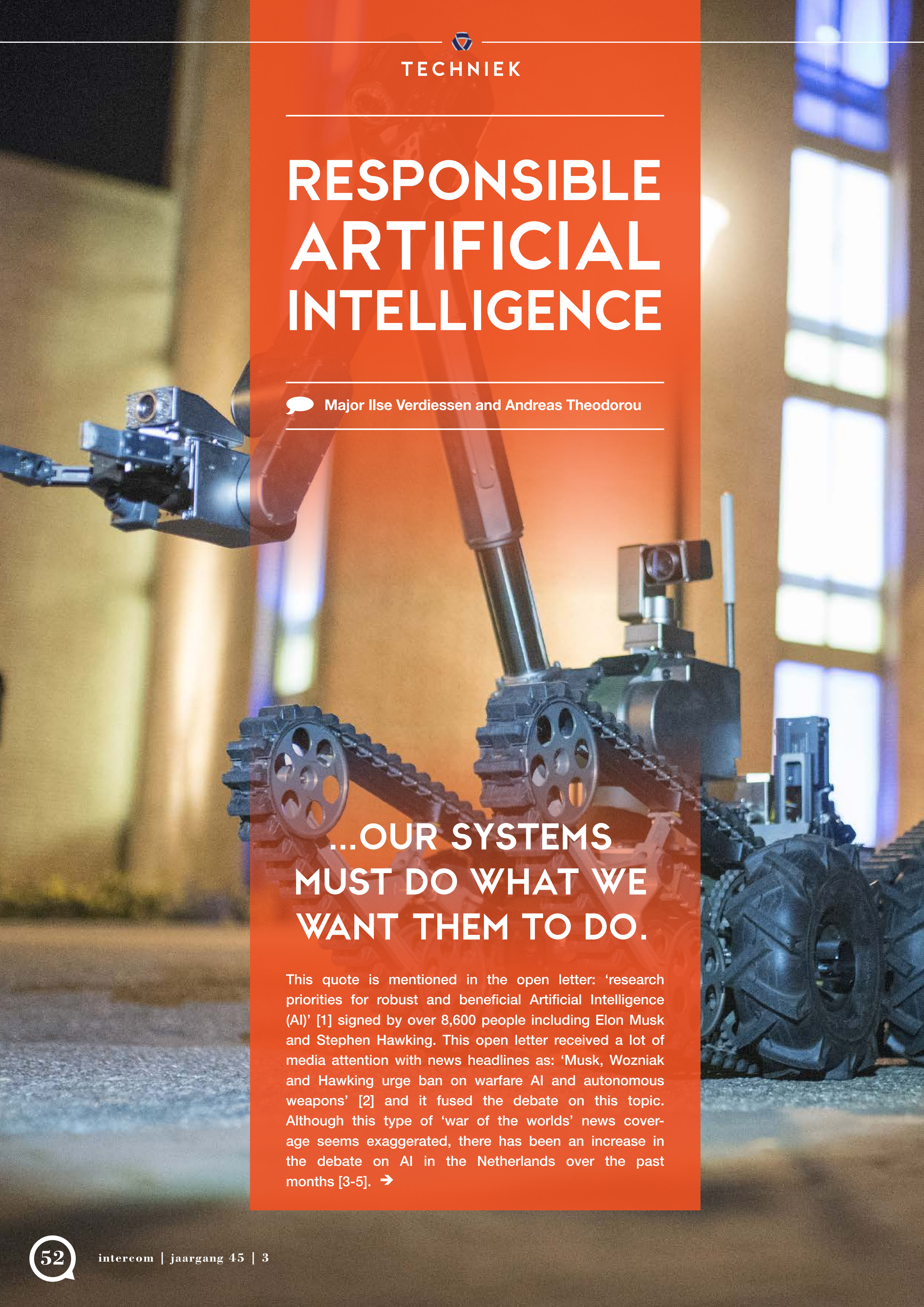


RESPONSIBLE ARTIFICIAL INTELLIGENCE

Major Ilse Verdiessen and Andreas Theodorou

...OUR SYSTEMS MUST DO WHAT WE WANT THEM TO DO.

