy. It has to do with morality in the robots themselves, but not just in a regulatory way, making sure that robots with a high level of autonomy behave ethically adequate, as described in the previous section. There are many discussions in the literature on whether robots can be considered agents, and actually be responsible for their choice of action.[20]

One might concede that it is *possible* that there might be artificial agents someday, but argue that this would still be a pointless fact or evoke meaningless questions, since all you have to do if the artificial agent misbehaves, is to switch it off. Even if we may be justified in holding the robot responsible for its actions the same way we hold humans responsible, there would be no *point* in doing it, partly because assigning praise or blame has no meaning for an artificial agent (e.g. Floridi and Saunders 2004). We hold humans responsible partly because it is an incitement to change future behavior.

Dodig-Crnkovic and Çürüklü (2011) argue that claims that artificial agents cannot be assigned responsibility is based on the fact that blame and punishment has no

[20] In the free will debate, there is traditionally a focus on moral responsibility rather than agency (Jeppsson 2012). In papers IV and V it is assumed that if the robot is responsible it is an agent, and that if it is an agent, it should be held responsible. Some, like Floridi and Sanders (2004) separate responsibility and agency. I did not agree with this when the papers on agency were written, but concede that a separation, where responsibility comes in degrees even if someone or something is considered an agent, might be more practical.

meaning for an artificial agent. They suggest that such arguments can be met by counter-arguments from the safety culture, where the primary interest is about learning from experience rather than assigning blame. Ascribing responsibility as a regulatory mechanism would be relevant for a learning robot and for "bottom-up"-methods, which were mentioned in section three.

In the case of bad or intolerable behaviour of an intelligent artificial agent a corrective mechanism equivalent to regret or remorse can be introduced, by synthetic emotions for instance, as mentioned in the previous section.[21]

But we would not be able to *punish* a robot, the argument continues; the robot would have no incitement to behave well because of fear of being punished. Or could we? The American Bar Association[22] held a mock trial in 2003, where an intelligent computer contacted a lawyer when it found out that its owner was about to switch it off (Rothblatt and Angelica 2003).

Countries using the death penalty in some sense "switch" *people* "off", which might be objected to not only on humanitarian grounds, but also according to some views on free will, such as incompatibilism. According to such views, at least in more extreme forms (like van Inwagen's Consequence Argument) we cannot be (morally) responsible for our actions: "If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us" (van Inwagen 1983). This is not a view societies endorse, however; we do hold people responsible for their actions unless they are mentally ill.[23] And it can be argued – as in paper V, "The Pragmatic Robotic Agent" – that we should hold robots to the same standards as humans in this respect – regarding free will – and that humans in a deterministic sense also are programmed, by genetics and environment.[24]

[21] The development being made in the area of artificial emotions such as guilt etc. indicates that scientists are working towards robots someday having cognitive *and* emotional capabilities equal to humans, acknowledging the biological importance of emotions.

[22] An association for lawyers and law students.

[23] The criteria for what falls under that category differ, however, as we could see in the trial against the Norwegian mass murderer Anders Behring Breivik.

[24] Free will can be understood in two basic senses: (i) you are free if you do what you wish to do (but there also has to be some control over your wish) or (ii) it must be true that you *could have*

Noorman (2012) points out that it might be argued that psychological reasons - the fact that humans have a strong urge to place blame - may warrant an investigation of whether computers could be held responsible and if so, when. "While blaming a gun is clearly inappropriate (i.e., guns don't kill people, people kill people), blaming a computer is just plausible enough to tempt us to do so under the illusion that we might get away with it" (Noorman, 2012). With more advanced robots, it will be important to *clarify* where the responsibility lies. With a learning robot, it might be a combination of programming, learning, and specific use, which together create something that lacks a clear path of responsibility.

When discussing artificial agency and responsibility, the main issues usually concern consciousness and free will.[25] In this thesis there is one paper more focused on consciousness – paper IV: "The Functional Morality of Robots", and one more focused on free will – paper V: "The Pragmatic Robotic Agent". They both have a pragmatic stance, in the sense that humans are considered "model" moral agents. This pragmatic stance may seem to implicate acceptance of rather controversial philosophical standpoints, such as functionalism and compatibilism, but it is a tentative or conditional acceptance, based on the assumption that a pragmatic stance is more fruitful when discussing these matters connected to robots.

CONSCIOUSNESS

The traditional understanding of artificial agency usually requires consciousness, including mental states, such as beliefs and desires.[26]

---

acted differently in a particular situation. *Could have acted differently* can be interpreted in different ways: logically, epistemically, physically, ability-wise (separate general and particular ability), hypothetically (if I had chosen to do A1 I would have done A1), categorically.

[25] Causal responsibility and mental states are often considered necessary in order to decide whether an agent is morally responsible for an action (e.g. Dodig-Crnkovic & Cürüklü 2011, Nissenbaum 1994).

[26] See, e.g. Dennett (1976). Dennett's criteria for moral personhood are the following: (i) the entity must have rationality, (ii) we must be able to take an intentional stance towards it, (iii) it must be the target of a certain kind of attitude, (iv) it must be capable of reciprocity and thereby return the attitude, (v) it must be capable of communicating with others, (vi) the entity must be capable of self-consciousness. For a summary and discussion of Dennett's condition see, e.g. Gallagher (2007). For other discussions on necessary criteria of moral agency see e.g. Sullins (2006). Sullins argues that robots can be moral agents. The requirements are significant autonomy in terms of programming, ascription of intention, and behaviour that shows understanding of responsibilities to other agents. See also Himma (2009). According to Himma the issue of

The discussions by Turing and Searle regarding whether machines can think and have mental states such as understanding, is one of the most debated topics in the philosophy of mind (Cole 2009). In Turing's famous paper of 1950 he introduced what is now called the Turing Test to support his claim that machines can think. This is a test of a machine's ability to demonstrate intelligence. A human judge engages in a conversation with a human and a machine, both trying to appear human. The judge cannot see the participants. If the judge cannot tell the machine from the human, the machine has passed the test. Turing's view, identifying thoughts with states of a system "defined solely by their roles in producing further internal states and verbal output", has much in common with functionalism (Levin 2010).

Functionalism is a doctrine in the philosophy of mind, regarding mental states. The idea is that whether or not something is a mental state is determined by the way it functions or the role it plays in a certain system. That is, it "does not depend on its internal constitution" (Levin 2010). Searle objected to this by using his famous Chinese Room thought-experiment, which is designed to show that it is possible for a computer to pass a Turing Test without possessing genuine understanding or intelligence. It is an argument against the possibility of "true artificial intelligence" (Cole, 2009). Since the computer is "executing a program" but yet does not genuinely understand the conversation in which it participates, Searle argued that executing a program is insufficient for genuine understanding. He distinguished between two different hypotheses about AI. Strong AI is the idea that an artificial intelligence system can think and have a mind, whereas such a system according to weak AI can (only) *act as if* it thinks and has a mind.

According to Allen et al. the goal of artificial morality is to design artificial agents that can act *as if* they are moral agents (Allen et al. 2006). This is connected to the distinction between weak and strong AI and whether weak AI – "as if" – is sufficient for moral agency. Luciano Floridi and J.W. Saunders took a step away from the traditional perspective – the debate on whether machines actually need mental states etc. – by focusing on "mind-less morality", with the concept of a moral agent that is not exhibiting free will, mental states or responsibility (Floridi

whether an artificial moral agent is possible depends on whether it is possible for it to be conscious.

and Sanders 2004). They argue that the idea that moral agency presupposes consciousness is problematic:

> [The view that only beings with intentional states are moral agents] presupposes the availability of some sort of privileged access (a God's eye perspective from without or some sort of Cartesian internal intuition from within) to the agent's mental or intentional states that, although possible in theory, cannot be easily guaranteed in practice (Floridi and Sanders 2004, p. 365)

Their suggestion is that moral agenthood depends on levels of abstraction, interactivity and adaptability (Floridi and Sanders 2004, 2008).[27]

James Moor (2006) also discusses the "as if" approach, and argues that we cannot be sure that machines in the future will lack the qualities we now believe to be unique for human ethical agents. He suggests three categories of ethical agents: "ethical impact agents" (any machine that can be evaluated for its ethical consequences), "implicit ethical agents" (machines whose designers have made an effort to design them so that they do not have negative ethical effects), and "explicit agents" (machines that reason about ethics using ethical categories as part of their internal programming), and beyond all those – full moral agency (Wallach and Allen 2009, pp. 33–34).

Wallach and Allen distinguish between instrumental, functional and full moral agents: "We take the instrumental approach that while full-blown moral agency may be beyond the current or future technology, there is nevertheless much space between operational morality and 'genuine' moral agency. This is the niche we identified as functional morality" (Wallach and Allen 2009, pp. 26–27).

Coeckelbergh (2009) agrees with Floridi and Saunders that we should move away from "the traditional approach". He suggests that we replace the question about how moral non-human agents *really* are – in terms of mental states – by the question about the moral significance of appearance, which means that this view is a

[27] Floridi and Sanders (2004) use a "Method of Abstraction" for analyzing levels of abstraction. The level of abstraction is determined by how one *chooses* to describe a system and its context. This has also been connected to the Turing Test (Floridi and Sanders 2010).

version of the "as if" approach.[28] Coeckelbergh believes that we might as well remain agnostic about what really goes on behind the scenes and focus on the "outer" scene, the interaction, and how this interaction is co-shaped and co-constituted by how artificial agents appear to humans. He argues that humans are justified in ascribing moral agency and moral responsibility to those non-humans that appear similar – but that we ascribe moral status and moral responsibility *in proportion* to the apparent features, and in order to do this, he coins the terms "virtual agency" and "virtual responsibility". Coeckelbergh believes that virtual responsibility should be followed by virtual blame and virtual punishment. These terms refer to "the responsibility humans ascribe to each other and to (some) non-humans on the basis of how the other is experienced and appears to them" (Coeckelbergh 2009, p. 184). It is unclear if he really means that we should say "I hold you virtually responsible for me being late" to humans, which the quotation above would indicate – because we do not know what actually goes on inside the minds of humans either. He also argues that it is important to include artificial agents in our moral discourse *without* giving up what he calls the "folk" intuition that humans are special with regard to morality.

