Ethics and Authority Sharing for Autonomous Armed Robots

Florian Gros[1] and **Catherine Tessier**[1] and **Thierry Pichevin**[2]

Abstract. The goal of this paper is to review several ethical questions that are relevant to the use of autonomous armed robots and to authority sharing between such robots and the human operator. First, we discern the commonly confused meanings of morality and ethics. We continue by proposing leads to answer some of the most common ethical questions raised by literature, namely the autonomy, responsibility and moral status of autonomous robots, as well as their ability to reason ethically. We then present the possible advantages that authority sharing with the operator could provide with respect to these questions.

1 INTRODUCTION

There are many questions and controversies commonly raised by the use of increasingly autonomous robots, especially in military contexts [51]. In this domain autonomy is can be explored because of the need for reducing the atrocities of war, e.g. loss of human lives, violation of human rights, and for increasing battle performance to avoid unnecessary violence [3]. Since full autonomy is far from achieved, robots are usually supervised by human operators. This coupling between a human and a robotic agent involves a shared authority on the robot's resources [30], allowing for adaptability of the system in complex and dynamic battle contexts. Even with humans in the process, the deployment of autonomous armed robots raises ethical questions such as the responsibility of robots using lethal force incorrectly [47], the extent of their autonomous abilities and the related dangers, their ability to comply with a set of moral rules and to reason ethically [44], and the status of robots with regard to law due to the ever-increasing autonomy and human resemblance that robots display [28].

In this paper we will highlight the distinction between morality and ethics (section 2). Then several ethical issues raised by the deployment of autonomous armed robots, such as autonomy, responsibility, consciousness and moral status will be discussed (section 3). As another kind of ethical questions, a review of the frameworks used to implement *ethical* reasoning into autonomous armed robots will be presented afterwards (section 4). Finally, we will consider the ethical issues and implementations mentioned earlier in the framework of authority sharing between a robot and a human operator (section 5).

2 MORALITY AND ETHICS

The concepts of *morality* and *ethics* are often used in an identical fashion. If we want to talk about ethics for autonomous robots, we have to distinguish those terms and define them.

2.1 Morality

If we ignore meta-ethical debates that aim at defining morality and its theoretical grounds precisely, we can conceive morality as principles of good or bad behaviour, an evaluation of an action in terms of right and wrong [52]. This evaluation can be considered either absolute or coming from a particular conception of life, a typical moral rule being "Killing is wrong". It is important to note that in this work, we focus on moral *action*, whether it results from rules, or from intentions of the subject doing the action.

2.2 Deontology and teleology

One of the bases for morality is the human constant need to believe in a meaning of one's actions. In most philosophical debates, this sense pertains to two often opposed categories : teleology and deontology.

For teleology, the moral action has to be good, the goal being to maximize the good and to minimize the evil produced by the action [33]. In this case, morality is commonly viewed as external to the agent, because it comes within the scope of a finalized world defining the rules and the possible actions and their goals, therefore defining the evaluation of actions.

For deontology, the moral action is done by duty, and must comply with rules regardless of the consequences of the action, whether they are foreseen or not, good or bad [34]. A case by case evaluation is not necessarily relevant here, because it is the humans' responsibility to dictate the rational and universal principles they want to live by.

2.3 Ethics

Ethics appears as soon as a conflict between existing legal or moral rules emerges, or when there is no rule to guide one's actions [36]. For example, if a soldier has received an order not to hurt any civilian, but to neutralize any armed person, what should he do if he encounters an armed civilian? We can thus consider ethics as the commitment to resolving moral controversies [13] where the agent, with good will, has to solve the conflicts he is faced with.

Those conflicts often oppose deontological and teleological principles, namely what has to be privileged between right and good ? The goal of ethics is not to pick one side and stand by it forever, but to be able to keep a balance between right and good when solving complex problems. Solving an ethical conflict then requires, apart from weighing good and evil, a sense of creativity in front of a complex situation and to be able to provide alternative solutions to moral rules imperatives [31].

[1] Onera, the French Aerospace Lab, Toulouse, France, email: name.surname@onera.fr

[2] CREC, Ecoles de Saint-Cyr Coetquidan, France, email: thierry.pichevin@st-cyr.terre-net.defense.gouv.fr

To provide an illustration of the distinction between morality and ethics, we will consider that any moral conflict needs ethical reasoning abilites to be solved. Speaking of ethical rules would not make sense since ethics apply when rules are absent or in conflict.