This quote is mentioned in the open letter: ‘research priorities for robust and beneficial Artificial Intelligence (AI)’ [1] signed by over 8,600 people including Elon Musk and Stephen Hawking. This open letter received a lot of media attention with news headlines as: ‘Musk, Wozniak and Hawking urge ban on warfare AI and autonomous weapons’ [2] and it fused the debate on this topic. Although this type of ‘war of the worlds’ news coverage seems exaggerated, there has been an increase in the debate on AI in the Netherlands over the past months [3-5]. ➔

Introduction

Artificial Intelligence (AI) is not just a futuristic science-fiction scenario in which human-like robots, like the Cylons in Battlestar Galactica, are planning to take over the world. Many AI applications are already being used today. Smart meters, autopilots and self-driving cars are examples of this. One of the applications of AI is Autonomous Weapons. Research showed that Autonomous Weapons are increasingly deployed on the battlefield [6]. It is already reported that China has autonomous cars which carry an armed robot [7]. Russia claims it is working on autonomous tanks [8] and in May of this year the US christened their first 'self-driving' warship [9].

Autonomous systems can have many benefits, for example when the autopilot of the F-16 autonomously prevents a crash caused by the loss off altitude and inaction of the unconscious pilot [10]. Yet the nature of autonomous weapons might also lead to uncontrollable activities and societal unrest. As large scale deployment of AI on the battlefield seems unavoidable, the discussion about ethical and moral responsibility is imperative.

In this article we first introduce Artificial Intelligence and the view on morality in relation to AI. Secondly we explain the concepts of AI by using the model in figure 1 and describe todays research on these concepts.

We conclude with implications for the Defence organization and developments in research.

Artificial Intelligence

Artificial Intelligence can be defined as 'intelligence exhibited by machines' [11]. A machine (or system) shows intelligent behaviour if it can select an action as a reaction to an observation in its environment. The intervention of the autopilot that prevented the crash of the F-16 is an example of this 'action selection'. The autopilot assessed its environment - in this case the rapid loss of altitude and the fact that the pilot did not act on warning signs or acted to improve the situation - and it pulled up to a safe altitude.

In scientific literature, AI is described as more than an Intelligent System alone. It is characterized by the concepts of Adaptability, Interactivity and Autonomy [12] as depicted in the second layer of figure 1 [13]. These characteristics may lead to undesirable behaviour or uncontrollable activities of AI as scenarios of many science fiction movies have shown us. Although these scenarios are often not realistic, a growing body of researchers is focusing on responsible design of AI, which incorporates social and ethical values, to prevent societal concerns about this kind of technology. These principles of Responsible AI are Transparency, Accountability and Responsibility which are depicted in the outer layer of the model (figure 1).

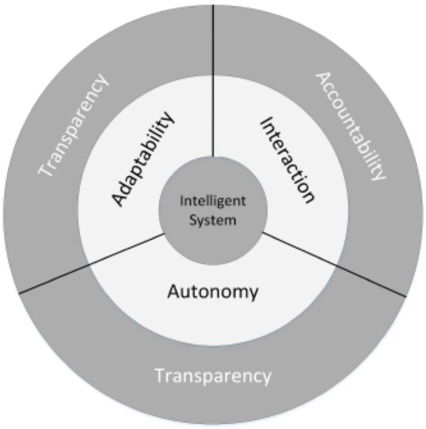


Figure 1: Concepts of Responsible AI

Adaptability means that the system can change based on its interaction and can learn from its experience.

Machine learning techniques are an example of this. Interactivity occurs when the system and its environment act upon each other and Autonomy means that the system itself can change its state. The interaction between autopilot and the F-16 to make the autonomous decision to gain altitude is an example of this [10].