The suggestion in paper IV of this thesis is that we should take an "as if" approach – that robots should be considered moral agents if they can act *as if* they are. I agree with Coeckelbergh that humans are special with regard to morality, but the idea in papers IV and V is that humans should be considered "model" moral agents, to which robots should be compared. That is, we should use the same criteria for robots as for humans, when deciding whether someone or something is a moral agent.

Coeckelbergh asks why we should take humans as the standard, the model for moral agency. One reason is that morality can be considered a human construction. Another important reason is the "other minds problem", as indicated by Floridi and Sanders (2004). We do not know whether fellow human moral agents have consciousness etc., since we cannot put ourselves inside their minds. In terms of the

[28] Torrance (2008) argues that appearance-based ethics (as if) can be a supplement to reality-based ethics, but not a replacement. Sparrow and Sparrow (2006, p. 141) argue that "[it] is not only misguided, but actually unethical, to attempt to substitute robot simulacra for genuine social interaction".

necessity of being organic in order to be a moral agent[29], there are, in paper IV, two examples that support the suggestion that we should not be biased against the potential agency of non-organic entities; the transferring of a human mind into a computer, and aliens which turn out to be robots. The conclusion in paper IV is that a robot should be considered a moral agent if it can pass a moral version of the Turing Test.

THE MORAL TURING TEST

Instead of offering a discussion on whether consciousness should be a prerequisite for moral agency, it is argued in paper IV that we should take a functional approach to this issue. That is, we should accept a functionalistic explanation of behaviour. The main reason is the problem of other minds. How do we justify the universal belief that others have minds like our own? We reach this conclusion from evidence we gather via communication and observation. The idea in paper IV is that if robots can behave as if they are moral agents in the same sense as humans, then this "as if" is necessary and sufficient for ascribing moral agency to robots.

Allen et al. (2000) first introduced the Moral Turing Test (MTT). This can be considered a functionalist method for deciding if someone, or something, is a moral agent. Analogous to the classic Turing Test, as spelled out by Kim (2006, p. 144) the MTT could be phrased as follows: "If two systems are input-output equivalent, they have the same moral status; in particular, one is a moral agent in case the other is".

Limitations of the MTT have to do with the emphasis on articulating moral judgments. Wallach and Allen discuss the MTT and argue that although such a test may bypass disagreements on particular ethical issues since there is no correct answer to moral questions – what the MTT judge is looking for is the ability of moral reasoning – the focus on reasoning and justification is inappropriate. They also believe that indistinguishability is an incorrect criterion for moral agency (Wallach and Allen 2009, pp. 206–207).

[29] For discussions of the importance of being organic in order to be a moral agent, see for instance Torrance (2008). Torrance advocates a so called organic view, according to which an artificial humanoid (today) cannot count as a full-blooded moral agent. Sentience and teleology require biologically based forms of self-organization.

There are claims that the MTT cannot be endorsed as a criterion for strong AI or genuine moral agency (Allen et al. 2005, p. 154). But the move away from the debate on whether strong AI is necessary for moral agency is supported by the "other minds problem". According to Stahl (2004) computers need to take part in moral discourses in order to pass the MTT. He argues that computers are, at least in their current form, unable to "capture the meaning of information" and therefore fail to reflect morality in anything but a most basic sense (Stahl 2004, p. 67). He argues that in order to be able to process information, a subject needs more than mathematical rules – it requires understanding of meaning, which in turn requires several other factors, among them a physical experience and a being-in-the-world. In paper IV it is argued, however, that *if* a robot passes a MTT, then it can be considered a moral agent. Stahl's claims are of an empirical nature. He may be right in his claims, but they do not affect the suggestion in paper IV that if a robot passes the MTT it should be considered a moral agent and held morally responsible for its actions.

It can, however, be argued that this is not a necessary and sufficient condition for moral agency (unless we are discussing the criterion of consciousness alone); we also need to look at the robot's freedom to act.

FREE WILL AND PROGRAMMING

Another criterion for moral agency, aside from that of consciousness, is free will. In paper V it is suggested that the relevant view on free will when it comes to determining whether robots can be moral agents, is compatibilism, since that – implicitly – is the view held in society in general. It can be objected that a computer is always programmed by a human. But then, it can be argued that humans are also programmed – in a deterministic sense – but still held responsible.

Strawson (1998) describes compatibilists[30] as, generally, believing that to have free will – to be a free agent in choice and action – is to be free of certain types of

[30] There are different versions of compatibilism, which we will not delve into. Dennett advocates one of these versions, and makes use of his intentional stance to argue for compatibilism. He believes that just as the decision to adopt the intentional stance towards a system is a pragmatic one, so is a pragmatic decision to adopt toward a system the stance that it is a morally responsible person. Dennett calls this The personal stance (Dennett 1973, pp. 157–8).

constraints. The focus is on not being forced; that freedom has to do with not being physically or psychologically forced or compelled to do what one does. The idea is that your "character, personality, preferences and general motivational set may be entirely determined by events for which you are in no way responsible (by your genetic inheritance, upbringing, subsequent experience and so on)" (Strawson 1998). Compatibilists generally argue that it is not necessary to be in control of any of these things in order to have compatibilist freedom; compatibilist freedom is about "being able to choose and act in the way one prefers or thinks best given how one is" (Strawson 1998). The general idea of a conditional analysis of the ability to do otherwise is as follows: I would be able to do otherwise/could have acted differently if and only if it were the case that…., where the space is filled with some elaboration of "I had an appropriately strong desire to do so" or "I had different beliefs about the best available means to satisfy my goal", etc. (O'Connor 2008).

The question in paper V is whether a compatibilist, conditional analysis can plausibly be applied to a robot, given that the robot has been programmed by a human.[31] If the elaboration would be "the robot could have acted differently, if it had been programmed differently", that might indicate a crucial difference between humans and robots, making it impossible to incorporate robots into compatibilist freedom. But for sophisticated robots, the elaboration might just as well be "if it had different beliefs" or even desires, it is argued. This has to do with the origin of the action, and higher-order intentionality. Can an action ever originate inside a robot? The idea in paper V is to, in a sense, compare the programming in the robot with the deterministic "programming" in the human.

According to the pragmatic (compatibilist) view, humans have free will, but considering determinism, it can be argued that we are also programmed. We are programmed by evolution, genetics, upbringing and environment. One might, for instance, program a robot to follow Kant's categorical imperative. If the robot follows that imperative, it might be argued that it would not matter whether the categorical imperative was programmed into the robot or, as for humans, appears through rationality. According to Kant every rational being knows, by being rational, that it is right to follow the categorical imperative. In that case, it can be

[31] Two main issues due to determinism are (i) alternative possibilities – how "the power to do otherwise" or "could have acted differently" should be interpreted, and (ii) ultimate responsibility/origin.

argued that humans are in some sense programmed (maybe by their nature – evolution, that is) to be rational – and to follow the categorical imperative, at least if they are sufficiently rational. It might be objected that this indicates a crucial difference between robots and humans – that robots would not be able to *not* follow the categorical imperative, while most people do not follow it. That is why, in paper V, there is a focus on norms and the ability to violate norms.

The main point according to the pragmatic view on autonomy and agency is that the origin should be within the agent, but the "ultimate origin" (God, evolution, genes, what have you) can be left aside, when we are being judged by fellow humans or stand trial in a court of law. If we wish to apply the same criterion of autonomy to robots, the ultimate origin should also be disregarded. It is impossible to actually decide which of the robot and human is more programmed in the relevant sense in terms of unpredictability – but it is doubtful whether unpredictability is relevant for free will (even though that often comes up in discussions on robots, programming and agency). When you know a human well, you know how that person will behave most of the time. But it is not just unpredictability per se that should be relevant when discussing the extent to which humans are programmed; it is rather the deliberative distance between thought and decision/action and the ability to violate norms.[32]

ROBOTS AND SHARED RESPONSIBILITY

The previous, pragmatically based discussion concerned consciousness and free will as criteria for agency and responsibility. This section will in a sense be even more pragmatic, describing suggestions on how to deal with the issue of robots and responsibility in the real world, so to speak; that is, with at least one foot outside the philosophy seminar room.

There seem to be two main approaches to the moral responsibility of machines: that they either cannot be assigned responsibility, or that they can be assigned responsibility to various degrees (Dodig-Crnkovic and Çürüklu 2011). The second category – responsibility in degrees – sometimes involves more or less semantic elaborations, or very subtle distinctions, such as ideas about virtual responsibility

[32] Verhagen (2004), (2000), discusses the dicotomy between reactive and deliberative agents.

(Coeckelbergh 2009), functional or artificial responsibility and distinctions between accountability and responsibility; that humans cannot always be found accountable for certain kinds of software or even hardware and that artificial agents should be acknowledged as moral agents and be held accountable, but not responsible (Floridi and Sanders 2004, p. 372).

The second category has, however, brought forward attempts to clarify notions of shared responsibility. This is based on a discontent with the way the concept of responsibility has been used in relation to computing, where the view that an artificial agent is primarily an isolated entity has been dominant (Noorman 2012). Some argue that organizations, such as corporations and similar sociotechnological systems, have a collective group responsibility and that this group responsibility differs from individual responsibility (e.g. Dodig-Crnkovic and Çürüklü 2011; Floridi and Saunders 2004; Coleman 2008; Silver 2005).

Several philosophers have argued that responsibility cannot be properly understood without recognizing the active role of technology in shaping human action (Jonas 1984; Johnson and Powers 2005; Verbeek 2006; Waelbers 2009). Johnson, for instance, claims that even though a computer may not be a moral agent, it should be the case that the designer, user and the artefact itself *all* should be the focus of moral evaluation, since they all influence an action (2006, p.195).

Dodig-Crnkovic and Çürüklü (2011) point out the importance of morally significant behaviour. They argue that "agents with morally significant behaviour should have moral responsibility. In the case of a robot/softbot, it may only be functional artifactual responsibility. This is very limited in comparison to corresponding human competence, and comes in varying degrees" (2011, p. 64). This functional responsibility can be understood as a network of distributed responsibilities within a socio-technological system, and it is suggested that engineers should be guided by the precautionary principle in order to ensure ethically acceptable behaviour of artificial agents. Hellström (2013) introduces the concept "autonomous power" and use it to identify the type of robots that call for moral considerations. The idea is that this autonomous power and the ability to learn are decisive for assignment of moral responsibility to robots.

