Responsible innovation, development, and deployment of automated technology

M.M.M. Peeters, T.J. Baar, and M. Harbers

1. Introduction

A heated international debate is taking place about the innovation, development, and deployment of automated military technology, such as remotely-controlled aerial vehicles. Recently, the scope of the debate is extended to the moral concerns about (future) automated technology possibly able to make decisions about the application of kinetic force (e.g. fire a bullet) without human intervention.

In this abstract, we will argue that it is hardly possible to have a discussion about the dangers of automated technology in general, because automated technology is specialist in nature, capable of performing specific tasks within an, often, narrow context [1]. Furthermore, we will argue that automated technology should be designed and developed in a way that supports responsible use from an early design stage all the way to its correct deployment.

2. Automated technology is part of a cognitive system

It is our opinion that (partially) automated systems cannot be regarded as limited to *just* the automated technology. In fact, automated technology is always part of a larger system consisting of both people and machines, collaborating on a given set of tasks [1] [2] [3]. Human-machine collaborative systems are generally referred to as a *cognitive systems* [4] [5]. In cognitive systems, machines carry out the tasks machines are good at (e.g. computations, data analysis) and people carry out the tasks people are good at (e.g. interpreting aggregated data patterns, decision making) [6] [7].

The reason for adding machines to the work environment is to optimally benefit from human capacities by supporting and, in some cases, enhancing them [8]. Exactly for which tasks machine automation provides a viable solution may vary over time and across situations. Therefore, an important aspect of cognitive systems engineering is how to properly and *dynamically* allocate tasks to either the machine(s) or the human(s) [7]. The goal of cognitive systems engineering, then, is not *just* to optimize the technology, but to optimize *the joint performance of the cognitive system* through 1) the design of the technology and 2) the organization and coordination of task allocation between the technology and the people involved.

Within cognitive systems, machines (i.e. automated technology) are used to perform specific (parts of) a task automatically, without human intervention. The 2012 DoD Taskforce Report describes this as follows: 'Autonomy is a capability (or a set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, "self-governing."' [9]. In this sense, automated technology can always be regarded as autonomous, as long as the scope of the system is sufficiently constrained [10]. However, by placing the scope of the system on the *cognitive system*, the technology is never considered to be fully autonomous: it is part of a man-machine collaboration.

By looking at technology as part of a cognitive system, designers do not merely design a piece of technology; they design work situations and task procedures in which both the human and the technology play their part. This means that the introduction of new technology to the work place changes the roles and tasks of people involved [1] [4]. Therefore, researchers emphasize that a good cognitive system design must take human values, such as *moral responsibility*, into account throughout the design process. This is done by 1) involving users in the design process, 2) extensively testing and evaluating designs, and 3) developing design practices that support responsible technological innovation [4] [11] [12] [13] [14] [15].

To support responsibility with automated technology and to promote responsible use of that technology within cognitive systems, a concerted effort is needed of designers, manufacturers, and users. We will now provide four possible measures that can be taken to support responsible

innovation. Firstly, designers should include users and other stakeholders in the design process from an early design stage. For reasons of simplicity and efficiency, designers generally make assumptions about the work environment for which the technology is designed, for instance, about the types of objects the machine must be able to recognize in its environment. Therefore, the design process should (a) closely involve users and other stakeholders in the formulation of assumptions made about the work environment in which the technology will be used, and (b) include an ongoing dialogue between designers and end-users on the stipulation of deployment guidelines. Secondly, when delivering a piece of technology to the user, the intended purposes, tasks, and context for which it was designed must be clearly stipulated, including the assumptions made about the work environment and task procedures for which the technology is used. Thirdly, automated (military) technology must include automated support for safety and security checks to verify that it is used in accordance with the deployment guidelines. And lastly, because the work environment and/or task procedures may change over time, it is important to perform systematic re-evaluations of the work environment to identify risks of unexpected, undesirable, or even dangerous situations.

The following example illustrates the implications of designing automated technology as part of a cognitive system. Consider a piece of automated technology that is to be used for sea mine identification on board of a frigate at open sea. It checks all objects in the frigate's surroundings for potential sea mines based on shape, location, and/or movement. If it detects a possible sea mine, the image is presented to the operator who then decides whether or not to take subsequent action. Let us assume that the machine contains a self-adaptive (learning) algorithm: it updates its reasoning rules based on corrections provided by the human operator.

Now imagine that recently a new type of sea mine has been produced which looks different from the old ones. The human operator knows that the system can learn to recognize these mines, but it must be taught before it is able to do so. Now, the operator has a different type of task to perform: he or she must scan the environment for this new type of sea mine and, at each encounter, mark the object as `sea mine' for the machine to adjust its internal reasoning rules. After multiple encounters, the system is able to detect the new type of sea mine without supervision and the operator can go back to his old job: making judgments about possible sea mines identified by the machine.

