AI, Human Values and Meaningful Human Control

In our everyday life, we are often exposed to outputs of AI-powered machines that impact our living environment, our way to see the world, and our life in general. Sometimes, machines are designed to allow just a predetermined range of possible reactions, limiting the users' decisions and their liberty to explore other ways to react to their output. Even in those cases in which machines are not designed with the intent of impacting human choices or values, their outcome might bear unexpected consequences for human values' formation and consolidation.

Different values might be affected by specific algorithmic outcomes in ways that are not always intuitive. Let's take the case of increased exposure to certain kinds of images on a social network as a result of automated recommendation – for example of people wearing a certain trendy outfit or replicating a certain behavior or pose. This might well influence the aesthetic preferences of a certain user and their purchasing habits when it comes to clothing and fashion. However, if the represented subject assumes a political meaning or transmits a value judgment concerning the depicted items or actions (regardless of whether this is intentionally planned or an unintended consequence), this might also influence the user's political or social values. Analogous reasoning applies to recommendations of free time activities, fitness activities, as well as to interests or users matching on networking or dating platforms.

Considerations addressing the impact of machines on human choices and value formation constitute one side of the coin when it comes to investigating AI and autonomy. Another important line of questioning concerns human oversight and control of autonomous systems. Contemporary methods in Artificial Intelligence (AI) like Machine Learning (ML) have significantly increased the autonomy of machines. Outputs from machines can now arrive due to features of the input, and the weighting of those features, that were not given to it by human beings. This raises concerns over how humans can retain meaningful human control (MHC) over these machines.

This debate began in the context of lethal autonomous weapons systems (LAWS). The idea that a machine could autonomously target and inflict lethal harm to people without a human being was met with objections from a variety of scholars. Though others have argued that it would be unethical *not* to use them because, for example, they would not get tired or seek revenge and commit war crimes. MHC has since been applied outside of LAWS to, for example, autonomous cars and surgical robots.

The concept of meaningful human control raises questions regarding human autonomy and oversight over the grounding, meaning, and expression of human values. We invite paper proposals from all scholars to explore these topics.

Possible topics and questions

- Analysis of value formation and consolidation as a result of an automated decision process, including cases in which this happens as an unintended consequence.
    - Contributions addressing either cases of intended or unintended value formation (or both) are welcome
    - How do we maintain human values in the case of so-called 'moral machines' which reason about values autonomously?
- Analysis of transformation of human values or intents through interaction with automated systems that were originally designed to support humans in their autonomous decision making

  - When do recommendations cease to support the user's intention and will and become value-transformative?
  - Contributions addressing either cases of intended or unintended value transformation (or both) are welcome
- Analysis of humans' choice limitation depending on predefined algorithmic outcome
  - When is this helpful to achieve autonomously pre-determined goals and when is this negatively affecting our freedom of choice?
  - Analysis of adaptive preference formation in the case of HMI or as a result of choice limitation by an autonomous system
- Exploring ways to maintain meaningful human control over the possible influence of algorithmic outcome on human values and over value-affecting possible scenarios resulting from automated decisions.
  - How does meaningful human control relate to the opacity problem of AI?
  - At what point are the outputs of AI manipulative and deceptive and how can we prevent that?
- Critical approaches to nudging techniques questioning their role in supporting personal goals and autonomous decisions

For consideration for the conference, please submit an abstract of no more than **500 words** to genovesi@uni-bonn.de by **Feb. 20th, 2023**. Please include your name and affiliation. We will notify you regarding the status of your submission by March 7th, 2023.