The relationship between the moral significance or status of artefacts and humans has been discussed by Verbeek – representing a strong view – and Illies and Meijers, representing a more moderate view.

According to the so called strong view on the moral status of technological artefacts, advocated by Verbeek (2006), both humans and technological artefacts can be moral agents, and technologies embody morality. This view involves claims that (1) technologies *actively* shape people's being in the world, (2) humans and technologies do not have a separate existence any more, (3) technologies have intentionality, leading to the conclusion that (4) moral agency is distributed over both humans and technological artefacts. As Peterson and Spahn (2011) point out, (2) would – alone – support (4), but then (1) and (3) would be difficult to make out. They also think that (2) is questionable. For instance, many technologies, such as sun-powered satellites, would continue to exist for decades even if humans were to go extinct. If (3) is true that would support (4), but Peterson and Spahn argue that Verbeek uses "intentionality" in a strange way and that the term should not be used to characterize the influence that technologies have on people's behaviour. They argue that we have no reason to believe (3). I will return to this in a moment, but let us just look briefly at the so called moderate view.

The moderate view, defended by Illies and Meijers (2009), is described as "an intermediate position that attributes *moral relevance* to artefacts without making them morally responsible or morally accountable for their effects". This does not entail that technological artefacts are (or can be part of) moral agents. As Peterson and Spahn (2011) point out, the notion of moral relevance needs to be rather strong and not just be about technological artefacts sometimes affecting the outcome of actions. Otherwise the moderate view would simply mean that technological artefacts – like natural phenomena, for instance – sometimes play a causal role in a chain of events.

It seems difficult to find a substantial middle ground between being a moral agent and not being a moral agent, in terms of having "relevance for moral actions". Peterson and Spahn (2011) advocate a third view, called the "Neutrality Thesis", concluding that technological artefacts are neutral tools that are at most bearers of instrumental value. They believe that technological artefacts can sometimes affect the moral evaluation of actions, but that artefacts "never figure as moral agents or are morally responsible for their effects". They continue to argue that technologies are not active in any reasonable sense (unlike Verbeek's claim (1)) – they are passive, and also, that it is clear that they do not possess intentionality. They actually argue that (3), the thesis about intentionality, is the least plausible of Verbeek's premises.

They agree that if Verbeek could really show that technologies have intentionality, this would support his claim that moral agency is distributed over both humans and technological artefacts (4). Of course "technologies" do not have intentionality, but I assume they mean technological artefacts. Peterson and Spahn may be too quick to dismiss the possibility of artefacts having intentionality, and they give no actual arguments to support their claim – they state that there is simply "no reason to accept" the claim in (3), and they thereby dismiss an entire debate in the philosophy of mind. Even if they are talking exclusively of the artefacts of today, they still seem to be running the risk of being biased against non-organic agency, which is discussed in paper IV in this thesis.

The claim that agency is distributed over both humans and technological artefacts seem to be a fruitful stance for a pragmatic view on responsibility. That responsibility should be seen as a regulatory mechanism is also in some sense a pragmatic stance – assuring proper behaviour in the future rather than assigning blame (Sommerville 2007).

At the beginning of this section it was mentioned that there are two main approaches: that robots either cannot be assigned responsibility or that they can be assigned responsibility to various degrees. An example of a very pragmatic stance when it comes to the responsibility of robots – applied to robots today and in the near future – is Miller's international Ad Hoc Committee for Responsible Computing (Dodig-Crnkovic and Çürüklü 2011). The first rule of five, draft 27:

> Rule 1: The people who design, develop or deploy a computing artifact are morally responsible for the artifact, and for the foreseeable effects of that artifact. This responsibility is shared with other people who design, develop, deploy or knowingly use the artifact as part of a sociotechnological system.

The question of robots and shared responsibility is something that needs constant discussion. Papers IV and V put forward rather strong claims, but in order to continue with a pragmatic and practical view, I side with the second approach – responsibility in various degrees. I do, however, believe that we need to have an open mind and consider the implications of the stronger claims, when robots get more advanced.

5. CONNECTING TODAY WITH THE FUTURE

When discussing future robots there is always a risk of collapsing into science fiction – or to neglect potential developments and risks altogether with a swift "that will never happen, so we do not have to think about it". Singularity – machines who outsmart humans – is one example of something that borders on science fiction.

Technological singularity is often described as a "greater than human" super intelligence or an intelligence *explosion* through technological means – where "technology appears to be expanding at infinite speed" – and intelligent machines design new machines without human intervention (Kurzweil 2005, pp. 23–24).

Some, like Nick Bostrom (2003a, b), argue not only that it *will* happen, but that it may happen this century, partly based on observations of exponential growth in various technologies. Others – like Gordon Moore, whose Moore's Law is often cited in support of singularity – argue that it never will.[33]

Bostrom (2003a) argues that it is likely that machines will outsmart humans within 50 years. In the article "When machines outsmart humans" he argues for the following: (1) Substantial probability should be assigned to the hypothesis that machines will outsmart humans within 50 years. (2) Such an event would have immense ramifications for many important areas of human concern and consequently, (3) serious attention should be given to this scenario.

Whether or not the singularity is possible, another discussion has to do with whether an intelligence explosion of this kind would be harmful (or beneficial) to

[33]Kurweil (2005, pp. 427–484) responds to different objections, like, for instance, the criticism from ontology (developing the Kurzweil's Chinese Room as an alternative to Searle's Chinese Room); criticism from incredulity; that exponential trends don't last forever; software stability; analog processing; the complexity of neural processing; the criticism from quantum computing; criticism from theism, holism and governmental regulation.

Pinker (2011) argues that the fact that you can visualize something in the future is not evidence that it is likely or possible and that there is no reason to believe in a coming singularity.

Another issue has to do with jobs. Ford (2012) argues that before the singularity could occur most routine jobs in the economy would be automated, since this would require a level of technology inferior to that of the singularity. This would cause massive unemployment and plummeting consumer demand, which in turn would destroy the incentive to invest in the technologies that would be required to bring about the singularity.

humans or even an existential threat.[34] When Bostrom (2002) discusses human extinction scenarios, he lists superintelligence as a possible cause.[35]

Reasons for the superintelligence being unfriendly are, for instance, scarce resources or that there would be no evolutionary motivation for friendliness (Berglas 2008)[36]. It is also argued that unfriendly AI would be easier to create. The friendly as well as the unfriendly AI would require large advances in recursive optimization process design (Yudkowski 2004), but friendly AI would also require an ability to make goal structures invariant under self-improvement so that it would not transform itself into something unfriendly. An unfriendly AI would be able to optimize for an arbitrary goal structure.

The issue of whether it is desirable or not to develop advanced robots is interesting as well as important.

Lintem (2007) talks of humans being seduced by the emergence of robots, and argues that there is a risk of being so caught up in the technological development that we do not see the structures that are being built at the same time – structures that captures the way we think, also luring us into putting ourselves in the back seat.

> It is imperative that we are not seduced by the techno-centric aura that constrains current development of socio-technical systems and it is important that our discussions do not encourage a techno-centric focus. There is a danger that technologists will find, in the notion of technological artifacts as team-players, justification for the perverse and fruitless pursuit of technological solutions at the expense of integrating

[34] In 2009, leading computer scientists, artificial intelligence researchers, and roboticists met at the Asilomar Conference Grounds in California. The goal was to discuss the potential impact of the hypothetical possibility that robots could become self-sufficient and able to make their own decisions. They discussed the extent to which computers and robots might be able to acquire autonomy, and to what degree they could use such abilities to pose threats or hazards. The conference attendees noted that self-awareness as depicted in science-fiction is probably unlikely, but that other potential hazards and pitfalls exist (Yudkowsky 2004).

[35] Bostrom (2002) points out that when we create the first superintelligent entity, we might make a mistake and give it goals that lead it to annihilate humankind, assuming its enormous intellectual advantage gives it the power to do so. For example, we could mistakenly elevate a subgoal to the status of a supergoal. We tell it to solve a mathematical problem, and it complies by turning all the matter in the solar system into a giant calculating device, in the process killing the person who asked the question.

[36] According to Yudkowski (2004) evolution has no inherent tendency to produce outcomes valued by humans, and there is a risk for AI behaving in a way not intended by its creators (like Bostrom's example of an AI originally programmed with the goal of manufacturing paper clips, but when achieving superintelligence, decides to convert the entire planet into a paper clip manufacturing facility (Bostrom 2004)).

> and supporting unique and critical human functionality. From the cognitive engineering perspective, we must combat this science fiction fantasy that technologists can somehow automate all critical human functions in case we end up with a system in which humans have no more than a peripheral role or even no roll at all (Lintem 2007).

This worry – about the structures that we unconsciously create when being too absorbed by developing advanced technology – and the discussion on the desirability of advanced robots, was somewhat addressed during the CopeTech workshop, indicating that such a methodology is useful for taking many different things into account. Singularity – which can be thought of as "the ultimate disruptive technology" – was not mentioned in the workshop however, showing that it is important to invite people from many different areas in order to cover as many future co-evolutionary scenarios as possible, but also to use relevant time frames.

The CopeTech methodology is based on a co-evolutionary scenario approach and the development of different evolutionary paths. It is based on four requirements of an assessment methodology for disruptive technologies, which are described in paper VI (Carlsen et al. 2010):

> 1. An assessment methodology for disruptive technologies should take several scenarios of society's development into account.
>
> 2. An assessment methodology for disruptive technologies should explore co-evolutionary paths for society and artefacts based on the technology.
>
> 3. Co-evolutionary scenarios to be used in the assessment methodology should be relevant from a policy perspective. They should therefore highlight ethically and politically controversial issues and put focus on situations in which a policy response is required.
>
> 4. A process should be designed for the involvement of relevant stakeholder groups and experts on the technology of interest as well as scenario methodology expertise.

