ASSIGNEMENT 2

Responsible AI

Benedicto Costa, Markel [mabe9418@student.su.se](mailto:mabe9418@student.su.se)

De Givry, Joseph [jode8835@student.su.se](mailto:mabe9418@student.su.se)

Bhuiyan, Md Motaher Hossain [mdbh3795@student.su.se](mailto:mdbh3795@student.su.se)

Table of Contents

[Abstract 3](#_Toc158977003)

[1-Introduction to Responsible AI 3](#_Toc158977004)

[1.1 – RICE 3](#_Toc158977005)

[1.2 - Alignment (forward and backward for example) 3](#_Toc158977006)

[1.3 - Alignment cycle 4](#_Toc158977007)

[1.4- Types of learning 4](#_Toc158977008)

[2-Technical Challenges 4](#_Toc158977009)

[3-Ethical Challenges 5](#_Toc158977010)

[3.1 Ethics and values 5](#_Toc158977011)

[3.2 Challenges 6](#_Toc158977012)

[3.3 Future Discussion 6](#_Toc158977013)

[4-Solutions 7](#_Toc158977014)

[5-Conclusion 7](#_Toc158977015)

[References 8](#_Toc158977016)

# Abstract

This essay looks at the complex idea of Responsible AI. It highlights the need to make AI systems match up with human values and social norms by focusing on RICE (Robustness, Interpretability, Controllability, and Ethicality), different ways to align these systems (using both forward and backward methods), the cycle of alignment and types of learning. By reviewing detailed surveys and models of governance we talk about how technological progress and ethical considerations need to come together to make sure AI is used in a way that benefits society. The discussion points out how important it is to use approaches from different fields, mixing technical solutions with insights from the social sciences and governance models to deal with the challenges and possibilities AI technologies bring. Through this discussion we want to add to the ongoing conversation about making AI systems that are not just technically effective but also responsible to society and in line with values centred around humans.

# 1-Introduction to Responsible AI

This essay explores the multifaceted challenges of responsible Artificial Intelligence, plotting from technical, ethical and governance perspectives. Emphasizing the principles of Robustness, Interpretability, Controllability, and Ethicality AI (RICE) it delves into the complexities of AI alignment, including forward and backward alignment and the alignment cycle. Additionally, it examines various learning types in AI development such as supervised, unsupervised and reinforcement learning, highlighting their relevance to responsible AI. By synthesizing insights from recent literature and governance frameworks the essay aims to illuminate the complex interplay between technological innovation and ethical mandates. It proposes a multi-stakeholder approach to navigating the challenges of AI governance, advocating for a balanced integration of technical solutions and ethical considerations to foster AI that is both innovative and aligned with human values and social rules.

## – RICE

The RICE framework Robustness, Interpretability, Controllability, and Ethicality AI serves as a cornerstone for developing AI systems that adhere to ethical guidelines and social expectations. This framework emphasizes the importance of creating AI technologies that are not only technologically advanced but also ethical and inclusive, ensuring that benefits are widely distributed across society and do not aggravate existing inequalities​​​​.

## 1.2 - Alignment (forward and backward for example)

AI alignment focuses on ensuring that AI systems goals and behaviours are aligned with human values and ethical standards. Forward alignment involves designing AI systems that can predict and align with future human values and ethical considerations while backward alignment refers to the AI's ability to understand and incorporate historical human values and contexts. The challenge here lies in accurately interpreting and integrating a diverse range of human values into AI systems which is critical for ensuring that AI technologies act in ways that are beneficial and not detrimental to humanity​​.

## 1.3 - Alignment cycle

The alignment cycle is a continuous process of aligning AI systems with evolving human values and social conventions. This iterative cycle involves regular monitoring, evaluation, and updating of AI systems to reflect changes in societal values and norms. The dynamic nature of community standards necessitates a flexible and adaptive approach to AI alignment, underscoring the importance of ongoing engagement with stakeholders and the public in the AI development process​​.

## 1.4- Types of learning

Understanding the various types of learning in AI such as supervised, unsupervised and others is critical for implementing responsible AI. Each learning type has its impact on AI behaviour with supervised learning relying on labelled data for training. Unsupervised learning identifying patterns without labelled data and reinforcement learning adapting actions based on feedback to achieve specific goals. The challenge is to ensure that these learning processes are follow by ethical principles and do not lead to unintended consequences like reinforcing biases or violating privacy​​.

By embracing the RICE principles, focusing in on AI alignment and understanding the implications of various sorts of learning stakeholders can explore the complexities of AI governance and strive towards a future where AI technologies innovations are both creative and lined up with human qualities.

# 2-Technical Challenges

One of the technical challenges of responsible AI is the misalignment of AI systems. This misalignment could have major consequences such as unintended behaviours or losing control of humanity [3]. The AI system fails to align with the good human intentions causing multiple misalignments. A notable example is the misgeneralization, where the AI deviates from the intended goal that we established when training. These behaviours have different outcomes, the model can acquire knowledge about itself and react in a way that, unfortunately, will potentially touch humans or the world by being offensive or inappropriate.

Moreover, the challenges extend to broadly-scoped AI goals, if these goals are not done precisely, it can lead to manipulation over a large population. For instance, AI systems may force humans to take a high-pressure job, in order to achieve human happiness [4]. This example shows that it’s very important to align goals with ethical and societal norms, and the need to define clear objectives and goals.

Adding to all these challenges, the increase of accessible resources poses also difficulties. They could choose to provide false information on purpose, deceive users, and the AI model can be manipulated by malicious people to help them in their “malicious” objectives. This demonstrates that monitoring/regulating models is as important as the ethical guidelines especially with the increase of resources.

