ROBONOMICS
The Journal of the Automated Economy
https://journal.robonomics.science

Book review of Umbrello, S. (2022). Designed for Death: Controlling Killer Robots. Trivent Publishing. 236 pp. ISBN: 978-615-6405-38-8.

Reviewed by

Jon-Arild Johannessen [1] ✉

Type: Book review

Citation: Johannessen, J.-A. (2023). Book review of Umbrello, S. (2022). Designed for Death: Controlling Killer Robots. Trivent Publishing. 236 pp. ISBN: 978-615-6405-38-8. *ROBONOMICS: The Journal of the Automated Economy,* 4, 40

[1] Professor, Kristiania University College, Norway; email: Jon-arild.johannessen@kristiania.no
✉ Corresponding author

The main point of the book "Designed for death" is that people can have full control over such weapon-systems when certain ethical conditions are met.

As I see it, it is precisely the idea of full control with a weapon system based on artificial intelligence that is the strength and weakness of this book. With a really good insight, Steven Umbrello has reflected in this book on the possibilities for humans to have control over this war technology. However, we all remember the strict systems the Soviet Union had for launching its nuclear programs and how they could fail, if human judgment had not intervened and ended the incipient nuclear war between the United States and the Soviet Union[1]. The point is that no matter how good control humans have with artificial intelligence in weapon systems and no matter how advanced this technology is, there will always be some weaknesses in this technology that can cause humans to lose control.

Umbrello has discussed these reflections in his book. My point is simply that his conclusion is that given some value judgments in the development of the design of this technology, then humans will have full control over the technology. Umbrello uses systems thinking in its reflections on the issue of human control of weapon systems based on artificial intelligence. This is certainly a strength of the book. The point, however, is that if we see these weapon systems in a systemic perspective, there is a connection that is decisive and which we think Umbrello under-communicates in his book.

This simple connection is that although values are important, technology affects these values so that they change when the technology is used. One can say that technology affects values and the changed values in turn affect technology. This is on the same level as the work environment affecting performance, but performance in turn affecting the work environment. If you have a model that says that it is the working environment that affects performance, then this model provides guidance on how to improve performance. If, on the other hand, you have a model that says that performance affects the working environment to a great extent, then you will have completely different strategies for developing the workplace. Transferred to the connection between military technology and values, the same type of considerations can be used as a basis. Therefore, it will be important when reflecting on artificial intelligence in weapon systems and values that you take both models into consideration. It is in this context that one can draw in the late Stephen Hawkins, and the letter he and twelve others wrote related to the control of artificial intelligence[2]. Their point was, and is, that it is desirable but almost impossible to manage and control artificial intelligence. If their point is correct, then one should tread carefully when claiming that artificial intelligence can be controlled through value design and value assessments. In this connection, it may be tempting to quote Gregory Bateson (2004): '*Where angels fear to tread, men should tread very carefully'*.

The criticism against Umbrello's reflections is largely related to what he himself calls fully autonomous air force systems. To a large extent, his arguments will hold when it comes to most signals that will then trigger a response from these systems. The point, however, is that there will always be cases that an algorithm designer has not thought through, and which can trigger responses from these weapon systems that are not desirable. We have seen this in the discussion of effective weapon umbrellas to defend the United States. Most people who had insight into these programming techniques stated that in some contexts these weapon systems that lay behind the umbrellas could be triggered by built-in errors in the programs. When the weapon system is of a type that has the potential to create catastrophic conditions, then one should think twice.

Umbrello's solution to this possible conflict is that the designer and the user have extensive communication. This sounds plausible, but the question is not whether this is necessary, but whether it is sufficient. Why should such communication do anything about built-in errors and unknown situations that can arise in a conflict situation? If one were to propose something to reduce the crisis that may arise, then it would have to be that both designers, users and experts in ethics and artificial intelligence should join in such communication. One can

also ask questions about what Umbrello means by users in this context. Is it the military that will use the systems? Are they the ones who may be exposed to the consequences of these weapons systems? Umbrello tries to help its arguments by saying that the technological systems must respond sensibly. The point is that artificial intelligence is designed to little if any degree for reason, but for rationality. Algorithms are rational, more rational than humans can credit. On the other hand, there is little to suggest that they are emotionally or socially upbeat. In this context, one could say that: If everything but the rational disappears, then human madness has reached its final peak. This is where the new technology has the greatest potential for ethical collapse. It becomes so rational that every reason is measured, weighed and numbered. If we then put the rational logic as input in an intelligent robot, then the output becomes the ethical and emotional wreck. If this statement has anything to do with it, then one should really be careful about hoping and believing that the rational weapon systems can be ethically and value-wise controlled.

When the objections shown above are taken into account, we must say that Umbrello's book is very important as a reflection on technology, ethics and weapon systems. In particular, this applies to weapon systems that are autonomous, semi-autonomous and respond to incoming signals. However, we are of the opinion that no matter how much value design and values are connected to autonomous weapon systems, the technology will always affect the values so that they are changed when the technology is applied. The opposite model that the values will influence the technology will apply in some cases, but not where the consequences of using such a technology exceed the dangers of using it.

Endnotes:

[1] See https://en.wikipedia.org/wiki/Stanislav_Petrov
[2] See https://en.wikipedia.org/wiki/Open_Letter_on_Artificial_Intelligence

References:

Bateson, G. (2004). *Angels fear: Towards an Epistemology of the Sacred*. London: Hampton Press.

Received: 26/01/2023
Accepted: 26/02/2023