In order to effectively assess challenges associated with future technologies, it was argued that it is necessary to be concrete regarding both the actual implementation of the technology in terms of actual prototypes of artefacts as well as the domain where these artefacts are put into use. It does not suffice to frame the problem like: "What are the future challenges of autonomous systems?" Concrete – but of course

hypothetical – prototypes of artefacts therefore have to be defined and the domain has to be determined.[37]

We noticed that none of the currently available methods satisfied these principles, but that some of them contain useful methodological elements which should be integrated in a more satisfactory methodology. The outlines of such a methodology, multiple expertise interaction, were proposed. It combines elements from several previous methodologies, including (1) interdisciplinary groups of experts that assess the potential internal development of a particular technology; (2) external scenarios describing how the surrounding world can develop in ways that are relevant for the technology in question; and (3) a participatory process of convergence seminars, which is tailored to ensure that several alternative future developments are taken seriously into account. In particular, CopeTech suggest further development of a bottom-up scenario methodology to capture the co-evolutionary character of socio-technical development paths.

The co-evolutionary methodology in paper VII consisted of three main steps:

Step 1: Developing prototype artefacts

The aim of this step was to develop prototypes of artefacts, based on the existing or expected user needs and applications that are deemed to be challenging in terms of opportunities, risks and potential ethical dilemmas. The goal was to create prototypes consisting of descriptions of robots that could be expected to be on the market within a time frame of ten years. The appropriate time frame is dependent on the application, including characteristics of the actual branch of technology (e.g., information and communication technologies vs. heavy infrastructures such as the railway sector). In the present case study, we arranged a workshop with a mix of participants: robotic experts (robotics researchers from universities and a robot producing firm), potential end-users from the sector of domestic security and safety (from private companies and governmental agencies) and members of the project group. The aim of the workshop was to identify possibilities, vulnerabilities and

[37] In this case study we focused on the domain of vital societal function within the sector "domestic security and safety" as defined by the Swedish Civil Contingencies Agency. Examples of services (cf. vital societal functions) in this sector include the rescue services, the police, supply of energy and supply of food.

ethical aspects related to the development and use of autonomous robots within the domestic security and safety sector.

The workshop dealt with the potential robotic systems in a ten year time frame, the market demand, ethical problems associated with robot applications and the need for society to react in one way or another. The ideas for potential artefacts were grouped into eight clusters. In order to prioritize among these clusters, the participants assigned votes to the clusters, based on the participants' judgments on the expected impact for the sector of domestic security and safety. The list consisted of eight ideas for artefacts of autonomous robots: 1) Fire-fighter robots to support human fire-fighters, 2) Safety controller robots, 3) Service and protection robots in shopping malls, 4) Portable robots that enhance human capabilities, 5) Unmanned aerial vehicle (UAV) robotic systems for detection and pursuit of criminals, 6) Fire detection and fire-fighting robots for use at home, 7) Life rescuing robots for accidents at sea, 8) Fire detection and fire-fighting robots for schools.

From this list of eight proposed prototypes, we selected three for further development. The prototypes were subsequently developed more in detail regarding advantages and future potential. The development of the prototypes was informed by a literature review, especially technology road mapping studies. The result of the creative prototyping was a description of the functionality of the prototype robotic systems.

Step 2: Constructing a hypothetical debate in society

The next step was to explore a hypothetical debate in society concerning security issues and ethical problems – already manifested or expected in the near future – associated with one or the other technological application of the robotic systems. After constructing the initial prototypes we evaluated advantages and potentials (expected developments) identified in the artefact prototypes. The prototypes formed the basis for outlining an ethical debate about the use of the robots. First, different attitudes and views on robots were identified as well as some plausible events triggering the debate. The evaluation was structured along the themes integrity, rights and conflicting interests, economy and security.

Step 3: Societal reaction

In this step, society's response to the experienced ethical problems and opportunities were discussed in the context of different future societies. In order to span a broader range of possible societal developments, we constructed simple "scenarios" of future societies. The idea was to explore how the different artefacts could develop in different societal contexts, depending for instance on how producers, the market and the public respond to the identified ethical and practical problems. Because we were interested in the acceptance or non-acceptance of technologies in society, we wanted to see how two opposite societal stances towards novel technologies would respond to the prototypes. Therefore, when examining the reaction from society, we looked at two different types of societies: 1) the *technology-skeptic (or conservative) society*, and 2) the *technology-positive society*. First, we made a description of these societies. Using these as a starting point, we then imagined different societal reactions, including governmental strategic principles, and the regulations and incentives used to influence the use of the technical artefacts for each scenario.

The next step would be to explore how future artefact designers and producers may respond to the regulations and incentives imposed by the different societies in the previous step. For example, new or updated artefacts may appear on the market, while others may disappear or become restricted to a certain application. This can be seen as an iteration of step one again, but starting from different prototype artefacts and a different society. The process can then continue for another loop, but now with at least two different development paths where new prototype artefacts are developed in different future societies. The scenario development process therefore leads to a tree-like scenario structure.

6. SUMMARIES OF THE PAPERS

These papers have been written at different occasions during the past five years. As a consequence, the thesis may contain some minor inconsistencies. Errata of previously published papers can be found after the summaries.

Paper I

Johansson, L. (2011) Is it Morally Right to Use Unmanned Aerial Vehicles (UAVs) in War? *Philosophy and Technology*, 24(3):279–291.

Several robotic automation systems, such as UAVs, are being used in combat today. This evokes ethical questions. This paper attempts to systemize those issues and provide a framework for organizing further ethical and policy inquiry regarding the use of UAVs in warfare.

In terms of *jus ad bellum*, there is a risk of increased inclination to start war since UAVs may make war seem more risk free. There may also be an increased sense of unfairness with the belligerent who does not have UAVs, leading to terrorism, for instance. Another *jus ad bellum*-issue has to do with the possibility to fight "secret" wars. When national security or intelligence agencies use UAVs aside from the military, there are difficulties in keeping the same level of transparency.

Arguments connected to *jus in bello* are, for instance, numbing. Operators may be based on the other side of the globe, making it all dangerously similar to a computer game.

Another ethical issue has to do with the fact that the laws of war are open for interpretation. Three important terms in the LOW can be interpreted differently depending on which normative theory you subscribe to. First, the term "just"—the reason for going to war must be just—is up for interpretation. The same applies to the terms "excessive" (an attack cannot be excessive in relation to the military advantage) and "necessary" (the attack must be necessary).

According to a utilitarian interpretation of a just cause, "just" might be connected to the maximizing of utility, and it is possible to argue for a local maximization rather than a global one. Kant's categorical imperative, "act only according to that maxim whereby you can, at the same time, will that it should become a universal law",

might very well provide an interpretation of "just" that would provide a country with the justification to start a war. Even if all six rules of *jus ad bellum* need to be fulfilled, the possession of UAVs might actually strengthen the case, in particular with regards to "reasonable chance of success". A virtue ethical interpretation of "just cause" is more difficult to determine. "Acting justly" would of course be virtuous – but virtue theory does not provide clear action guidance. This indicates that the term "just" may, based on any of the three normative theories mentioned above, be "hijacked" by a UAV possessor. That is, a UAV possessor inclined to start a war, perhaps against a country that does not possess UAVs, might find a suitable interpretation of "just", allowing him to start a war, with the LOW on his side.

The term necessary is also possible to interpret to one's own advantage. What is necessary may, if utilitarianism were to be interpreted in its purest form, mean almost anything, since this is a theory often connected to proverbs such as "the end justifies the means". It is, however, important to note that utilitarianism, strict or not, is seldom applied in war. It is certainly not likely that a country would consider the lives of the enemy as having the same value as those of one's own people, for instance. But it is also likely that most belligerents do not wish to use *any* means to obtain victory. However, when asking whether it is, in a certain situation, really necessary to kill so many, and in a certain manner, a utilitarian may say yes, if the total sum of utility is greater than if they are not killed. Proponents of deontology may act in the same way, by using the defense from unforeseen consequences and double effect, according to which it is morally indefensible to intend to harm an innocent person, but morally defensible to perform actions with good intended consequences, where harm is a foreseen but unintended consequence. For instance, in the waging of a just war civilian casualties might be permitted, but it is always forbidden to target civilians. Virtue ethics may, in terms of necessity, seem the safeguard from atrocities when belligerents are blinded by the wish for victory, but it is not easy to derive a determinate answer regarding what action is morally right. "Excessive" is a term closely connected to "necessary" in the LOW, and might function as a moral guardian for a belligerent who is too focused on the goal to care too much about the means, but since that is another unspecified term, it does not actually provide much guidance. Compare a utilitarian judgment with the ones of a Kantian or perhaps a virtue ethicist; it seems clear that judgments on what is necessary may differ radically.

To sum up, a country with the advantage of being in possession of UAVs might—based on the choice of a suitable normative theory – argue that the laws of war are on one's side. This points to the importance of revising the LOW, or adding some

rules that focus specifically on UAVs. For instance, if a country that possesses UAVs intends to start a war with a country that does not, then it is particularly important to determine that the cause really is just.

Paper II

Johansson, L. Autonomous Robots in War: Undermining the Ethical Justification for Killing? *Submitted manuscript*

The increased use of robots in war raises ethical questions. In this paper it is argued that the traditional view on the justification for killing in war - which employs an idea of moral equality between combatants - is undermined by the use of unmanned robots. The more autonomous the robot, and the higher the level of replacement, the more the justification is undermined. The reason for this is that when used to replace humans (rather than simply assist them) robots remove the reciprocal imposition of risk, which, it is argued, is a necessary condition for the moral equality of combatants.

According to the traditional view – implicit in the laws of war – the killings committed both by just and unjust combatants within a war are morally permissible. This indicates a "moral equality of soldiers". War is taken, by the traditional view, to be a large-scale exercise of the right of self-defence. Some, like Kahn (2002), argue that the assumption of moral equality of combatants is based on the principle of reciprocal imposition of risk. An important issue when discussing the ethical justification for killing in war, especially connected to robots, is whether reciprocal imposition of risk is a necessary condition for the moral equality of combatants. In this paper it is argued that highly autonomous robots substantially undermine the principle of reciprocal imposition of risk. It is also argued that this principle *is* a necessary condition for the moral equality of combatants.