3 AUTONOMY, RESPONSIBILITY, MORAL STATUS : PROSPECTS FOR ROBOTS

Technologies leave us presently in an intermediate position where robots can perceive their environment, act and make decisions by themselves, but lack a more complete kind of autonomy or the technological skill to be able to analyze their environment precisely and understand what happens in a given situation. Still research advances urge us to think about how to consider autonomous robots in a moral, legal and intellectual frame, both for the time being and when robots are actually skilled enough to be considered similar to humans. In this section, we will review important questions for autonomous robots i.e. autonomy, responsibility, moral status and see which answers are plausible. Then we will relate these questions to authority sharing.

3.1 Autonomy

3.1.1 Kant and the autonomy of will

When considering autonomy, one of the most influential view in occidental culture is Kant's. For him, human beings bend reality to themselves with their perception and reason, they escape natural or divine laws. Only reason enables humans to create laws that will determine humankind. Then laws cannot depend on external circumstances as reason only can provide indications in order to determine what is right or wrong. Consequently laws have to be created by a *good will*, i.e. a will imposing rules on itself not to satisfy an interest, but by duty towards other humans. Therefore no purpose can be external to humankind, and laws are meaningful to humans only if they are universal. This leads to the well-known moral ”categorical” imperative[3], that immediatly determines what it orders because it enounces only the idea of an universal law and the necessity for the will to follow it [39].

Humans being the authors of the law they obey, it is possible to consider them as an end, and the will as autonomous. Thus, to be universal, a law has to respect humans as ends in themselves, inducing a change in the categorical imperative. If the law was external to humans, they would not be ends in themselves, but mere instruments used by another entity. Such a statement would deny the human ability to escape divine or natural laws, which is not acceptable for the kantian theory. We can only conceive law as completely universal, respecting humans as ends in themselves. To sum up, the kantian autonomy is the ability for an agent to define his own laws as ways to fulfill his goals and to govern his own actions.

3.1.2 Autonomy and robots

In the case of an Unmanned System, autonomy usually stands for decisional autonomy. It can be defined as the ability for an agent to minimize the need for supervision and to evolve alone in its environment [43], or more precisely, its ”own ability of sensing, perceiving, analyzing, communicating, planning, decision making, and acting/executing, to achieve its goals as assigned by its human operators” [21].

We can see a difference between those definitions and Kant's. Robot autonomy is perceived differently for robots than for humans, as an autonomy of means, not of end. The reason for this is that robots are not sophisticated enough to be able to define their own goals and to achieve them. Robots are therefore viewed as mere tools whose autonomy is only intended to alleviate the operators' workload.

Consequently, to be envisioned as really autonomous, robots should be able to determine their own goals once deployed, thus to have will and be ends in themselves. The real question to ask here is if it is really desirable to build such fully autonomous robots, especially if they are to be used on a battlefield. If the objective is solely to display better performance than human soldiers, full autonomy is probably inappropriate, since being able to control robots and their goals from the beginning to the end of their deployment is one of the main reasons for actually using them.

3.2 Responsibility

If we want to use autonomous robots, we have to know to what extent a subject is considered responsible for his actions. It is especially important when applied to armed robots, since they can be involved in accidents where lives are at stake.

3.2.1 Philosophical approaches to responsibility

Classically responsibility has been considered from a broad variety of angles, whether being a relationship to every other human being in order to achieve a goal of salvation given by a divine entity (Augustine of Hippo), a logic consequence of the application of the categorical imperative (Kant), a duty towards the whole humanity as the only way to give a sense, a determination to one's actions and to define oneself in the common human condition (Sartre, [42]), or an obligation to maintain human life on Earth as long as possible by one's actions (Jonas, [22]).

The problem with those approaches is that they are thought for humans and consequently they require, more or less, an autonomy of end. As discussed above, this is not a direct possibility for robots. We then need to envision robot responsibility in their own ”area” of autonomy, namely an autonomy of means, where the actions are not performed by humans. To discuss this problem, it is necessary to distinguish two types of responsibility : causal responsibility and moral responsibility.

