Responsibility is a term that signifies opportunity to perform a task under a certain circumstance. In the context of AI, it refers to the role of people in their relation with AI systems [Markel-1]. However, that role is performed on an individual level i.e. it depends on personal beliefs, which can differ from one to other, or simply not represent the whole population. The following sections introduce the concept of ethics, highlights the main ethical challenges when developing AI and proposes future discussion about this topic.

Ethics and values

One of the main challenges of ethical reasoning is to determine which moral values to consider and how to prioritise them in a given circumstance [Markel-2]. Values differ from human to human; therefore, it is close to impossible to determine a list of values that can be handled to a system and expect it to work correctly for everyone.

Schwarz [Markel-3] proposes a model where 10 basic values are placed in a circle (Figure x.1), with proximity correlates to similarity, therefore, distant values will be on the opposite side. These values are grouped in 4 different categories (*Openness to Change, Self-transcendence, Conservation* and *Self-enhancement*).

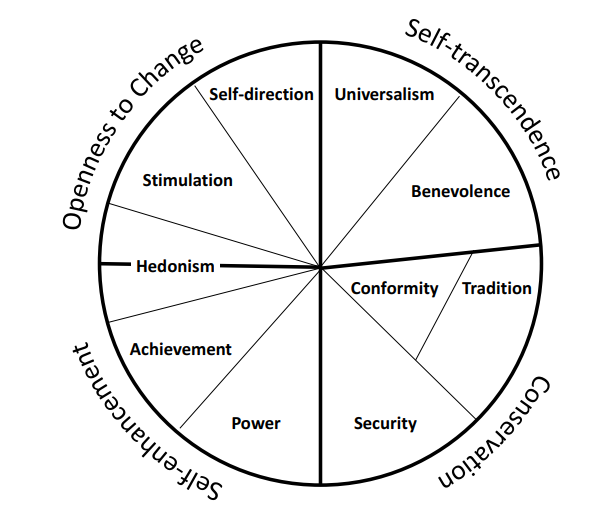


Figure 1: Schwarz value classification

[Markel-7] also proposes a short list of ideal human values, such as ‘*respect and caring for others*. The problem lies in the interpretation of them; for instance, in some cultures, the respect for God weights more, meaning that they must obey certain rules (to respect God) that might disrespect certain groups depending on their gender or ethnicity.

Such values should be used to determine a society global moral conduct, yet they cannot represent the whole moral conduct of society. Moreover, the technique used for evaluating the results seems to unanimously be *generalization*, as democracy has been the system chosen for centuries across many countries. However, the acquisition of a generalized rule is still not enough to solve certain *dilemmas* as different ethical theories (metaethics, normalized-ethics) will lead to distinct solutions [Markel-2]. The capacity to understand such difference plays an important role when designing AI systems that will face situations like the popular *train-trolley* dilemma, where a trolley is facing a group of 5 people attached to a rail track and a system must decide if it should press a switch that will change the train’s track to one where only one person lies.

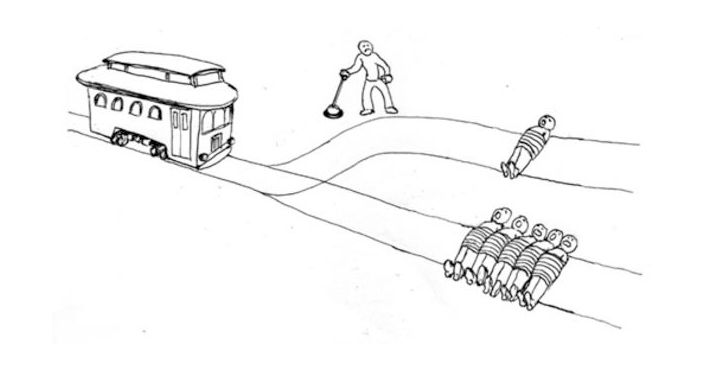


Figure 2: Train trolley dilemma  
Extracted from: https://nymag.com/intelligencer/2016/08/trolley-problem-meme-tumblr-philosophy.html

Challenges

When developing AI, there will be a number of ethical challenges that computer scientists will face. [Markel-4] defines these ethical challenges into 4 categories:

*Discrimination of people due to bias*. This refers to the unfairness provoked by the bias of either the training data or the algorithm employed. An example of this [Markel-5] occurred when Amazon’s recruitment system developed a bias against woman.

Then next challenge is to *ensure a system fulfils minimal conditions*, similar to the socially established sanitary conditions. The point is to prevent absurd situations, like a system which recommends lyrics for songs that does not create songs about beaches as it is phonetically close to an obscene word.

Thirdly, to *stop the usage of AI for unethical purposes*. That expresses the need to consider situations where users can explode a mechanism to deliver an unethical product, such as the recently scandal of Taylor Swift’s deepfake pictures, or potentially using Chat-GPT for obtaining links to pages that offer movies.

Finally, the last category is the *indiscriminate use of resources* to train large models, that is the environmental impact produced by the power consumption required for the vast number of operations carried out.

These scenarios will occur most times; meaning mutual consensus and responsible design is a must, and actions should be taken to prevent them.

Future Discussion

AI is something humans have not experienced yet. We are unaware of the potential risks of developing a machine that can be smarter than us, as we have never been in such context. Value alignment is, therefore a necessary measure to ensure that the actions carried out by AI corresponds to the actions humans would take.

To carry out that alignment, we must first analyse ourselves to determine are our values (and keep on doing so like the alignment training cycle described in [Markel-6]). By doing this task individually and discussing it with others, by the means of education, we can slowly build a more homogeneous ethic space where these values can have a specific shape.

References

[Markel-1] Dignum, V. (2020). Responsibility and artificial intelligence. *The oxford handbook of ethics of AI*, *4698*, 215.

[Markel-2] Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Vol. 2156). Cham: Springer.

[Markel-3] Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, *2*(1), 11.

[Markel-4] Baeza-Yates, R. (2022, February). Ethical challenges in AI. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (pp. 1-2).

[Markel-5] Mu, W. (2023). How Artificial Intelligence Affects Workforces: The Impact of Biased Recruitment and Job Displacement Risk. *Highlights in Business, Economics and Management*, *23*, 19-25.

[Markel-6] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

[Markel-7] Kinnier, R. T., Kernes, J. L., & Dautheribes, T. M. (2000). A short list of universal moral values. *Counseling and values*, *45*(1), 4-16.