The laws of war codify rules surrounding killing in war, but the ethical justification for killing is not explicit. It seems like the implicit justification for killing – in the laws of war – has to do with ideas on rights and self-defence, based on an assumption of moral equality between combatants. The idea is that soldiers on both sides in a war lose their right not to be attacked by each other, but retain their right of self-defence. Since the soldier risks his life, and his enemy risks his, they are allowed to kill each another. Note that this symmetry is supposed to exist between just and unjust combatants alike, that is, it does not matter whether the cause is just. These assumptions constitute the traditional view on the justification for killing in war, most famously represented by Walzer.

It is argued that the symmetry between combatants is considered to be the deep origin of their moral equality. Kahn (2002) argues that this discrimination rule is central to the ethics of warfare not because it separates the morally guilty from the innocent but because it delineates a domain of *threat* and that if combatants are no longer a threat they are no more appropriate targets than noncombatants. This, and quotes from Anscombe and Nagel, support the idea of reciprocal imposition of risk as a necessary condition for moral equality of combatants according to the traditional view. According to Anscombe "innocent" is not a term referring to personal responsibility (a central notion in the challenging view) but rather means "no harming". But the people fighting are harming, so they can be attacked, she argues. According to Thomas Nagel, innocent means "currently harmless" and is opposed not to "guilty" but to "doing harm." *Threat* can be interpreted as *risk*, and *harming* plausibly also may include *killing*.

Kahn suggests that the reciprocal imposition of risk disappears in asymmetric warfare, especially when warfare turns into what he calls policing. There are different types of asymmetry in war; in terms of strength and in terms of goals. It is not easy to measure difference in strength if there is asymmetry in goals; terrorism, for instance. Robots would be placed in the category of contributing to asymmetry in terms of strength or technology (rather than asymmetry in terms of goals). They may certainly contribute to the level of asymmetry – policing – that Kahn believes might undermine the justification for killing. Since the LOW are open to interpretation, there is a risk that a strong country may have too much freedom under the laws of war as they are formulated today. It is not, however, entirely clear that this would mean that the justification for killing would be substantially undermined if the laws of war as they stand are followed, according to some interpretation. If this reasoning is applied to advanced unmanned robots, however, a stronger argument emerges. If robots replace humans entirely in the battlefield – which is the very idea with many robots that are autonomous to a high degree – and the belligerent employing such robots thereby does not risk or lose any lives, there is such a profound asymmetry that one can argue that it is asymmetry of a third kind – aside from asymmetry based on (1) strength or (2) goals. This third type of asymmetry – which may be called risk imposition – can lead to a substantial undermining of the justification for killing in war that is implicit in the LOW. The principle of reciprocal imposition of risk is based on the fact that you yourself risk what you aim to take from someone else – your life. Even in asymmetric warfare of the "strength-category" (or "goal-category) there is a risk of losing one's life in war, although it may be so diminished that the justification for killing is threatened as Kahn suggests. It may, however, not be a question of a special type of asymmetry

but rather a subcategory to that of strength. But. advanced autonomous robots may call for revisiosor additions to the LOW.

Paper III

Johansson, L. (2013). Robots and the Ethics of Care. *International Journal of Technoethics* 4(1): 67–82.

In this paper, the moral theory *ethics of care* – EoC – is investigated and connected to care robots. The aim is twofold: first, to provide a plausible and ethically relevant interpretation of the key term *care* in EoC (which is, it is argued, slightly different from the everyday use of the term) suggesting that we should distinguish between "natural care" and "ethical care". The second aim is to discuss whether EoC may be a suitable theory to implement in care robots.

In this paper it is suggested that it is possible to make a distinction between *care ethics* and *ethics of care (EoC)*; that *care ethics* may be interpreted as "ethics in or surrounding the health care profession" on a general, wide note, whereas *ethics of care* refers to a normative theory that might be a rival to the other normative theories, such as utilitarianism and deontology.

There are a few key aspects that are common to all versions of EoC: (i) the conception of the moral agent as *relational* rather than sufficient and independent, (ii) permissibility to be *partial* and valuing particular others rather than seeing all persons as having equal moral worth, (iii) a *lack of principles* or right making criteria that are present in most other normative theories, making the theory particularistic, (iv) an epistemology based on emotions, (v) care is viewed as a moral value, prior to or replacing that of justice, for instance.

When looking at the characteristics of EoC, the crucial issue for evaluating it is to determine whether it is plausible to consider "care" as prior to justice as a moral value. In order to do that, a proper understanding or definition of the term "care" is needed. There are differences within EoC regarding how to understand care. In order to solve this, and come up with a relevant understanding of the term, we need to discuss whether care can be considered a moral value at all. The suggestion is that we separate two notions of care: *natural care* (care in the biological sense), and *ethical care* (care as a morally action-guiding notion). In the moral domain, it is care in the ethical sense – ethical care – that is relevant. The main characteristic of ethical care is that the focus – regarding feelings – always lies with the care receiver, not the care giver. In order to put more justice into EoC, and the ethically relevant

concept of care, the focus should always be on the feelings on the care receiver rather than – as Held suggests – using EoC in certain geographic domains (such as the home), and always use the relation and the emotional setup of the care giver as part of the right-making features of moral actions.

If EoC should be used in healthcare professions, care should be interpreted as ethical care – and as already suggested (the ability for) natural care, which may require emotions, is not necessary when determining whether an action is morally right or wrong. Some patients might prefer human nurses, while some might prefer robot nurses (which might be an argument for focusing on the well being of the care receiver rather than the emotional ability of the care giver.) If the care receiver feels cared for, it does not matter whether the care giver possesses *actual* feelings or actual empathy. We want the patient who suffers the most to get relief first. But there should also be some objective measure, so that a spoiled person (getting elevated stress levels for the tiniest thing) will not get help before a more stoic, patient person. A robot might detect this better than a human since it would not be biased, and also would be able to measure things instantly; measure blood pressure, take blood and measure stress levels, and thereby make an accurate, unbiased assessment of how the patient's status on an objective (actual levels of stress, blood samples, etc.) and subjective level (by reading facial expressions, conversing, etc.). The focus would be a mix of objective and subjective criteria for who gets help first. The conclusion is that EoC may be a theory that would work well for robots in a health care setting.

Paper IV

Johansson, L. (2010). The Functional Morality of Robots. *International Journal of Technoethics* 1(4): 65–73.

In this paper it is suggested that one should use the same criteria for robots as for humans, regarding the ascription of moral agency. When deciding whether humans are moral agents we look at their behaviour and listen to the reasons they give for their judgments in order to determine that they understood the situation properly.

In order to be a moral agent it is necessary to have a certain understanding of what is important to the "moral community". That is, (M1): to be able to discern morally relevant information. This includes the ability to, for instance, discern motives, foresee consequences, and predict the reactions of others, (M2) make moral judgments based on that information, and then (M3) initiate an action based on the judgment. (M1) and (M2) are connected to the possession of internal states, which

are for instance *desires* (to act on what is good), and *beliefs* (about what is good for humans, what is important for humans etc.).

As mentioned above, one of the generally accepted conditions for agency is that the potential agent has mental states. How can we tell that another human being has mental states? And how can we tell that he or she understands? This touches upon *the other minds problem*. The problem of other minds is the problem of how to justify the almost universal belief that other people have minds like our own. This involves two issues: an epistemological one (how can beliefs about mental states other than our own be justified?) and a conceptual one (how is it possible for us to form a concept of mental states other than our own?)

One theory of mind that deals with the other minds problem in a way that suits the common opinion in the moral community is f*unctionalism*. According to functionalism regarding mental states, what makes something a mental state of a particular type is not its internal constitution but rather the way it functions, or the role it plays, in the system of which it is a part. A functionalist method for deciding if another human agent is a moral agent is a so called moral Turing Test (MTT). Analogously to the classic Turing Test, the MTT would be: "If two systems are input-output equivalent, they have the same moral status; in particular, one is a *moral agent* in case the other is". A MTT can be constructed by restricting the standard TT to conversations about morality.

MTT is similar to the way in which we decide whether humans are moral agents or not, in the sense that we check their reasoning and perceptive abilities before deciding if it seems to be the case that they "think" or "understand. Regarding organic issues, one example is to imagine that we – in the future – will be able to transfer a human mind into a robot (a computer). Not the brain per se, since it is organic, but the contents of the brain in terms of all the functions, memories, personality and the sense of self. The idea is that everything would be intact, and identical to the contents of the person's mind, but that this content would be transferred into a machine. There would be nothing organic left. This may seem farfetched and too science fictional, but in principle it is possible; most cognitive scientists and philosophers agree that there is no such thing as an immaterial soul. And the point is: if the mind of a person is intact, but transferred into a machine, we would be biased if we because of that reason alone—the fact that the mind was no longer organic—would not hold the transferred person morally responsible for her actions.

Another example is imagining aliens coming to earth. We can speak to them, also in moral matters. We realize that they seem to have an idea of morality similar to

ours, since they make moral statements such as "that is wrong" when someone is causing pain to a sentient being, for instance. We would most likely accept these aliens as moral agents—based on their passing a MTT. But what if it would, after a while, turn out that these aliens are actually robots? If we decided to consider them moral agents at first, while we believed them to be organic, why would anything change – regarding their moral ability – if we realized that they were *not* organic? There does not seem to be any plausible reason for this. To say that you need to be made out of organic material, and have an organic brain, in order to be a moral agent, is biased. If you pass a MTT, that should be sufficient for moral agency.

Paper V

Johansson, L. (2013). The Pragmatic Robotic Agent. Forthcoming in *Techné* 17(3).

Can artefacts be agents in the same sense as humans? This paper endorses a pragmatic stance to that issue. The question is whether artefacts can have free will in the same pragmatic sense as we consider humans to have a free will when holding them responsible for their actions. The origin of actions is important. Can an action originate inside an artefact, considering that it is always programmed by a human?