In this example, the operator is able to closely and effectively collaborate with the machine due to awareness of the machine's (in)capacities. Moreover, the operator is aware of the shift in roles between the machine and the operator resulting from a change in the work environment and is able to mitigate those changes by employing the machine's capacity to learn. Without this awareness the cognitive system, consisting of the operator and the machine, would malfunction. This example illustrates the need for clear communication and documentation about the machine's (in)capacities and inner workings, and about the assumptions made during the design phase regarding the work environment and task procedures for which the machine will be used. Clarity about these features allows the operator to use the machine as intended and to ensure that the machine is only used in contexts where the respective assumptions are valid.

3. Conclusion

This abstract aims to raise awareness of the role technology plays in our work environments and the specialist nature of (most) technological devices. Automated technology is generally designed to perform specific tasks within specific contexts. Taking technology out of the context for which it was designed can result in misuse or malfunctioning, leading to irresponsible or, at the least, unexpected behaviour, causing dangerous situations.

To prevent irresponsible use of (military) automated technology, we call for the specification of guidelines and policies for the design, development, and deployment of such technology. The design process of automated should follow strict design methods that extensively test and evaluate the performance of the joint human-machine (i.e. cognitive) system before putting it to use in the actual work place. The effects on the roles and work flows of people involved must be thoroughly investigated from an early stage. In addition, the assumptions that have been made about the work

environment and task procedures must be clearly formulated and communicated to the users. The use of automated technology for purposes - or in contexts - other than its original design is, in our view and without further investigation and re-evaluation, irresponsible.

Bibliography

[1] J. M. Bradshaw, R. R. Hoffman, M. Johnson and D. D. Woods, "The Seven Deadly Myths of "Autonomous Systems"," *IEEE Intelligent Systems,* vol. 28, no. 3, pp. 54-61, 2013.

[2] M. Johnson, J. M. Bradshaw, P. J. Feltovich, R. R. Hoffman, C. Jonker, B. Van Riemsdijk and M. Sierhuis, "Beyond cooperative robotics: The central role of interdependence in coactive design," *IEEE Intelligent Systems,* vol. 26, no. 3, pp. 81-88, 2011.

[3] J. M. Bradshaw, V. Dignum, C. Jonker and M. Sierhuis, "Human-agent-robot teamwork," *IEEE Intelligent Systems,* vol. 27, no. 2, pp. 8-13, 2012.

[4] E. Hollnagel and D. D. Woods, Joint cognitive systems: Foundations of cognitive systems engineering, CRC Press, 2005.

[5] A. Leveringhaus and T. De Greef, "Tele-operated Weapons Systems: Safeguarding Moral Perception and Responsibility," in *Hitting the target?*, 2013, pp. 57-64.

[6] M. L. Cummings and K. M. Thornburg, "Paying Attention to the Man behind the Curtain," *IEEE Pervasive Computing,* vol. 10, no. 1, pp. 58-62, 2011.

[7] T. De Greef and A. Leveringhaus, "Safeguarding moral perception and responsibility via a partnership approach," in *International Conference on Naturalistic Decision Making*, 2013.

[8] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors,* vol. 39, pp. 230-253, 1997.

[9] U.S. Department of Defense, "The role of autonomy in DoD systems," 2012.

[10] M. Noorman and D. G. Johnson, "Negotiating autonomy and responsibility in military robots," *Ethics and Information Technology,* vol. 16, no. 1, pp. 51-62, 2014.

[11] B. Friedman, P. H. Kahn and A. Borning, "Value sensitive design and information systems.," *Human-computer interaction in management information systems: Foundations,* vol. 5, pp. 348-372, 2006.

[12] M. A. Neerincx and J. Lindenberg, "Situated cognitive engineering for complex task environments," in *Naturalistic Decision Making and Macrocognition*, Aldershot, UK, Ashgate Publishing Limited, 2008, pp. 373-390.

[13] M. Noorman, "Responsibility Practices and Unmanned Military Technologies," *Science and engineering ethics,* pp. 1-18, 2013.

[14] G. R. Lucas, "Legal and Ethical Precepts Governing Emerging Military Technologies: Research and Use," *Amsterdam Law Forum,* vol. 6, no. 1, pp. 23-34, 2014.

[15] M. J. Van den Hoven, "Moral responsibility, public office and information technology," in *Public administration in an information age: a handbook*, 1998, pp. 97-112.