Handout on the topic of ethics at DFKI

1. About this document

This handout serves the purpose of introducing the ethics team at DFKI. To describe our field of work, we address a number of fundamental topics that are relevant for a discourse on AI and ethics. Concrete measures, guidelines or even recommendations for action are not (yet) to be expected here. These require an intensive examination of concrete cases, which will take place in the future. From this discussion, we hope to be able to incorporate generalizable aspects into future versions of this handout.

2. Introduction and motivation

While human individuals measure their actions against moral concepts, ethics reflects moral action. There is no consensus on the question of whether one can speak of (possibly autonomous) action in AI systems at all. What is certain, however, is that AI systems can learn from human actions and that (autonomous) AI systems can influence the lives of individuals and society as a whole, for example by supporting decisions or by the systems physically interacting with humans. For this reason, it is necessary to consider ethical aspects in the research and development of AI systems, although the scope of consideration varies greatly depending on the research project.

3. Goal, scope, mode

The aim of this short handout is to inform DFKI-external people about this policy of DFKI. From our point of view, it is not possible to formulate detailed ethical requirements for the work of more than 20 research groups and hundreds of completely different projects. Therefore, we only want to give some impulses in the following.

4. Ethical principles for research and development at DFKI

The following principles govern research and development on AI systems at DFKI:

AI for the benefits of humans
In all our thoughts and actions within DFKI, the focus is on the well-being of people as individuals and of humanity as a whole. Human rights serve as a basis; we consider people as subjects at all times, never as objects. This is also expressed, among other things, by DFKI's motto "Human Centric AI".

AI and sustainability
With its research and development, DFKI strives to actively advance the UN's global sustainability goals [Bundesregierung 2019].

The principle of acceptability
For any interaction of a human with an AI system, the principle of acceptability [Alpsancar 2018] should apply: Humans must be enabled to take an approving or disapproving position towards the AI system. For this, it must be transparent to the human user at all times whether a person is in an interaction situation with the AI system or not. It must also be clear when and how a person can leave this interaction situation, if necessary including the possible consequences of this decision. Furthermore, it must be possible to explain the

processes within the system to the human if necessary. The principle of acceptability should underlie the design of every system developed at DFKI.

Creating transparency even beyond the boundaries of the systems
In order to adequately evaluate the reliability and usability of an AI system in every respect, it is also necessary to know its limitations. Therefore, for each system developed at DFKI, it should be documented in sufficient granularity for which tasks it is suitable and for which tasks it is not. In the case of a neural network, for example, it should be specified with what kind of data it was trained (e.g., "This system was trained exclusively with photos of Europeans."). This allows a user to understand what outputs to expect for certain inputs, or whether meaningful outputs can be expected at all for certain inputs. Where possible, potential edge cases in particular should be included in the testing of such systems.

AI and consciousness
DFKI's research focuses on the field of weak AI that can be used by humans in the sense of a tool. DFKI does not support research aimed at creating an artificial intelligence with consciousness or with its own complex goals.

Research on AI for military purposes is restricted
DFKI participates only in research and development of AI for military purposes within very strict and limited boundaries, which is described in a separate internal document. There is a corresponding decision by the steering committee on this.

5. Support and collaboration

The DFKI Ethics Team is appointed by the DFKI Steering Committee and legitimized by the Executive Board. It currently consists of Aljoscha Burchardt, Christiane Plociennik, Christian Müller, Iris Merget and Mihai Maftei and is supported by Antonio Krüger, Gesche Joost and Paul Lukowicz. It is available to the workforce as a contact for all ethics issues, whether related to a specific project or generally concerning the work at DFKI. DFKI can draw on experience in involving external ethics experts in projects with a corresponding need. All employees of the DFKI are invited to further develop the handout together with the ethics team.

6. References

[Alpsancar 2018] Alpsancar, Suzana (2018): The ethics of artificial intelligence. https://www-docs.b-tu.de/fg-technikwissenschaft/public/BTU_News_09_2018_The_Ethics_of_k%C3%BCnstliche_Intelligenz.pdf, retrieved 16.01.2020.

[Bundesregierung 2019] Die Bundesregierung: Nachhaltigkeitsziele verständlich erklärt. https://www.bundesregierung.de/breg-en/themen/nachhaltigkeitspolitik/nachhaltigkeitsziele-verstaendlich-erklaert-232174, accessed Jan. 16, 2020.