In this paper it is argued that autonomy with respect to *norms* is crucial for artificial agency, and that *ascription* of mental states is sufficient for satisfying a pragmatic view on agency. The idea in this paper is to compare robots to humans, which were considered model agents since the stance is pragmatic; we do hold other humans responsible (under certain conditions) despite the fact that we cannot be certain that they have consciousness, and despite the fact that humans may not have free will in the more extreme philosophical sense. Humans decide, on a common sense basis, whether other humans are agents. In order to determine whether robots can be agents in the relevant sense we needed, however, to spell out the (pragmatic) requirements for free will and autonomy, and see if it is possible to incorporate robots into such a view.

A traditional requirement for agency is mental states, something that robots supposedly lack. But the debate surrounding the principle of explanatory exclusion indicates problems surrounding mental causation, and that the *ascription* of mental states is sufficient for a pragmatic view on agency. The deliberative agent in AI has desires and beliefs, and the example of aliens shows that the requirement of mental states being (or stemming from something) organic is biased and lacks support. What remains to be accounted for from the compatibilist account of free will is the

requirement of origin. Where the action originates seems crucial for the agency of robots. Can an action ever originate inside a robot, considering that it is programmed by a human? To answer this we need to compare the programming of a robot, and the "programming" in a human. According to the pragmatic (compatibilist) view, humans have free will, but we are, in some sense, also programmed. We are programmed by evolution, upbringing and genetic codes. One might, for instance, program a robot to follow Kant's categorical imperative. The categorical imperative can be considered a norm. If the robot follows that norm, it might be argued that it would not matter whether it (the categorical imperative) was *programmed* into the robot or, as for humans, it comes via rationality. According to Kant every rational being knows, by being rational, that it is right to follow the categorical imperative. In that case, it can be argued that humans are in some sense "programmed" (maybe by their nature – evolution, that is) to be rational – and to follow the categorical imperative.

A Kantian robot's most important trait would, as Powers (2006) points out, be the deliberative abilities – including an ability to use some sort of practical reasoning or common sense. But with that ability in place, it would not matter how the Kantian morality entered into the robot. The main point according to the pragmatic view on autonomy and agency is that the origin should be within the agent, but the ultimate origin (God, evolution, genes, what have you) can be left aside, when we are being judged by fellow humans or stand trial in a court of law. If we wish to apply the same criterion to autonomy of robots, the ultimate origin should also be disregarded when it comes to robots. It is impossible to actually decide which of the norm-autonomous robot and the human is more programmed in the *relevant* sense in terms of unpredictability. But it is not just unpredictability per se that should be relevant when discussing the extent to which humans are programmed; it is rather the deliberative distance between thought and decision/action.

Paper VI

Carlsen, H., Dreborg, K.H., Godman, M., Hansson, S.O., Johansson, L., Wikman-Svahn, P. (2010) Assessing Socially Disruptive Technological Change. *Technology in Society* 32(3): 209–218.

The co-evolution of society and potentially disruptive technologies makes decision guidance on such technologies difficult. In this paper, we outline an assessment approach for decision guidance on the disruptive social effects of new technologies. Such assessments should be used to inform social decision making, e.g., whether to promote, regulate, or restrict an emerging technology or technological device. Since

our focus is on social effects of new technologies, we need to consider the linkages between society and technologies. This leads us to emphasize the need for co-evolutionary scenarios when assessing the effects of new technologies.
Four basic principles are proposed for such decision guidance. None of the currently available methods satisfies these principles, but some of them contain useful methodological elements that should be integrated in a more satisfactory methodology. Principle 1: An assessment methodology for disruptive technologies should take into account several scenarios of society's development. Principle 2: An assessment methodology for disruptive technologies should explore co-evolutionary paths for society and artefacts based on the technology. To our knowledge a developed methodology of this kind is not readily available; therefore it needs to be developed. Our second principle raises the question of how to select co-evolutionary scenarios out of a vast set of possible scenarios. While it is important to span relevant uncertainties, it is also important to keep the number of scenarios at a manageable level. So far we have focused on developments unaffected by political interventions. But the methodology should be a tool for assessing the effects on societal goals and values of conceivable policy measures. Therefore, the choice of co-evolutionary scenarios should preferably be made in such a way that policy-relevant issues and options are highlighted. In order to achieve this, specific emphasis should be put on constructing scenarios that represent plausible decision nodes, i.e., situations in which a policy response will be required. Principle 3: Co-evolutionary scenarios to be used in the assessment methodology should be relevant from a policy perspective. They should highlight ethically and politically controversial issues and put focus on situations in which a policy response is required. Of course, this selection criterion is value dependent and presupposes the involvement of relevant stakeholders in the assessment process. The same applies to the assessment of the scenarios and the scenario-based deliberations on possible policy responses. In order to develop the scenarios and other background material, expertise on the relevant technologies and social mechanisms will be needed, as well as persons with a competence in scenario development.

In summary, this calls for the inclusion of multiple types of expertise. Principle 4: A process should be designed for the involvement of relevant stakeholder groups and experts on the technology of interest as well as scenario methodology expertise. The outlines of such a methodology, multiple expertise interaction, are proposed. It combines elements from several previous methodologies, including (1) interdisciplinary groups of experts that assess the potential internal development of a particular technology; (2) external scenarios describing how the surrounding world can develop in ways that are relevant for the technology in question; and (3) a

participatory process of convergence seminars, which is tailored to ensure that several alternative future developments are taken seriously into account. In particular, we suggest further development of a bottom-up scenario methodology to capture the co-evolutionary character of socio-technical development paths.

Paper VII

Johansson, L., Carlsen, H., Dreborg, K-H., Wikman-Svahn, P. Co-evolutionary Scenarios for Creative Prototyping of Future Robot Systems for Civil Protection. Forthcoming in *Technological Forecasting and Social Change*.

Abstract: Co-evolutionary scenarios for creative prototyping are used in order to assess the potential implications of future autonomous robot systems on civil protection. Opportunities, threats and ethical aspects in connection with the introduction of robotics in the domestic security and safety sector are identified using an iterative participatory workshop methodology. The first step was to develop prototype artefacts based on the existing or expected user needs and applications that are deemed to be challenging in terms of opportunities, risks and potential ethical dilemmas. The goal was to create prototypes consisting of descriptions of robots that could be to be on the market within a time frame of ten years. The appropriate time frame is dependent on the application, including characteristics of the actual branch of technology (e.g., information and communication technologies vs. heavy infrastructures such as the rail sector).

In the present case study, we arranged a workshop with a mix of participants: robotic experts (robotics researchers from universities and a robot producing firm), potential end-users from the sector of domestic security and safety (from private companies and governmental agencies) and members of the project group. The aim of the workshop was to identify possibilities, vulnerabilities and ethical aspects related to the development and use of autonomous robots within the domestic security and safety sector. Applications that in the long run might have an impact on vital societal functions (cf. section 1 above) within this sector were of special interest. The workshop dealt with the potential robotic systems in a 10 years' time frame, the market demand, ethical problems associated with robot applications and the need for society to react in one way or another.

The ideas for potential artefacts were clustered into eight clusters. In order to prioritize among these clusters, the participants assigned votes to the clusters, based on the participants' judgments on the expected impact for the sector of domestic security and safety. The list consisted of eight ideas for artifacts of autonomous robots: 1) Fire-fighter robots to support human fire-fighters, 2) Safety controller robots, 3) Service and protection robots in shopping malls, 4) Portable robots that

enhance human capabilities, 5) Unmanned aerial vehicle (UAV) robotic systems for detection and pursuit of criminals, 6) Fire detection and fire-fighting robots for use at home, 7) Life rescuing robots for accidents at sea, 8) Fire detection and fire-fighting robots for schools. From this list of eight proposed prototypes, we selected three for further development. The prototypes were subsequently developed more in detail regarding advantages and future potential. The development of the prototypes was informed by a literature review, especially technology road mapping studies. The result of the creative prototyping was a description of the functionality of the prototype robotic systems: "RoboMall", "RoboButler" and "SnakeSquad".

Step two was to construct a hypothetical debate in society concerning security issues and ethical problems – already manifested or expected in the near future – associated with one or the other technological application of the robotic systems. After constructing the initial prototypes we evaluated advantages and potentials (expected developments) identified in the artefact prototypes. The prototypes formed the basis for outlining an ethical debate about the use of the robots. First, different attitudes and views on robots were identified as well as some plausible events triggering the debate. The evaluation was structured along the themes integrity, rights and conflicting interests, economy and security.

In the third step society's response to the experienced ethical problems and opportunities were discussed in the context of different future societies. In order to span a broader range of possible societal developments, we constructed simple "scenarios" of future societies. The idea was to explore how the different artefacts would develop in different societal contexts, depending for instance on how producers, the market and the public respond to the identified ethical and practical problems. Because we were interested in the acceptance or non-acceptance of technologies in society, we wanted to see how two opposite societal stances towards novel technologies would respond to the prototypes. Therefore, when examining the reaction from society, we looked at two different types of societies: 1) the *technology-skeptic society*, and 2) the *technology-positive society*. First, we made a description of these societies. Using these as a starting point, we then imagined different societal reactions, including governmental strategic principles, and the regulations and incentives used to influence the use of the technical artifacts for each scenario. The next step would be to explore how future artefact designers and producers may respond to the regulations and incentives imposed by the different societies in the previous step. For example, new or updated artefacts may appear on the market, while others may disappear or become restricted to a certain application. This can be seen as an iteration of step one, but starting from different prototype artefacts and a different society. The process can then continue another loop, but now with at least two different development paths where new prototype artefacts are developed in different future societies, leading to a tree-like scenario.

7. ERRATA PREVIOUSLY PUBLISHED PAPERS

Paper III

p. 68, left column, line 21: ...are to be designed

p. 68, right column, line 15: This seems to be a good...

p. 69, right column, line 30: All forms of EoC are partial in a way that utilitarianism and ~~consequentialism~~ deontology are not.

p. 71, left column, line 22: The very same empathy that leads us to respond differently to different kinds of situation enters into an understanding of and claims about what is morally better or worse, and it is no wonder...

p. 73, right column, lines 29 and 36: ~~labor~~ --> labour

Paper IV

p. 66, left column, line 35: (which ~~are~~ is a part of the other minds problem).

p. 69, left column, line 19: (Kim ~~2008~~ 2006)

p. 70, left column, line 37: ..they can have no ~~intension~~ intention to act...