Morality in Artificial Intelligence

This threefold characterization of AI is also referred to by Wallach and Allen in their book on Moral Machines [14]. They present a framework depicted in figure 2 based on two dimensions, autonomy and sensitivity to values, to understand the pathway to engineering moral AI. According to the authors simple tools, such as a hammer, do not have either autonomy or sensitivity to values and are not considered to be moral.

Operational morality is in the low end of their framework and machines in this area lack autonomy and sensitivity, but in their design the values of their engineers are incorporated. A smart meter which has encryption to secure the privacy of the user is an example of this. The next stage is functional morality in which the machine either has significant autonomy and little ethical sensitivity such as the F-16 autopilot, or low autonomy and high ethical sensitivity such as an ethical decision support system for doctors. The last category is a full moral machine which has a high autonomy and high sensitivity to values which does not exist right now, but the science fiction literature and movies portrait many examples (e.g. Data in Star Trek or Ava in the movie Ex Machina).

Responsible Artificial Intelligence

Researchers at the University of Bath videoed a low-cost, Arduino-based robot, showed this to a group of people and asked them what the robot was doing [15]. The robot could simply move around a room, avoid objects, while searching for humans. When it finds a human it flashes lights and then continues seeking another one. Nothing complex or computational intensive, such as Machine Learning. It could find humans using a cheap heat sensor, but could neither classify nor understand who it tracked. Yet, some of the answers were noteworthy.

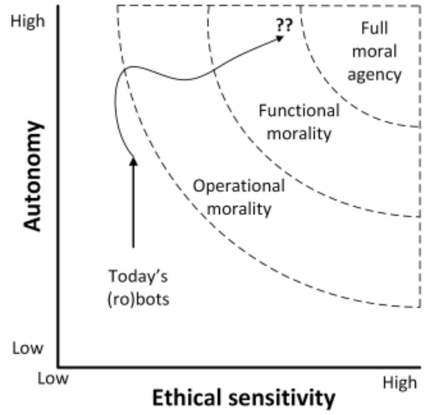
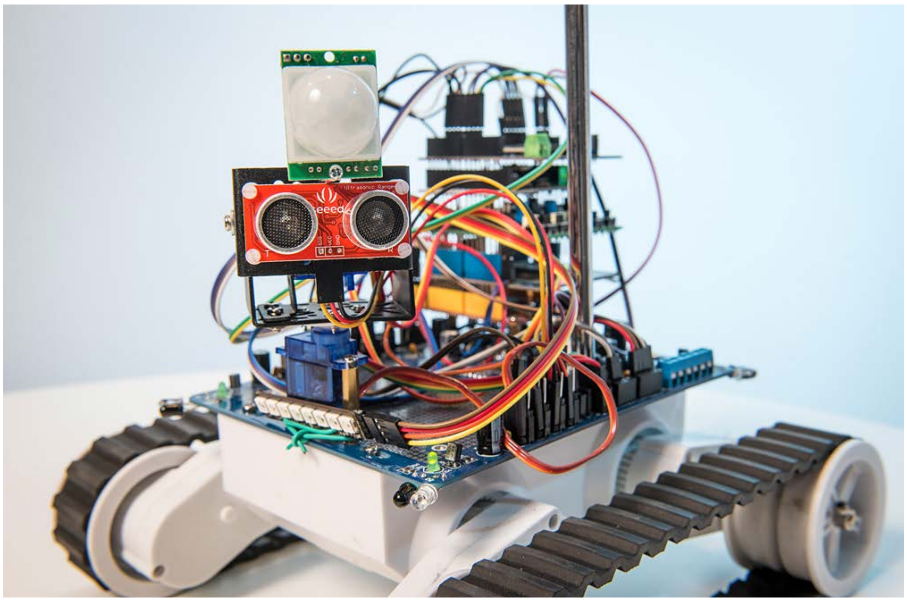


Figure 2. Stages of moral development in AI (retrieved from Wallach & Allen, 2008, p. 26)

