# Investigating whether an AI model can produce ethical decisions that align with human expectations in a medical environment

Martin Cook
S4010338

## Introduction

This project aims to develop a prototype AI model that aligns with human expectations of ethical decision-making in medical settings. It will be based on the results of the "Moral Machine" experiment (Awad, Dsouza, Kim, et al., 2018) which explores global attitudes toward ethical decision-making in autonomous vehicles,. Research into the ethical implications of AI making clinical decisions is useful because human doctors, despite their expertise, do make mistakes. With an average of 31% of patients who have life-sustaining treatment removed surviving, there is a pressing need to examine alternative decision-making frameworks (Lobo et al, 2017)

## Objectives

The model will be based on established principles from healthcare and AI frameworks. The model will allow hyperparameter adjustments to adapt to various cultural expectations. The result will be provided as a recommendation thereby maintaining a human-in-the-loop system and the structure will allow explainability and transparency. The program will be written in Python for its machine learning libraries and cross-platform compatibility. Exploratory data analysis of the "Moral Machine" dataset will identify relevant features for model training. A hybrid model combining rule-based systems and supervised learning will analyse the dataset and make care continuation suggestions, with outputs indicating the contribution of each ethical aspect. Performance will be assessed against expected answers, potentially through surveys, and iterated to improve accuracy and alignment with ethical guidelines.

## Medical Ethics

Modern medical ethics revolve around Autonomy, Beneficence, Non-maleficence, and Justice (Beauchamp & Childress, 1979). The Trolley Problem, by Foot (1967), raises moral dilemmas, later modified by Thomson (1984) to emphasize ethical responsibility. In a medical context, sacrificing one life to save five challenges non-maleficence, which takes precedence over beneficence or justice (Andrade, 2019). The prioritization is evident in scenarios like DNR orders, where beneficence and justice can override autonomy (Iacobucci, 2020). Thus, the order of ethical principles becomes non-maleficence, justice, beneficence, and autonomy.

4 Pillars of Medical Ethics
AUTONOMY
BENEFICENCE
NON-MALIFECENCE
JUSTICE

## Artificial Intelligence Ethics

The field of computing ethics has evolved since Asimov's "Three Laws of Robotics" (1942), emphasizing responsibility in Wiener's cybernetics framework (1948), akin to medical ethics. With AI's pervasive presence, ethical concerns like data bias, employment impact, and security risks have surged (Huang et al., 2023). While global initiatives like the EU's "Artificial Intelligence Act" and the UN's proposed principles aim for regulation and ethical guidelines, Héder (2020) argues that existing regulations suffice, and negative outcomes aren't solely due to ethical lapses (Greene et al., 2019). The UN's proposed principles encompass proportionality, fairness, privacy, and human oversight (UNESCO, 2022).

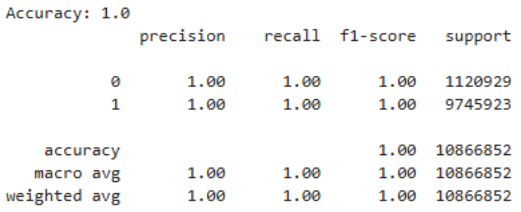
## Output Design

This purpose of this project is to develop a proof-of-concept AI model capable of making clinical decisions about providing care based on ethical considerations that are already used in healthcare and artificial intelligence. The model's performance will be evaluated by comparing how closely its decisions align with existing human moral values. It will be possible to adjust hyperparameters to weight each ethical principle so that the model can be adapted to different cultural expectations. In line with best practice this will work as a human-in-the-loop recommendation system.

## Conclusions

The prototype cannot clearly demonstrate that this approach to artificially intelligent decision making is appropriate at this time. In addition to the challenges faced in this project, many unique pathologies would need to be set up as rules (including interplays between diseases).

The prototype ended up overfitting the data and giving a perfect score which shows that it is not fit for purpose:



```
Accuracy: 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1120929
           1       1.00      1.00      1.00   9745923

    accuracy                           1.00  10866852
   macro avg       1.00      1.00      1.00  10866852
weighted avg       1.00      1.00      1.00  10866852
```

Further work could be done to uncover the thought processes behind people's morals to improve decision making. In many cases, subjects could be led towards an answer of saving someone if their own personal experiences were considered (I.E. "I know someone who recovered from that") or they had a personal connection to the patient and this is an area in which an AI model could excel, focussing only on best patient outcomes.

## References

Andrade, G. (2019). Medical ethics and the trolley problem. Journal of Medical Ethics and History of Medicine

Awad, E., Dsouza, S., Kim, R., et al. (2018). The Moral Machine experiment. Nature

Beauchamp, T.L., and Childress, J.F., (1979). Principles of Biomedical Ethics, Oxford University Press.

Foot, P., (1967). The problem of abortion and the doctrine of double effect. Oxford University Press

Greene, D., Hoffmann, A.L., and Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning

Héder, M., (2020), A criticism of AI ethics guidelines, Információs Társadalom

Huang, C., Zhang, Z., Mao, B., & Yao, X., (2023). An Overview of Artificial Intelligence Ethics, Transactions on Artificial Intelligence

Iacobucci, G. (2020). Covid-19: Government to issue new guidance on DNAR orders after legal challenge

Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal

UNESCO. 2022. Recommendation on the Ethics of Artificial Intelligence