8. REFERENCES

Allen, C., Varner, G., Zinser, J. (2000). Prolegomena to any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12(3): 251–261.

Allen, C., Smit, I., Wallach, W. (2005). Artificial Morality: Top-Down, Bottom-Up and Hybrid Approaches. *Ethics and New Information Technology* 7(3): 149–155.

Allen, C., Wallach, W., Smit, I. (2006). Why Machine Ethics? *IEEE Intelligent Systems* 21(4): 12–17.

Arkin, R. (1998). *Behavior-Based Robotics*. Cambridge: MIT Press.

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous systems*. Boca Raton, FL.: CRC Press.

Asaro, P. (2008). How Just Could a Robot War be? In (eds.) Briggle, A., Waelbers, K., Brey, P.A.E. *Current Issues in Computing and Philosophy*. Amsterdam: IOS Press.

Asaro, P. (2009). Modeling the Moral User: Designing Ethical Interfaces for Tele-Operation. *IEEE Technology & Society* 28(1): 20-24.

Asaro, P. (2012). A Body to Kick, but no Soul to Damn. In (eds.) Lin, P., Abney, K., Bekey, G.A. *Robot Ethics – The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Asimov, I. (1942). Runaround. *Astounding Science Fiction*. March, 94-103.

Barribeau, T. (2011). Robots at the University of Tsukuba learn to read facial expressions. *The Verge*. URL= <http://www.theverge.com/2011/11/7/2543952/robot-learns-from-facial-expressions> Accessed May 2013.

Becker, B. (2006). Social robots – emotional agents: Some remarks on naturalizing man-machine interaction. *International Review of Information Ethics* 6(12): 37–45.

Bekey, G.A. (2012). Current Trends in Robotics: Technology and Ethics. In (eds.) Lin, P., Abney, K., Bekey, G.A. *Robot Ethics – The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Berglas, A. (2008). Artificial Intelligence will Kill our Grandchildren. URL= <http://berglas.org/Articles/AIKillGrandchildren/AIKillGrandchildren.html> Accessed May 14, 2013.

Bostrom, N. (2002). Analyzing Human Extinction Scenarios and Related Hazards. Journal of Evolution and Technology 9(1): [first version 2001: URL= <http://www.nickbostrom.com/existential/risks.pdf> Accessed May 2013]

Bostrom, N. (2003a). When Machines Outsmart Humans. *Futures* 35(7): 759–764.

Bostrom, N. (2003b). Taking intelligent machines seriously? Reply to my critics. *Futures* 35(8): 901-906.

Bostrom, N. (2004). The Future of Human Evolution. In (ed.) Tandy, C. *Death and Anti-Death, volume 2: Two Hundred Years after Kant. Fifty Years after Turing.* Palo Alto: Ria UP.

Bringsjord, S. and Taylor, J. (2012). The Divine-Command Approach to Robot Ethics. In (eds.) Lin, P., Abney, K., Bekey, G.A. *Robot Ethics – The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Carlsen, H., Dreborg, K.H., Godman, M., Hanson, S.O., Johansson, L., Wikman-Svahn, P. (2010). Assessing Socially Disruptive Technological Change. *Technology in Society* 32(3): 209–218.

Christensen, Clayton M. (2003). *The Innovator's Solution: Creating and Sustaining Successful Growth*. Boston, MA.: Harvard Business School Press.

Christensen, Clayton M. (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Boston, MA.: Harvard Business School Press.

Chung, M. (2011). Charles the GPS car robot can read your facial expressions. *Übergizmo*. URL= <http://www.ubergizmo.com/2011/01/charles-the-gps-car-robot-can-read-your-facial-expressions/> Accessed May 2013.

Coeckelbergh, M. (2009). Virtual Moral Agency, Virtual Moral Responsibility: On the Significance of the Appearance, Perception and Performance of Artificial Agents. *AI and Society* 24(2): 181–189.

Coeckelbergh, M. (2010). Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations. *Studies in Ethics, Law, and Technology* 4(3), art. 2.

Coeckelbergh, M. (2012). Can We Trust Robots? *Ethics and Information Technology* 14(1): 53–60.

Cole, D. (2009). "The Chinese Room Argument". *The Stanford Encyclopdia of Philosophy.* (Winter 2009 Edition). Edward N. Zalta (ed.). URL= <http://plato.stanford.edu/archives/win2009/entries/chinese-room/> Accessed May 2013.

Coleman, K.G. (2008). "Computing and Moral Responsibility". *The Stanford Encyclopedia of Philosophy*. (Fall 2008 Edition). Edvard N. Zalta (ed.) URL=<http://plato.stanford.edu/archives/fall2008/entries/computing-responsibility/> Accessed May 2013.

Cummings, M. (2006). Integrating ethics through value sensitive design. *Science and Engineering Ethics*, 12(4): 701–715.

Dennett, D. (1973). Mechanism and Responsibility. In T. Honderich (Ed.) *Papers on Freedom of Action*. Boston, MA.: Routledge & Keegan Paul.

Dennett, D. (1976). Conditions of personhood. In A. Rorty (ed). *The Identities of Persons* (175–96). Berkeley, CA.: University of California Press.

Dodig-Crnkovic G. and Çürüklü B. (2011). Robots – Ethical by Design. *Ethics and Information Technology* 14(1): 61–71.

Fellous, J-M. and Arbib, M.A. (2005). *Who Needs Emotions? The Brain Meets the Robot*. New York: Oxford University Press.

Floridi, L. and Sanders, J.W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14(3): 349–379.

Floridi, L. (2008). The Methods of Levels of Abstraction. *Minds and Machines* 18(3): 303–329.

Floridi, L. (2010). Levels of Abstraction and the Turing Test. *Kybernetes* 39(3): 423–440.

Ford, M. (2012). Will Robots and Automation Make Human Workers Obsolete? *Huffington Post*. URL= <http://www.huffingtonpost.com/martin-ford/robots-human-workers_b_1604965.html> Accessed May 2013.

Gallagher, S. (2007). Moral Agency, Self-Consciousness and Practical Wisdom. *Journal of Consciousness Studies* 14(5–6): 199-223.

Gates, B. (2007). A robot in every home. *Scientific American* 296: 58–65.

Grossman, D. (1996). *On Killing: The Psychological Cost of Learning to Kill in War and Society*. New York: Little, Brown and Company.

Hellström, Y. (2013). On the moral responsibility of military robots. *Ethics and information technology* 15(2):99–107.

Himma, K.E. (2009). Artificial agency, consciousness and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11(1): 19–29.

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Bloomington, MN.: Indiana University Press.

Illies, C.F.R. and Meijers, A.W.M. (2009). Artefacts without Agency. *The Monist* 92(3): 420–440.

Jeppsson, S. (2012). *Practical Perspective Compatibilism*. Stockholm Studies in Philosophy 35. Stockholm: ACTA Universitatiss Stockholmiensis.

Johnson, D.G. (2006). Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology* 8(4): 195–204.

Johnson, D.G. and Powers, T.M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology* 79(2): 99–107.

Jonas, H. (1984). *The Imperative of Responsibility. In Search of an Ethics for the Technological Age.* Chicago: The Chicago University Press.

Kim, J. (2006). *Philosophy of Mind*. Boulder, CO.: Westview Press.

Kurzweil, R. (2005). *The Singularity Is Near*. New York: Viking.

Levin, J. (2010). "Functionalism". *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/functionalism/> Accessed May 2013.

Lin, P., Bekey, G.A., Abney, K. (2008). Autonomous military robotics: Risk, ethics and design, a US department of defense office of naval research-funded report. <http://ethics.calpoly.edu/ONR_report.pdf> Accessed May 2013.

Lin, P., Abney, K., Bekey, G.A. (2012). *Robot Ethics – The Ethical and Social Implications of Robotics*. Cambridge, MA.: MIT Press.

Lintem, Gavan (2007). What is a Cognitive system? *Procedures of the Fourteenth International Symposion on Aviation Psychology* (pp. 398–402) Dayton, OH. <http://www.cognitivesystemsdesign.net/Papers/What%20is%20a%20Cognitive%20System.pdf> Accessed April 2013.

Matthias, A. (2004). The Responsibility Gap. Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6(3): 175–183.

McLaren, B. (2006). Computational Models of Ethical Reasoning: Challenges, Initial Steps and Future Directions. *IEEE Intelligent Systems* 21(4): 29-37.

Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* New York: Simon & Schuster Inc.

Moor, J. (2006). The Nature, Importance and Difficulty of Machine Ethics. IEEE *Intelligent Systems* 21(4): 18–21.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM* 37(1): 73–80.

Noorman, M. (2012). "Computing and Moral Responsibility". *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/computing-responsibility/> Accessed May 2013.

O'Connor, T. (2008). "FreeWill". *The Stanford Encyclopedia of Philosophy* (Fall 2008 edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/freewill/> Accessed May 2013.

Orend, B. (2008). "War". The Stanford Encyclopedia of Philosphy (Fall 2008 Edition), Edward N. Zalta (ed.), URL= <http://plato.stanford.edu/archives/fall2008/entries/war/>. Accessed May 2013.

Peterson, M. & Spahn, A. (2011). Can Technological Artefacts Be Moral Agents? *Science and Engineering Ethics* 17(3): 411–424.

Pinker, S. (2011). Tech luminaries address Singularity. *IEEE Spectrum*. URL= <http://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity> Accessed May 2013.

Powers, T.M. (2006). Prospects for a Kantian Machine. *IEEE Computer Society* 21(4): 46–51.

Quintana, E. (2008). The Ethics and Legal Implications of Military Unmanned Vehicles. RUSI Occasional Paper. URL= < http://www.rusi.org/downloads/assets/RUSI_ethics.pdf > Accessed May 2013.

Roberts, A. and Guelef, R. (2010). *Documents on the Laws of War*. Oxford: Oxford University Press.

Rothblatt, M. and Angelica, A.D. (2003). Biocyberethics: Should we stop a company from unplugging an intelligent computer? URL= <http://www.kurweilai.net/biocybernetics-should-we-stop-a-company-from-unplugging-an-intelligent-computer> Accessed May 2013.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3(3): 417–457.