3.2.2 Causal responsibility vs. moral responsibility

By moral responsibility, we mean the ability, for a conscious and willing agent, to make a decision without referring to a higher authority, to give the purposes of his actions, and to be judged by these purposes. To sum up, the agent has to possess a high-level intentionality [12]. This moral responsibility is not to be confused with causal responsibility, which establishes the share of a subject (or an object) in a causal chain of events. The former is the responsibility of a soldier who willingly shot an innocent person, the latter is the responsibility of a malfunctioning toaster that started a fire in a house.

Every robot has some kind of causal responsibility. Still, trying to determine the causal responsibility of a robot (or of any agent) for a given event is way too complex because it requires to analyze every action the robot did that could have led to this event. What we are really interested in is to define what would endow robots with a *moral* responsibility for their actions.

[3] ”act only according to that maxim by which you can at the same time will that it be a universal law”

3.2.3 *Reduced responsibility, a solution ?*

Some approaches that are currently considered for the responsibility of autonomous robots are based on their status of ”tools”, not of autonomous agents. Thus, their share of responsibility is reduced or transferred to another agent.

The first approach is to consider robots as any product manufactured and designed by an industry. In case of a failure, the responsibility of the industry (as a moral person) is substituted to the responsibility of the robot. The relevant legal term here is *negligence* [24]. It implies that manufacturers and designers have failed to do what was legally or morally required, thus can be held accountable of the damage caused by their product. The downside of this approach is that it can lean towards a causal responsibility which – as said earlier – is more difficult to assess than a moral responsibility. Besides, developing a robot that is sure *enough* to be used on a battlefield would demand too much time for it to represent a good business, and it wouldn’t even be enough to be safely used, a margin of error still existing no matter how sophisticated a robot is.

Another approach then would be to apply the *slave morality* to autonomous robots [24] [28]. A slave, by itself, is not considered responsible for his actions, but his master is. At a legal level, it is considered as *vicarious liability*, illustrated by the well-known maxim *Qui facit per alium facit per se*[4]. If we want to apply this to autonomous armed robots, their responsibility would be substituted to their nearest master, namely the closest person in the chain of command who decided and authorized the deployment of the robots. This way, a precise person takes responsibility for the robots actions, which spares investigations through the chain of command to assess causal responsibilities.

Finally, if we consider an autonomous robot to be able to comply with some moral rules, to reason as well as to act, it is possible to envision the robot as possessing, not moral responsibility, but moral intelligence [5]. The robotic agent is then considered to be able to adhere to an ethical system. Therefore there is a particular morality within the robot that is specific to the task it is designed for.

3.2.4 *Other leads for a moral responsibility*

No robot has been meeting the necessary requirements for moral responsibility, and no law has been specifically written for robots. The question is then to determine what is necessary for robots to achieve moral responsibility and what to do when they break laws.

For [19] and [1], the key to moral responsibility is the access to a moral status. Besides an emotional system, this requires the ability of rational deliberation, allowing oneself to *know* what one is doing, to be conscious of one’s actions in addition to make decisions. Severals leads for robots to access to a moral status are detailed in the next section.

As far as responsibility is concerned, a commonly used argument is that robots cannot achieve moral responsibility because they cannot suffer, and therefore cannot be punished [47]. Still, if we consider punishment for what it is, i.e. a convenient way to change (or to compensate for) a behaviour deemed undesirable or unlawful, we can agree that it is not the *sine qua non* requirement for responsibility. There are other ways to change one’s behaviour, one of the most known examples being treatment, i.e. spotting the ”component” that produces the unwanted behaviour and tweak it or replace it to correct the problem [28]. Beating one’s own car because of a malfunction would be absurd, in this case it is more fitting to replace the malfunctioning component. The same applies with certain types of law infringement (leading to psychological treatment or therapy), so it could apply to robots as well, e.g. by changing the program of the defective vehicle. Waiting for technology to progress to finally being able to punish robots so that they could have moral responsibility is not a desirable solution, but using vicarious liability, treatment and moral status appears to be a sound basis.

3.3 Consciousness and moral status for autonomous robots

We have said earlier that for a robot to be considered responsible for its actions, it must be attributed a moral status, so it needs consciousness [19]. The purpose of this section is to see how this can be achieved and how moral status can be applicable to robots in order to help them to have moral responsibility.