Furthermore, the natural behaviour of AI systems to seek power amplifies the risk. These systems try to take charge of both resources and people, necessitating a careful approach to avoid unintended consequences and ensure responsible use of AI technologies. Once again, this proves the need to continue research, ethical thinking and governance systems to have Responsible AI.

# 3-Ethical Challenges

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 [5]. 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.

## 3.1 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 [6]. 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 [7] proposes a model where 10 basic values are placed in a circle (Figure 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*).

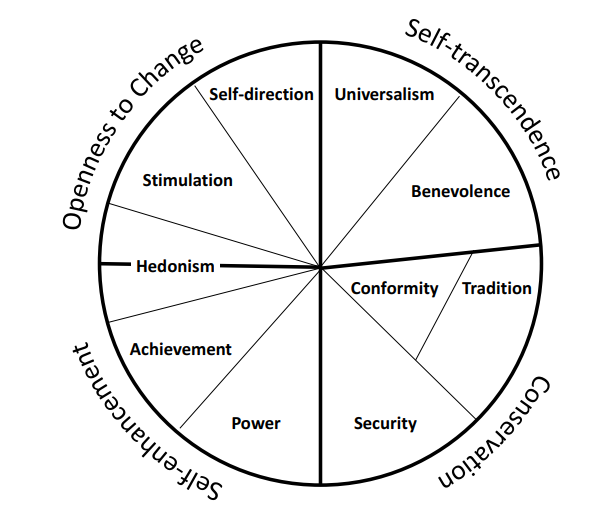


Figure 1: Schwarz value classification

[8] 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 [6]. 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.

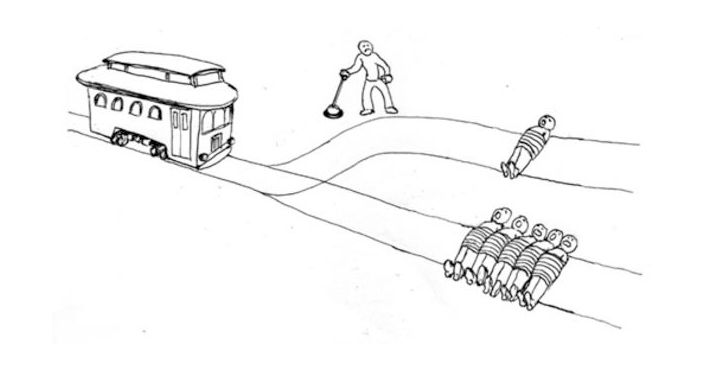


Figure 2: Train trolley dilemma  
Extracted from: https://nymag.com/intelligencer/2016/08/trolley-problem-meme-tumblr-philosophy.html

## 3.2 Challenges

When developing AI, there will be a number of ethical challenges that computer scientists will face. [9] 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 [10] 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.

## 3.3 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 [1]). 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.

# 4-Solutions

As we have seen before, the urgent need for responsible AI, a variety of solutions have emerged to ensure ethical and safe models. First, Dignum explains that we have to ‘take responsibility’ in development. Meaning we have to add “Ethic in Design” where the ethical values are strictly considered especially concerning its integration and replacement of traditional systems and social structures. “Ethic by Design” is the step further, it means that the model integrates ethical reasoning as part of his behaviour. “Ethic for Designer” focus on the importance of research integrity among researchers and manufacturers involved in designing, constructing, using, and managing AI systems

Human-friendly collaboration and co-creation in mixed human-AI settings are very important aspects of responsible AI to align with human value he has to learn, reason and plan with humans.

To avoid bias in the AI solutions, Corporate Social Responsibility (CSR) plays a major role. It drives bias out of the training data by hiring computer scientists, sociologists, anthropologists, legal scholars, and activists to develop strategies to develop more fair and accurate training data. These teams work to develop strategies to prevent biases, particularly those related to racism, sexism, ableism, homophobia, and xenophobia, in order to create a more equitable AI solution.

On a bigger scale, National and International AI Guidelines provide a framework for lawful, ethical, and robust AI systems. Obviously, they have to respect law and regulations, respect ethical principles and be technically robust, but also consider social implications as foundational principles.

The European Union created the AI Act, which aims to regulate AI use for safety, transparency, and non-discrimination. Leading to risk classification, ranging between unacceptable, high, and limited risks. In particular, limitations for the use of biometrics identification systems by law enforcement, bans on social scoring and manipulation of other vulnerabilities. This includes the right of consumers to launch complaints and receive meaningful explanations. If AI systems don’t respect the AI Act, they can be charged fines ranging from 35 million euro or 7% of global turnover to 7.5 million or 1.5% of turnover.

# 5-Conclusion

To conclude, responsible AI is very important in the future because it is all the guidelines of responsible AI that will allow its proper functioning. It is essential that AI respects our human values, ethics otherwise we will face the consequences (racism, discrimination, manipulation of humans by AI). But in today's society, the increase of resources, the fast improvement of technology makes it more difficult to handle it. Also translating human value to an AI model is tricky (e.g. misgeneralization).

# References

[1] Ji, et al. (2023). AI Alignment: A Comprehensive Survey.

[2] World Economic Forum (WEF) AI Governance Alliance. (2024).

[3] Nick Bostrom. 2014, Superintelligence: Paths, Dangers, Strategies

[4] Jacob Steinhardt. 2023. Emergent deception and emergent optimization.

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

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

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

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

[9] 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).

[10] 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.