**Can machines be (regarded as) people?** *[Robot Ethics, Lin&Abney et. Al., 2014]*

In philosophical ethics - and especially in applied ethics - questions about the wrongness of killing are now debated in the context of a distinction between - human beings‖ and - persons‖ (Kuhse and Singer 2002). Human beings are – unsurprisingly - members of the species Homo sapiens and the extension of this term is not usually a matter of dispute. However, in these debates, "persons" functions as a technical term to describe all and only entities that have (at least) as much moral standing as we ordinarily grant to a healthy adult human being. "Moral standing" refers to the power that certain sorts of creatures have to place us under an obligation to respect their interests. Thus, persons are those things that it would be at least as wrong to kill as a healthy adult human being. human beings and that - artificial intelligence‖ would involve the production of such intelligence in a machine. Questions about the moral standing of machines will only arise if researchers succeed in creating such strong AI. The question the Turing Triage Test is designed to answer, then, is "when will machines become persons". Here is the test, as originally described:

**The Turing Triage Test (Robert Sparrow)**

*Imagine yourself the Senior Medical Officer at a hospital which employs a sophisticated artificial intelligence to aid in diagnosing patients. This artificial intelligence is capable of learning, of reasoning independently and making its own decisions. It is capable of conversing with the doctors in the hospital about their patients. When it talks with doctors at other hospitals over the telephone, or with staff and patients at the hospital over the intercom, they are unable to tell that they are not talking with a human being. It can pass the Turing Test with flying colours. The hospital also has an intensive care ward, in which up to half a dozen patients may be sustained on life support systems, while they await donor organs for transplant surgery or other medical intervention. At the moment there are only two such patients.*

*Now imagine that a catastrophic power loss affects the hospital. A fire has destroyed the transformer transmitting electricity to the hospital. The hospital has back up power systems but they have also been damaged and are running at a greatly reduced level. As Senior Medical Officer you are informed that the level of available power will soon decline to such a point that it will only be possible to sustain one patient on full life support. You are asked to make a decision as to which patient should be provided with continuing life support; the other will, tragically, die. Yet if this decision is not made, both patients will die. You face a 'triage' situation, in which you must decide which patient has a better claim to medical resources. The diagnostic AI, which is running on its own emergency battery power, advises you regarding which patient has the better chances of recovering if they survive the immediate crisis. You make your decision, which may haunt you for many years, but are forced to return to managing the ongoing crises.*

*Finally, imagine that you are again called to make a difficult decision. The battery system powering the AI is failing and the AI is drawing on the diminished power available to the rest of the hospital. In doing so, it is jeopardising the life of the remaining patient on life support. You must decide whether to 'switch off' the AI in order to preserve the life of the patient on life support. Switching off the AI in these circumstances will have the unfortunate consequence of fusing its circuit boards, rendering it permanently inoperable. Alternatively, you could turn off the power to the patient's life support in order to allow the AI to continue to exist. If you do not make this decision the patient will die and the AI will also cease to exist.*

*The AI is begging you to consider its interests, pleading to be allowed to draw more power in order to be able to continue to exist. My thesis, then, is that machines will have achieved the moral status of persons when this second choice has the same character as the first one. That is, when it is a moral dilemma of roughly the same difficulty. For the second decision to be a dilemma it must be that there are good grounds for making it either way. It must be the case therefore that it is sometimes legitimate to choose to preserve the existence of the machine over the life of the human being. These two scenarios, along with the question of whether the second has the same character as the first, make up the 'Turing Triage Test'.*

**The importance of the Turing Triage Test**

I noted above that the question of the moral standing of machines will arise with great urgency the moment scientists claim to have created an intelligent machine. Having switched their AI on, researchers will be unable to switch it off without worrying whether in doing so they are committing murder! Presuming that we do not wish to expose AI researchers to the risk that they will commit murder as part of their research, this is itself sufficient reason to investigate the Turing Triage Test. However, the question of when, if ever, AIs will become persons is also important for a number of other controversies in roboethics and the philosophy of artificial intelligence.

As intelligent systems have come to play an increasingly important role in modern industrialised economies and in the lives of citizens living in industrial societies, the question of the ethics of the operations of these systems has become increasingly urgent. At the very least, we need to be looking closely at how these systems function in the complex environments in which they operate and asking whether we are happy with the consequences of their operations and the nature of human interactions with such systems (Johnson 2009; Veruggio and Operto 2006). This sort of ethical evaluation is compatible with the thought that the only real ethical dilemmas here arise for the people who design.

This formulation of the Turing Triage Test introduced the Test in the context of the discussion of the role played by the original Turing Test in the historical debate about the prospects for machine intelligence, which accounts for the reference to the Turing Test in this passage. In particular, in an earlier section of the paper I had argued that in order to be a plausible candidate for the Turing Triage Test, a system would first have to be capable of passing the Turing Test: this assumption is not, however, essential to the Test. It is arguable that killing an artificial intelligence because of a lack of appreciation of its moral standing should be categorised as manslaughter or some other lesser category of offence, rather than murder, on the grounds that it would not involve the deliberate intention to take a life that is essential to the crime of murder.

A crucial question here will be whether a lack of awareness of the moral standing of the entity towards whom one's lethal actions were directed is sufficient to exclude the conclusion that the killing was intentional: in the scenario we are imagining, the actions taken to kill‖ the AI would be deliberate and the intended result would be the destruction of the AI, but the knowledge that the AI was a moral person would be absent. In any case, regardless of whether the appropriate moral or legal verdict is murder, manslaughter, negligent homicide, or some other conclusion, clearly this scenario is one we should strive to avoid or make use of these systems.

[…]