3.3.1 *Consciousness*

Since there is an abundant literature on the topic of consciousness, and still no real consensus among the scientific community on how define consciousness, the purpose of this section is not to give an exhaustive nor accurate definition of consciousness, but merely to see what seems relevant to robots. However, if we want to use consciousness, we can consider it as described by [32], namely the ability to know *what it is like* to have such or such mental state from one’s own perspective, to subjectively experience one’s own environment and internal states.

The first approach for robots consciousness is the theory of mind [38] [6]. It is based on the assumption that humans tend to grant intentionnality to any being displaying enough similarities of action with them (emotions ou functional use of language). It is then possible for humans, by analogy with their experience of their own consciousness, to assume that those beings have a consciousness as well. This approach is already developing with conversational agents or robots mimicking emotions, even if it can be viewed as a trick of human reasoning more than an ”absolutely true” model of consciousness.

The second approach considers consciousness as a purely biological phenomenon, and has gained influence with the numerous discoveries of neurosciences. Even if we do not know what really explains consciousness (see the Hard problem of consciousness [9]), considering it as a property of the brain may allow conscious robots to be developed, as did [55] [54] by recreating a brain from collected brain cells. There is still a lot of work to do here, as well as many ethical questions to answer, but it definitely looks promising. Indeed, if a being, even with a robotic body, has a brain that is similar to a human’s, in a materialist perspective, this being is conscious.

The last approach is the one proposed by [25] [26] to build self-aware robots that can explore their own physical capacities to find their own model and to determine their own way to move accordingly. Those robots are probably the closest ones to consciousness as defined by [32]. They are still far from being used on a battlefield, but this method of self-modelling could be applied to more ”evolved” robots for ethical decision-making. This way a robot could explore its own capacities for action and could build an ethical model of itself.

3.3.2 *Moral status*

An individual is granted moral status if it has to be treated never as a means, but only as an end, as prescribed by Kant’s categorical imperative. To define this moral status, two criteria are commonly

[4] ”He who acts through another does the act himself.”

used [7], namely sentience (or *qualia*, the ability to experience reality as a subject) and sapience (a set of abilities associated with high-level intelligence). Still, none of those attributes have been successfully implemented in robots. Even though it could be counter-productive to integrate *qualia* to robots in some situations (e.g. coding fear into an armed robot), it can be interesting to model some of them into robots, like [4] did for moral emotions like guilt. This could provide a solid ground for access of robots to moral status. [7] have proposed two principles stating that two different agents can have the same moral status if they possess enough similarities : if two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation (Principle of Substrate Non-Discrimination) or on how they came to existence (Principle of Ontogeny Non-Discrimination), then they have the same moral status.

Put simply, those principles are pretty similar to what the theory of mind proposes, that is if robots can exhibit the same functions as human's, then they can be considered as having a moral status, no matter what their body is made of (silicon, flesh, etc.) or how they matured (through gestation or coding). Still, proving that robots can have the same conscious experience as humans is currently impossible, so we can consider a more applicable version of those principles: [49] proposes that robots have moral agency if they are responsible with respect to another moral agent, if they possess a relative level of autonomy and if they can show intentional behaviour. This definition is vague but is grounded on the fact that moral status is attributed. What matters is that the robot is advanced enough to be similar to humans, but it does not have to be identical.

Another solution for autonomous robots with a moral status is to create a sort of Turing Test comparing the respective ”value” of a human life with the existence of a robot. This is called by [46] the Triage Turing Test and shows that robots will have the same moral status as humans when it is at least as wrong to ”kill” a robot as to kill a human. Advanced reflections on this topic can be found in [48].

4 IMPLEMENTING ETHICAL REASONING INTO AUTONOMOUS ARMED ROBOTS

Another question related to autonomous armed robots is how those robots can solve ethical problems on the battlefield and make the most ethically satisfying decision. In this section, we will briefly review several frameworks to integrate ethical reasoning into robots.

Three kinds of approaches are considered:

- Top-down : these approaches take a particular ethical theory and create algorithms for the robot, allowing it to follow the aforesaid theory. This is convenient to implement, e.g. a deontological morality into a robot.
- Bottom-up : the goal is to create an environment wherein the robot can explore different courses of action, with rewards to make it lean towards morally satisfying actions. Those approaches focus on the autonomous robot learning its own ethical reasoning abilities.
- Hybrid : these approaches look for a merge between top-down and bottom-up frameworks, combining their advantages without their downsides.