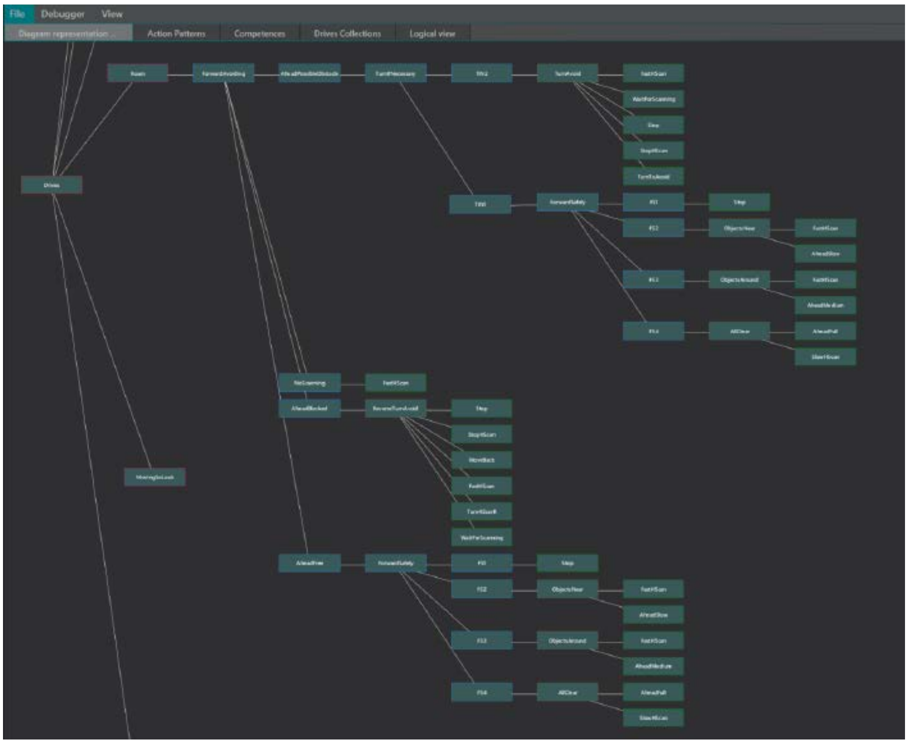


The R5 Robot used at the experiment

Based on cues from the environment, and the imaginations of the people, they came up with all sorts of ideas about what the robot was up to – views that were generally quite wrong.

For instance, there is a bucket in the room, and several people were sure the robot was trying to throw something into it. Others noticed an abstract picture in the room and wondered if the robot was going to complete the picture. These results are from people were mainly graduates in professional jobs, and several had science and technology degrees. Almost all used computers every day [15].

Although we did not program the robot, nor create the room explicitly to mislead, the observers were deceived. It's almost as if the participants believed that machines have minds of their own. The fact that we perceive them as intelligent is partly why they have such potential. Robots are moving beyond industrial, commercial and scientific applications, and are already used in hospitals and care homes. Imagine an autonomous robotic system built for providing health-care support to the elderly, who may be afraid of it, or simply distrust it. They may not allow the robot to interact with them. In such a scenario human lives are at risk, as they may not get the required medical treatment in time, as a human overseeing the system must detect lack of interaction and intervene. Conversely, if the human user places too much trust in an intelligent system, it could lead to misuse, over-reliance, and disuse of the system. In our example of a health-care robot, if the agent malfunctions and its patients are unaware of its failure to function, the patients may continue using the robot, risking their own health. Where errors occur, they must be addressed, in some cases redressed, and in all cases used to reduce future



A plan editor allowing real-time display of transparency-related information

mishaps. This is nice in principle, but probably impossible to implement.

While we can strive to make it true, we must ensure legal paths which should be addressed by ownership of our responsibility. Otherwise we the same problem as we the one have with militias, lack of effective accountability.