Sharkey, N. and Sharkey, A. (2010). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*. doi:10.1007/s10676-010-9234-6 (2012, 14: 27-40).

Silver, D.A. (2005). Strawsonian defense of corporate moral responsibility. *American Philosophical Quarterly* 42(4): 279–295.

Singer, P.W. (2009). *Wired for War – The Robotics Revolution and 21st Century Conflict*. New York: Penguin Press.

Sommerville, I. (2007). Models for responsibility assignment. In (Eds.) Dewsbury, G. and Dobson, J.: *Responsibility and dependable systems*. Kluwer: Springer.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy* 24(1): 62–77.

Sparrow, R. (2009). Building a Better Warbot. Ethical Issues in the Design of Unmanned Systems for Military Applications. *Science and Engineering Ethics* 15(2): 169–187.

Sparrow, R. and Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines* 16(2): 141–161.

Stahl, B.C. (2004) Information, Ethics and Computers: The Problem of Autonomous Moral Agents. *Minds and Machines* 14(1): 67–83.

Strawson, G. (1998, 2011). Free Will. In (Ed.) Craig, E.*: Routledge Encyclopedia of Philosophy*. London: Routledge. URL= <http://www.rep.routledge.com/article/vol14SECT1> Accessed May 2013.

Sullins, J.P. (2006). When is a robot a moral agent? *International Review of Information* 6(12): 23–30.

Torrance, S. (2008). Ethics and Consciousness in Artificial Agents. *AI & Society - Special Issue: Ethics and Artificial Agents* 22(4): 495–521.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* 59(236): 433–460.

Vallverdú, J. and Cascuberta, D. (2009). *Handbook of Research on Synthetic Emotions and Sociable Robotics: New applications in Affective Computing and Artificial Intelligence.* Hershey, PA.: IGI Global.

Vallor, S. (2011). Carebots and caregivers: Sustaining the ethical ideal of care in the 21st century. *Journal of Philosophy and Technology* 24(3): 251–268.

Van den Hoven, J. (2007). ICT and value sensitive design. *International Federation for Information Processing* 233: 67–72.

Van Inwagen (1983). *An Paper on Free Will*. Oxford: Clarendon Press.

Van Wynsberghe, A. (2012). Designing Robots for Care: Care-Centered Value Sensitive Design. *Science and Engineering Ethics*. DOI 10.1007/s11948-011-9343-6 (June 2013, Volume 19(2): 407–433.)

Verbeek, P.-P. (2005). *What Things Do.* Pennsylvania: Penn State University Press.

Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology & Human Values* 31(3): 361–380.

Verhagen, H.JE. (2000). *Norm Autonomous Agents*. PhD thesis, Department of Computer and Systems Sciences. Stockholm University/Royal Institute of Technology.

Verhagen, H.J.E. (2004). Autonomy and Reasoning for Natural and Artificial Agents. In (eds.) Nickles, M., Rovatsos, M., Weiss, G.: *Agents and Computational Autonomy.* Lecture notes in Computer Science, Volume 2969/2004, pp. 83–94. Berlin: Springer.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Waelbers, K. (2009). Technological Delegation: Responsibility for the Unintended. *Science and Engineering Ethics* 15(1): 51–68.

Warwick, K. (2009). Today it's a cute friend. Tomorrow it could be the dominant life form. *Times of London* 2009, URL= <www.noonehastodietomorrow.com/tech/transhumanism/865-865> Accessed May 2013.

Wrenn, E. (2012). Almost human: The robot called FACE who can display dozens of life-like emotions (and is based on one of the researcher's wives). *Daily Mail* 13 July 2012. URL= <http://www.dailymail.co.uk/sciencetech/article-2172990/Almost-human-The-robot-called-FACE-display-dozens-life-like-emotions-based-researchers-wife.html> Accessed May 2013.

Yudkowski, E. (2004). Coherent Extrapolated Volition. *Machine Intelligence Research Institute*. URL= <http://intelligence.org.files/CEV.pdf> Accessed May 2013.

OTHER SOURCES

HRW 2012. (Human Rights Watch Report). "Losing Humanity: The case against killer robots" http://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf . Accessed May 2013.

Robotnyheter 2012: (in Swedish) URL= <http://www.robotnyheter.se/2012/10/24/sjalvkorande-bilar-fran-volvo-redan-2014> Accessed May 2013.

US Dept of Defense 2011. "Unmanned systems Integrated Roadmap FY 2011-2036; "Reference Number 11-S-3613, 2011, URL= <http://www.defenseinnovationmarketplace.mil/resources/UnmannedSystemsIntegratedRoadmapFY2011.pdf > Accessed May 2013.

9. SAMMANFATTNING PÅ SVENSKA (SUMMARY IN SWEDISH)

Denna doktorsavhandling startade i ett samarbetsprojekt mellan KTH och FOI: CopeTech. Syftet med detta projekt var att utveckla en metod för att hantera så kallade disruptiva teknologier med hjälp av co-evolutionära element. En workshop med efterföljande analys indikerade att tvärvetenskaplighet och deltagande av olika intressegrupper är essentiellt för att identifiera och värdera framtida risker och hot gällande disruptiva teknologier. Grunden till workshopen lades i artikel VI i denna avhandling "Assessing socially disruptive technological change". Inför workshopen valde projektet att fokusera på autonoma system, en potentiellt disruptiv teknologi. Resultatet från workshopen utvecklades och presenterades i text VII: "A co-evolutionary creative prototyping approach for the future of autonomous robot systems and potential implications on the civil protection sector".

Det finns främst tre kategorier av filosofiska frågor gällande autonoma system. En kategori handlar om etiska frågor rörande själva användandet. I denna avhandling ligger huvudfokus gällande denna kategori på robotar i krig. I artikel I, "Is it morally right to use UAVs in war?" diskuteras huruvida det är etiskt att använda obemannade flygande farkoster – UAVer – i krig med krigets lagar – fastlagda ibland annat Genève och Haag-konventionerna – som utgångspunkt. Exempel på etiska frågor rörande *jus ad bellum* (regler för att starta krig) är att UAVer kan göra krig mer riskfritt och därmed sänka tröskeln, att det blir enklare att föra "hemliga" krig samt att användandet kan skapa en stark känsla av orättvisa hos den part som inte har UAVer. Etiska frågor rörande *jus in bello* (regler inom kriget) handlar om att de krigförande operatörerna, som i vissa fall sitter på andra sidan jorden, blir känslomässigt avdomnade eftersom de befinner sig så långt från krigszonen. En annan fråga handlar om att krigets lagar är väldigt öppna för tolkning, vilket gör att ett land med UAVer kan tolka lagarna till egen fördel. Exempel på termer som kan tolkas olika beroende på vilken etisk teori man utgår från är "rättvis" eller "giltig" (*just* cause), "nödvändig" (necessary) och "överdriven" (excessive). I artikeln undersöks hur utilitarism, deontologi och dygdetik kan ge skilda svar. Slutsatsen är att man bör se över krigets lagar med hänsyn till användandet av UAVer.

I text II, "Autonomous robots in war: undermining the ethical justification for killing?", diskuteras huruvida användandet av robotar i krig underminerar det etiska rättfärdigandet att döda, som är implicit i krigets lagar. I texten argumenteras för att ju mer autonoma robotarna är, desto mer undermineras rättfärdigandet att döda. Anledningen till detta är att om robotar används för att ersätta människor snarare

än att assistera, försvinner det ömsesidiga utsättandet för risk ”the reciprocal imposition of risk” – som är ett nödvändigt villkor för den jämlikhet mellan kombattanter som är implicit i krigets lagar. Detta indikerar att man behöver se över krigets lagar gällande robotar som är autonoma i hög grad.

En andra kategori av frågor, som är unik för autonoma system på grund av deras autonomi, är hur man försäkrar sig om att autonoma robotar beter sig på ett etiskt tillfredsställande sätt. Olika förslag finns i litteraturen. Ett alternativ är att programmera in en moralteori ”top-down”, ett annat är att låta roboten lära sig vad som är moraliskt önskvärt, ”bottom-up”, eller att blanda dessa metoder. I artikel III diskuteras en ”top-down” metod för detta; huruvida den etiska teorin ”ethics of care” skulle kunna vara en lämplig teori för robotar i vården. Förslaget är att man bör skilja på ”care” i en vardaglig mening och i en etiskt relevant mening. Olika varianter av ”ethics of care” ser på begreppet ”care” på olika sätt, och många menar att den som ägnar sig åt ”care” måste ha känslor. I denna artikel argumenteras för att det etiskt relevanta ”care” är hur den som får ”care” uppfattar denna, snarare än huruvida den som ger ”care” verkligen har mänskliga känslor och att denna teori skulle kunna passa för robotar i vårdmiljö.

En tredje kategori av frågor handlar om agentskap och ansvar. Sådana frågor är ganska kontroversiella och ibland spekulativa eftersom man diskuterar huruvida en robot kan vara moraliskt ansvarig för sina handlingar – huruvida den kan vara en moralisk agent i samma mening som en människa är det. Agentskap och ansvar diskuteras i två texter i denna avhandling. I artikel IV, “The Functional Morality of Robots”, argumenteras för att en robot skulle kunna hållas moraliskt ansvarig om den kan klara ett moraliskt Turingtest. I text V, ”The Pragmatic Robotic Agent”, undersöks huruvida robotar kan ha den frihet att agera som krävs för att man ska hållas moraliskt ansvarig för sina handlingar. Denna avhandling har en pragmatisk utgångspunkt; människor hålls ansvariga för sina handlingar – moraliskt och i juridisk mening – om vissa kriterier är uppfyllda, och tanken är att robotar bör bedömas enligt samma standard. En viktig fråga gäller friheten att agera; en robot är ju programmerad. Det kan emellertid hävdas att även människor är programmerade i någon mån. Den kritiska punkten tycks handla om handlingens ursprung, och förslaget är att man ska skilja på ”ultimate origin” och ursprung som har tillräckligt avstånd mellan tanke och handling. Om en robot kan trotsa normer, ska den kunna anses vara en agent.