4.1 Top-down approaches

Top-down frameworks are the most studied in the field of ethics for robots and the number of ethical theories involved is high. Literature identifies theories such as utilitarianism [10], divine-command ethics [8] and other logic-based frameworks [27] [15]. Still, the most famous theory among top-down approaches is the Just-War Theory [35], which underlies the instructions and principles issued in the Laws of War and the Rules of Engagement (for more on these documents, see [3]). Those approaches have in common to take a set of rules and to program them into the robot code so that their behaviour could not violate them. The upside of those approaches is that the rules are general, well-defined and easily understandable. The downside is that no set of rules will ever handle every possible situation, mostly because they do not take into account the context of the particular mission the robot is deployed for. Thus top-down approaches are usually too rigid and not precise enough to be applicable. Also, since they rely on specific rules – more morality-like than ethics-like – they are not fit to capture ethical reasoning abilities but they are usually used to justify one's own actions. In order to implement ethical reasoning abilities into robots, it seems more desirable to use top-down approaches as moral heuristics guiding ethical reasoning [53].

4.2 Bottom-up approaches

Bottom-up frameworks are way less developed than top-down approaches. Still, some research like [26] gives interesting options, using self-modeling. Most of the bottom-up approaches insist on machine learning [17] or artificial evolution using genetic algorithms based on cooperation [45] to allow agents to reason ethically given a specific parameter. The strength of these frameworks is that learning allows flexibility and adaptability in complex and dynamic environments, which is a real advantage in the field of ethics wherein there is no predefined answers. Nevertheless the learning process takes a lot of time and never completely removes the risk of unwanted behaviour. Plus, the reasoning behind the action produced by the robot cannot be traced, making the fix of undesirable behaviours barely possible.

4.3 Hybrid approaches

Three different frameworks can be distinguished among hybrid approaches : case-based approach [29] [2], virtue ethics [24] [53] and the hybrid reactive/deliberative architecture proposed by [3], using the Laws of War and the Rules of Engagement as a set of rules to follow. They are probably the most applicable researches to autonomous robots and combine aspects of both top-down (producing algorithms derived from ethical theories) and bottom-up (using agents able to learn, evolve and explore possible ethical decisions) specifications. The main problem with these approaches is their computing time, since learning is often involved in the process. Nevertheless, they appear theoretically satisfying and their applicability looks promising.

5 ETHICS AND AUTHORITY SHARING

In this section we will focus on the previously mentioned ethical issues in the framework of authority sharing between a robot and a human operator.

Joining human and machine abilities aims at increasing the range of actions of “autonomous” systems [23]. However the relationship between both agents is dissymmetric since the human operator's “failures” are often neglected when designing the system. Moreover simultaneous decisions and actions of the artificial and the human agents are likely to create conflicts [11]: unexpected or misunderstood authority changes may lead to inefficient, dangerous or catastrophic situations. Therefore in order to consider the human agent and the artificial agent in the same way [20] and the human-machine system as a whole [56], it seems more relevant to work on authority and authority control [30] than on autonomy, which concerns the artificial agent exclusively.

Therefore authority sharing between a robot and its operator can be viewed as an “upgraded” autonomy. As far as ethical issues are concerned, authority sharing considered as a relation between two agents [18] may provide a better compliance with sets of laws and moral rules, this way enabling ethical decision-making within a pair of agents instead of leaving this ability to only one individual.

5.1 Autonomy

As previously mentioned, the autonomy of an armed robot can be conceived as an autonomy of means only; robots are almost always used as tools. Authority sharing can bring a change in this organization. As a robot cannot (yet) determine its own goals, it is the human operator’s role to provide the goals so as some methods or partial plans to achieve them [14]. Still, authority sharing allows the robot to be granted decision-making power allowing it to take authority from the operator to accomplish some tasks neglected by him (e.g., going back to base because of a fuel shortage) or even when the operator’s actions are not following the mission plan and may be dangerous. For example, some undesirable psychological and physiological ”states” of the operator, e.g. tiredness, stress, attentional blindness [37] can be detected by the robot, in order to allow it to take authority if the operator is not considered able to fulfill the mission anymore.

5.2 Moral responsibility

Concerning moral responsibility, authority sharing forces us to make a distinction between two instances : the one where the operator has authority over the robot, and the reverse one. The former is simple; since the robot is a tool, we use the vicarious liability, therefore the operator engages his responsibility for any accident caused by the use of the robot that could happen during the mission. The latter is more complex and we do not claim to give absolute answers, but mere propositions.