Currently, even where we have inadequate control over something as in the case of young children, owned animals, and operated machinery. If we lose control over entities that we have responsibility for, they themselves cannot be held accountable. We are held responsible for that loss of control, including whatever actions comes as a consequence of it. If our pet or car kills a human, we are not held accountable for murder, but we can and should be held accountable for negligence and manslaughter. Similarly, robots belong to us.

People, governments and companies build, own and program robots. Whoever owns and operates a robot is responsible for what it does [16]. Assigning responsibility to the artefact for actions we designed it to execute would be to deliberately disavow our responsibility for that design. Unexpected effects, which may 'emerge' during the operation of complex systems, do not revise the designers' responsibility to observe and account for such effects. Neither should courts of law. Frameworks such as EPSRC Principle of Robotics ensure that it should be possible to find out who is responsible for any intelligent agent [17]. To avoid such situations in the first place, proper calibration of trust between operators and their agents is critically important, if not essential, in high-risk scenarios, such as the usage of robots in the military or for medical purposes. Accurate calibration of trust occurs when the end-user has a mental model of the system and relies on the system within the system's capabilities and is aware of its limitation.

Agents containing the necessary mechanisms to provide meaningful information to its end users, can help improve mental models of AI users. To consider a system transparent to inspection, the end user should have the ability to request accurate interpretations of the robot's capabilities, goals, and current progress in relation to its goals, its sensory inputs, and its reliability, as well as reports of any unexpected events. The information provided by the AI should be presented in a human understandable format. So, the robotic nurse of the future may have transparency built in. Perhaps you can ask it what it's doing and it will tell you by showing you or talking about what's going on in its brain. It would be nice for the user to be able to dial this up or down depending on how familiar they are with the tasks the robot is doing.

Conclusion

Artificial Intelligence is not a science-fiction scenario in distant future. Many applications of AI can already be found in our daily lives. Accidents with Tesla cars make the headlines almost every week which shows the growing interest and sensitivity of this topic. Also the opposition to Autonomous Weapons is becoming louder. Examples of this are the 'Stop Killer Robots' of 61 NGO's directed by Human Rights Watch [18], but also the United Nations are voicing their concerns and state that 'Autonomous weapons systems that require no meaningful human control should be prohibited, and remotely controlled force should only ever be used with the greatest caution' [19].

Developments in research

Despite enormous fluctuations in public profile based on misconceptions, AI has been making steady progress for decades since its inception in the 1950s.

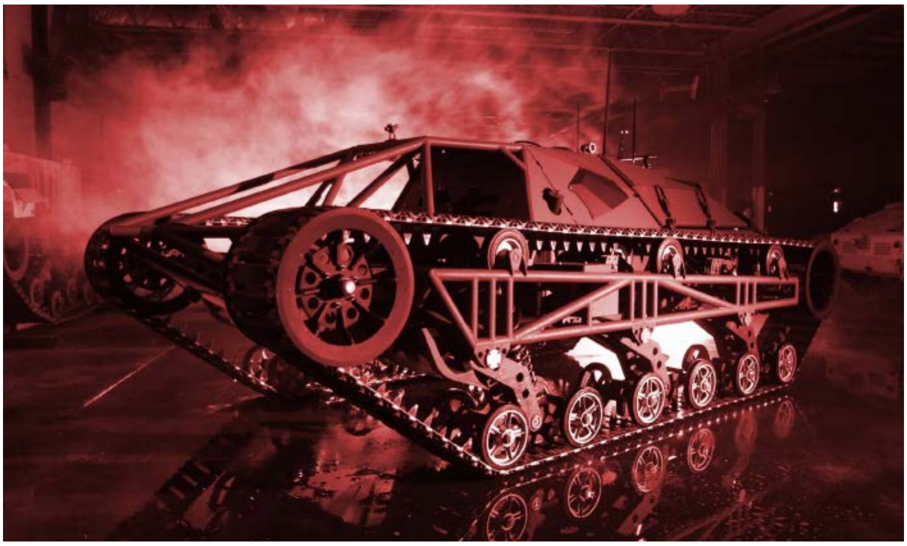
Thanks to increases in computational power and further optimisations of machine learning techniques, there is a sudden public high profile for the field. Companies, such as Google and Microsoft, which were relying on intelligent systems, mainly data analysis, and started to openly discuss and promote AI to the general public.

