**2023-10-17 Practice Exam, Outline of answers**

1. Human rights

What are the main reasons why the human rights enumerated in the Universal Declaration of Human Rights have authority as normative premises in argumentation about policy in the Netherlands?

*The following answers would be acceptable (some are mentioned in the intro lecture):*

- *The Netherlands is a member of the United Nations, and in the charter of the UN, fundamental human rights are affirmed. These have been articulated in the UDHR.*
- *Because it is a universal declaration of moral and legal rights, the rights that are enumerated in the UDHR hold for humans in all countries, including the Netherlands.*
- *European law has largely adopted the rights from the UDHR as fundamental legal rights in EU countries.*
- *Liao writes that such rights protect the fundamental conditions for pursuing a good life, which is something that we all value.*

2. Four principles

Which fifth principle has been added to the four principles of ethics derived from the Belmont Report, as an extra principle that is needed for trustworthy AI in a European context?

*The principle of "Explicability" is the extra ethics principle (alongside respect for autonomy, non-maleficence, and fairness/ justice) mentioned in the HLEG guidelines on trustworthy AI. This is translated into requirements of "transparency" and "accountability". (Lecture 26 Sept, also mentioned verbally in the intro lecture)*

3. Goodall

Goodall argues that it may be possible to use machine learning to train autonomous vehicles to make some ethical decisions. What data does he think should be used for the training?

*Humans would score potential actions in crash similations/ recordings as more or less ethical, with a diverse set of data (Goodall p. 63). The model would be trained on this data. A similar approach can be seen in MIT's Moral Machines project. (Lecture 12 Sept)*

4. HCAI

What values does Schneiderman claim will be promoted if we adopt human-centered AI instead of full technological automation? In your opinion, would these values actually be promoted by HCAI? Why or why not?

*Schneiderman's article is subtitled "Trustworthy, Safe and Reliable" and those are the target values promoted by HCAI, according to him. He also mentions "self-efficacy, mastery, and responsibility" as intermediating values. (Lecture 14 Sept). However, we saw some reasons to doubt whether the connection between HCAI and those values is as strong as Schneiderman argues. The isolated cases mentioned by Schneiderman hardly prove that these values are promoted by HCAI more than the alternatives. Also, there are tricky responsibility issues with HCAI related both to "handoffs" between human operators and automation, as well as to humans who are influenced by automation bias (as mentioned by Johnson).*

5. Human skills and work

Explain one of Zoller's arguments for the claim that AI might lead to a loss of human value. Give his premises and conclusion in a way that is not obviously invalid.

*Achievement of excellence, practical identity, and the goods of agency all depend on skills (see Zoller p. 87-90).*

*For this reason, skill displacement --- the loss of skills through non-development or failure to maintain them --- implies a significant loss of value within lives.*

*AI has a tendency to take over skills that take time to acquire (see Zoller p. 85), removing the "niche" in which they are developed and maintained.*

*Therefore, AI tends to lead to a significant loss of human value.*

6. Environmental impact of AI

Lin argues that we should take several steps to protect the environment from the impact of machine learning. What are his main proposals?

*Monitoring emissions impact (control leads to measurement)*
*Climate-aware guidelines for AI training (inc. open innovation)*
*Clean energy for data centers*
*(Lecture 21 Sept, also in reading)*

7. Explainability

Why do Hatherley et al. say that there is not just one problem of explainability in medical AI, but multiple problems of explainability? Briefly explain.

Hatherley et al. are referring to Zednik's account of opacity. Quoting Hatherley et al.:

"the aims of post hoc explanation methods are often under-specified, particularly once the problem of agent relativity in explanations is considered. Explanations often need to be tailored to a particular audience in order to be of any use. As Zednik has expressed, "although the opacity of ML-programmed computing systems is traditionally said to give rise to the Black Box Problem, it may in fact be more appropriate to speak of many Black Box Problems—one for every stakeholder." An explanation that assumes a background in computer science, for instance, may be useful for the manufacturers and auditors of medical AI systems, but is likely to deliver next to no insight for a medical professional that lacks this technical background." (p. 4).

8. Self-driving cars

What is meant by *mixed traffic* in the article of Nyholm & Smids? What are their arguments that there is no practical way to avoid serious ethical questions about how to design for mixed traffic? Explain your answers.

*Their implicit definition is a situation in which "there [are] different types of vehicles on our roads with different levels and types of automation," with a particular focus in the article on the case where human-operated vehicles are co-existing with autonomous vehicles.*

*The authors write that "even if highly or even fully automated vehicles will at some later time come to dominate the roads, there will still be a long transition-period during which mixed traffic will be a problem that needs to be dealt with (van Loon and Maartens 2015). Nor should we assume that full automation in all vehicles is an end-point towards which we are moving with necessity (Mindell 2015); mixed traffic may come to mean a mix of vehicles with different levels and types of automation interacting with each other on the road (Wachenfeld et al. 2015; Yang et al. 2016)." The particular ethical issue they seek to address is which "style" of driving --- human or robotic --- should set the ethical norm for driving behavior.*

9. Medical AI

What are three reasons in favor of doing an RCT to test the effect of an expert system in a medical context? Briefly explain.

*Clinician trust is based on RCTs; professional guidelines are evidence-based, which means that efficacy has been proven. Since expert systems only work if there is trust, an RCT is an important means to make the system work. (Lecture 3 Oct).*

*The effect of an expert system considered as a technology in the broad sense (a socio-technical system) can only be known if it is tested for causation in the context of a hospital or clinic, not just in a laboratory. (Argument using points from earlier lecture that were mentioned verbally on 3 Oct)*

*There are reasonable doubts about how well an expert system trained on one set of data will generalize to the messy data of a clinic or hospital (external validity), and the RCT would be likely to address these doubts. (Grote & Genin, mentioned in Lecture 3 Oct)*

10. Generative AI

Khosrowi et al. (2023) argue that we should adopt a "collective-centered creation" model of credit for artistic works created using generative AI. Give an example of how the CCC theory would allocate credit for a Stable Diffusion artwork created by an amateur using the prompt "Cat on a mat". Be sure to consider all the relevant candidate stakeholders who are sometimes part of the "collective".

*See the Khosrowi et al. article, p. 896, where they discuss this case. You could work this out using the principles for assigning credit that we discussed in lecture of 10 Oct. The candidate stakeholders include the prompt user, Stable Diffusion, and those whose artwork was used in the training data for the network behind the GAI.*

11. AI in military applications

What is meant by “meaningful human control”, and what is Johnson’s (2023) main criticism of the idea of meaningful human control?

*There is disagreement about what the term means, but the core idea is that “humans not computers and their algorithms should ultimately remain in control of, and thus morally responsible for, relevant decisions” (Santoni di Sio & van den Hoven 2018). This is more than merely a “human in the loop”, or a human causal control condition. (Lecture 5 Oct)*

*Johnson argues that when humans make decisions with technology, psychology and case studies predict an over-willingness to outsource moral decisions to machines. This is because of various psychological effects such as the Illusion of control, responsibility avoidance, the technological imperative (Johnson: “techno-rationalization”), and automation bias (in relation to “human in the loop” systems). (Lecture 5 Oct)*

12. Superintelligence

Not covered for credit on the exam. An extra credit question about the readings will appear on the exam.