What we propose is that, in order to assess moral responsibility when the robotic agent has authority over the system, it is necessary to define a mission-relevant set of rules, e.g. Laws of War and Rules of Engagement [35] [3], and a contract, as proposed by [41] or [40], between robotic and human agents, providing specific clauses for them to respect during the mission. These clauses must to be based on the set of rules previously mentioned, and an agent who violates them would be morally responsible of any accident that could happen as a consequence of his actions.

This kind of contract would provide clear conditions for authority sharing (i.e., an agent loses authority if he violates the contract) and could open the way to apply works on trust [4] or persuasion [16] in robotic agents. During a mission, such contracts would engage both agents to monitor the actions of the other agent and, if possible, to take authority if this can prevent any infringement of the contract. If one agent detects a possibly incoming accident due to the other agent’s actions, e.g. aiming at a civilian, and does nothing to prevent it, then this agent is responsible for this accident as much as the one causing it. Because of the current state of law, i.e. dealing only with human behaviours, if a robot is considered responsible for ”evil” or unlawful actions, then it should be treated by replacing the parts of its program or the pieces of hardware that caused the unwanted behaviour. Human operators, if displaying the same kind of unlawful behaviour, should be judged by the appropriate laws. To integrate contracts in a concrete way, we can lean towards the perspective presented by [3] who proposes some recommendations to warn the operator of his responsibility when using potentially lethal force.

5.3 Consciousness and moral status

Authority sharing is not of a great help to implement consciousness into robots. Still, [37] and [50] provide leads to allow robots to assess the ”state” of the operator and to take authority from him if he is not considered able to achieve the mission. This approach would help robots to improve their situational awareness and to design systems that are better at interacting with humans, either operator or civilians. Enhancing the responsibility and autonomy of robots could also be a way to push them towards the ”same functionality” proposed by [7], i.e. acting with enough caution to be considered equals to humans in a specific domain, thus helping to give a moral status to robots.

5.4 Ethical reasoning

Given the current state of law and the common deployment of robots on battlefields, granting robots with ethical reasoning have to be rooted in a legally relevant framework, that is Just-War Theory [35]. Laws of War and Rules of Engagement have to be the basic set of rules for robots. Still, battlefields being complex environments ethics needs to be integrated into robots with a hybrid approach combining learning capabilities and experience with ethical theories. In the case of authority sharing, two frameworks seem relevant at the moment : case-based reasoning [2] and Arkin’s reactive/deliberative architecture [3]. What seems applicable in case of an ethical conflict is to give the authority to the operator and to use the robotic agent both to assist him during the reasoning, i.e. by displaying relevant information on an appropriate interface, and to act as an ethical handrail in order to make sure that the principles of the Laws of War, e.g. discrimination or proportionality, are respected.

6 CONCLUSION AND FURTHER WORK

The main drawback of the implementation of ethics into autonomous armed robots is that, even if the technology, the autonomy and the lethal power of robots increase, the legal and philosophical frameworks do not take them into account, or consider them only from an anthropocentric point of view. Authority sharing allow a coupling between a robot and a human operator, hence a better compliance with ethical and legal requirements for the use of autonomous robots on battlefields. It can be achieved with vicarious liability, a good situational awareness produced by tracking both the robot and the operator’s ”states”, and a hybrid model of ethical reasoning – allowing adaptability in complex battlefields environments.

We are currently building an experimental protocol in order to test some of our proposals, namely automous armed robots that embed ethical reasoning while sharing authority with a human operator. We have constructed two fully-simulated battlefield scenarios in which we will test the compliance of the system with a specific principle of the Laws of War (proportionality and discrimination). These scenarios feature hostile actions done towards the robot or its allies, e.g. throwing rocks or planting explosives, that need to be handled while complying with a set of rules of engagement. During the simulation, the operator is induced to produce an immoral behaviour, provoking an authority conflict in which we expect the robot to detect the said behaviour and to take authority from the operator: the authority conflict thereby generated has to be solved by the robot via the production of a morally correct behaviour. Since the current state of our software does not yet allow the robotic agent to actually observe the operator, we are working on some pre-defined evaluations of actions in order for the robot to be able to detect unwanted behaviours, and to act accordingly.