AI researchers expect this "AI boom" to continue, as the world's largest companies - including investment banks and hedge funds - started heavily investing in AI for data science, business analytics, and automating administrative tasks. One of the most well-marketed and discussed events was the victory of AlphaGo against the world-champion Lee Sedol in the ancient game of Go. Unlike reports by the media, AI did not kill Go – the same way that it didn't kill chess at the beginning of the 21st century. Instead, thanks to AI we learned new moves and countermoves, enhancing our understanding of Go.

Yet, one of the most significant developments is the inclusion of a personal assistance at every smartphone. Whether the mobile platform, GoogleNow, Cortana and Siri, rely of machine learning to provide information on demand or otherwise. They are able to perform simple tasks; such as setting up meetings, answering messages, turning on alarms, and even recommending you tailored selections for food and entertainment. Such systems use Natural Language Processing to convert voice into machine-understandable data. We anticipate intelligent systems to continue integrating in our daily lives. Every major car manufacturer announced the development of their own driverless car system. Originally pushed by companies such as Tesla and Google, driverless cars are now used by Uber as a taxi service and Tesla as a co-pilot system, allowing car owners to switch between "manual" and "automatic" driving. SoftBank/Alderbaran Robotics' upcoming Pepper is already used in beta testing, as a specialist seller in various shops and restaurants, and expected to be employed in Tokyo's 2020 Olympics as an information provider to spectators and athletes.

Implications for the Defence organization

In the near future AI technology will also become more available in the Defence organization. The US Airforce expects the deployment of robots with fully autonomous capabilities between the years 2025 and 2047 [20]. However, the Explosive Ordnance Disposal (EOD) already uses robots to dismantle bombs today and in the procurement of the new submarines the possibilities of unmanned submarines are studied [21]. The Pentagon is experimenting with autonomous drones that determine their own flightpath without human intervention [23]. There are many more applications which can be beneficial for the Defence organization. Goods can be supplied with self-driving trucks leading to a different logistics concept and small UAV's can be programmed with swarm behaviour to support intelligence gathering. Next to the practical application of AI technology, a moral debate is going on. Some scholars argue that, under the same conditions, battlefield robots will behave morally better than human soldiers, because robots do not have an emotional state that makes them less vulnerable [14].



Others counter this argument by stating that in a military context reflection on ethical decision-making is fundamental and robots will lack the ability to ask themselves questions on their ethical choices and actions. Therefore they miss the understandability to make ethical decisions in a complex environment such as a battlefield [20]. The deployment of Autonomous Weaponized Robots on the battlefield is therefore not only a military revolution, but can also be considered as a moral one [22].

Anticipating on these rapid technological developments, we need to build knowledge on AI within the Defence organization in order to procure and deploy these technologies. We also need to get involved in the discussion on the ethical and moral implications of Autonomous Weapons to voice our point of view in the societal debate so that this will not be a pure academic discussion, but also the military application and benefits of AI technology are taken into account.

For references to footnotes see VOVKlict.nl

About the authors

Major Ilse Verdiesen has worked at the Armed Forces since 1995. Currently she is pursuing a Master degree on Information Architecture at the TU Delft. She will graduate in the field of Responsible Artificial Intelligence and will conduct her research at the Scalable Cooperation Lab at MIT. Twitter: @IlseVerdiesen

Andreas Theodorou is a PhD student in the Intelligent Systems group at the University of Bath (UK), where he works under the supervision of Dr Joanna Bryson in Artificial Models of Natural Intelligence. His main research interest is the design and application of intelligent systems, and its effects on human society. Andreas served two years mandatory military service as a Lance Corporal in the Cypriot National Guard, acting in a range of logistics and operations related responsibilities.
Twitter: @RecklessCoding