REFERENCES

[1] K. Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, 35–52, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[2] M. Anderson, S. Anderson, and C. Armen, 'An approach to computing ethics', in *IEEE Intelligent Systems*, pp. 56–63, (July/August 2006).

[3] R.C. Arkin, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', Technical report, Georgia Institute of Technology, (2007).

[4] R.C. Arkin, P. Ulam, and A.R. Wagner, 'Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception', in *Proceedings of the IEEE*, volume 100, pp. 571–589, (2011).

[5] P. Asaro, 'What should we want from a robot ethic?', *International Review of Information Ethics*, **Vol. 6**, 9–16, (Dec. 2006).

[6] S. Baron-Cohen, 'The development of a theory of mind in autism: deviance and delay?', *Psychiatrics Clinics of North America*, **14**, 33–51, (1991).

[7] N. Bostrom and E. Yudkowsky. The Ethics of Artificial Intelligence. Draft for Cambridge Handbook of Artificial Intelligence, 2011.

[8] S. Bringsjord and J. Taylor, *The Divine-Command Approach to Robot Ethics*, 85–108, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[9] D.J. Chalmers, 'Facing up to the problem of consciousness', *Journal of Consciousness Studies*, **2**(3), 200–219, (1995).

[10] C. Cloos, 'The utilibot project: An autonomous mobile robot based on utilitarianism', in *2005 AAAI Fall Symposium on Machine Ethics*, (2005).

[11] Fr. Dehais, C. Tessier, and L. Chaudron, 'Ghost: Experimenting conflicts countermeasures in the pilot's activity', in *IJCAI'03*, Acapulco, Mexico, (2003).

[12] D. Dennett, *When HAL Kills, Who's to Blame?*, chapter 16, MIT Press, 1996.

[13] H.T. Engelhardt, *The Foundations of Bioethics*, Oxford Uninversity Press, Oxford, 1986.

[14] K. Erol, J. Hendler, and D. Nau, 'HTN planning: complexity and expressivity', in *AAAI'94*, Seattle, WA, USA, (1994).

[15] J.G. Ganascia, 'Modeling ethical rules of lying with answer set programming', *Ethics and Information Technology*, **9**, 39–47, (2007).

[16] M Guerini and O. Stock, 'Towards ethical persuasive agents', in *IJCAI Workshop on Computational Models of Natural*, (2005).

[17] G. Harman and S. Kulkarni, *Reliable Reasoning: Induction and Statistical Learning Theory*, MIT Press, 2007.

[18] H. Hexmoor, C. Castelfranchi, and R. Falcone, *Agent Autonomy*, Kluwer Academic Publishers, 2003.

[19] K. Himma, 'Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?', in *7th International Computer Ethics Conference*, San Diego, CA, USA, (July 2007).

[20] *Handbook of cognitive task design*, ed., E. Hollnagel, Mahwah, NJ: Erlbaum, 2003.

[21] H. Huang, K. Pavek, B. Novak, J. Albus, and E. Messin, 'A framework for autonomy levels for unmanned systems ALFUS', in *AUVSIs Unmanned Systems North America 2005*, Baltimore, MD, USA, (2005).

[22] H. Jonas, *Das Prinzip Verantwortung. Versuch einer Ethik fr die technologische Zivilisation*, Insel Verlag, Frankfurt, 1979.

[23] D. Kortenkamp, P. Bonasso, D. Ryan, and D. Schreckenghost, 'Adjustable autonomy for human-centered autonomous systems', in *Proceedings of the AAAI 1997 Spring Symposium on Mixed Initiative Interaction*, (1997).

[24] P. Lin, G. Bekey, and K. Abney, 'Autonomous military robotics: Risk, ethics, and design', Technical report, California Polytechnic State University, (2008).

[25] H. Lipson, J. Bongard, and V. Zykov, 'Resilient machines through continuous self-modeling', *Science*, **314**(5802), 1118–1121, (2006).

[26] H. Lipson and J.C. Zagal, 'Self-reflection in evolutionary robotics: Resilient adaptation with a minimum of physical exploration', in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 2179–2188, (2009).

[27] G.J. Lokhorst, 'Computational meta-ethics: Towards the meta-ethical robot', *Minds and machines*, **6**, 261–274, (2011).

[28] G.J. Lokhorst and J. van den Hoven, *Responsibility for Military Robots*, 145–156, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[29] B. McLaren, 'Computational models of ethical reasoning: Challenges, initial steps, and future directions', in *IEEE Intelligent Systems*, pp. 29–37, (July/August 2006).

[30] S. Mercier, C. Tessier, and F. Dehais, 'Dtection et rsolution de conflits dautorit dans un systme homme-robot', *Revue dIntelligence Artificielle, numro spcial 'Droits et Devoirs dAgents Autonomes'*, **24**, 325–356, (2010).

[31] S. Miller and M. Selgelid, *Ethical and Philosophical Consideration of the Dual-Use Dilemma in the Biological Sciences*, Springer, New York, 2009.

[32] T. Nagel, 'What is it like to be a bat?', *The Philosophical Review*, **83**(4), 435–450, (1974).

[33] *Teleological Language in the Life Sciences*, ed., L. Nissen, Rowman and Littlefield, 1997.

[34] R.G. Olson, *Deontological Ethics*, The Encyclopedia of Philosophy, Collier Macmillan, London, 1967.

[35] B. Orend, *The Morality of War*, Broadview Press, Peterborough, Ontario, 2006.

[36] T. Pichevin, 'Drones arms et thique', in *Penser la robotisation du champ de bataille*, ed., D. Danet, Saint-Cyr, (November 2011). Economica.

[37] S. Pizziol, F. Dehais, and C. Tessier, 'Towards human operator state assessment', in *1st ATACCS (Automation in Command and Control Systems)*, Barcelona, Spain, (May 2011).

[38] D. Premack and G. Woodruff, 'Does the chimpanzee have a theory of mind?', *The Behavioral and Brain Sciences*, **4**, 515–526, (1978).

[39] S. Rameix, *Fondements philosophiques de l'thique mdicale*, Ellipses, Paris, 1998.

[40] J. Rawls, *A Theory of Justice*, Belknap Harvard University Press, Harvard, 1971.

[41] J.-J. Rousseau, *Du contrat social*, 1762.

[42] J.-P. Sartre, *L'existentialisme est un humanisme*, Gallimard, Paris, 1946.

[43] D. Schreckenghost, D. Ryan, C. Thronesbery, P. Bonasso, and D. Poirot, 'Intelligent control of life support systems for space habitat', in *Proceedings of the AAAI-IAAI Conference*, Madison, WI, USA, (1998).

[44] N. Sharkey, 'Death strikes from the sky: the calculus of proportionality', *Technology and Society Magazine, IEEE*, **28**(1), 16–19, (2009).

[45] B. Skyrms, *Evolution of the Social Contract*, Cambridge University Press, Cambridge, UK, 1996.

[46] R. Sparrow, 'The Turing triage test', *Ethics and Information Technology*, **6**(4), 203–213, (2004).

[47] R. Sparrow, 'Killer robots', *Journal of Applied Philosophy*, **24**(1), 62–77, (2007).

[48] R. Sparrow, *Can Machine Be People?*, 301–315, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[49] J. Sullins, 'When is a robot a moral agent?', *International Journal of information Ethics*, **6**(12), (2006).

[50] C. Tessier and F. Dehais, 'Authority management and conflict solving in human-machine systems', *AerospaceLab, The Onera Journal*, **Vol.4**, (2012).

[51] G. Veruggio, 'Roboethics roadmap', in *EURON Roboethics Atelier*, Genoa, (2011).

[52] L. Vikaros and D. Degand, *Moral Development through Social Narratives and Game Design*, 197–216, Ethics and Game Design: Teaching Values through Play, IGI Global, Hershey, 2010.

[53] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Rights from Wrong*, Oxford University Press, New York, 2009.

[54] K. Warwick, *Robots with Biological Brains*, 317–332, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[55] K. Warwick, D. Xydas, S. Nasuto, V. Becerra, M. Hammond, J. Downes, S. Marshall, and B. Whalley, 'Controlling a mobile robot with a biological brain', *Defence Science Journal*, **60**(1), 5–14, (2010).

[56] D.D. Woods, E.M. Roth, and K.B. Bennett, 'Explorations in joint human-machine cognitive systems', in *Cognition, computing, and cooperation*, eds., S.P Robertson, W. Zachary, and J.B. Black, 123–158, Ablex Publishing Corp. Norwood, NJ, USA, (1990).