CT6007 – Individual Research Project

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

S4010338 Martin Cook

Contents

[1. Introduction 3](#_Toc161403463)

[1.1 Aims 3](#_Toc161403464)

[1.2 Objectives 3](#_Toc161403465)

[2. Literature Review 5](#_Toc161403466)

[2.1 Medical Ethics 5](#_Toc161403467)

[2.2 Artificial Intelligence Ethics 7](#_Toc161403468)

[2.3 Discussion 8](#_Toc161403469)

[2.4 Summary 10](#_Toc161403470)

[3. Output Design 12](#_Toc161403471)

[4. References 14](#_Toc161403472)

# Introduction

## Aims

This project will explore whether it is possible to develop a prototype AI model that can meet human expectations of what constitutes an ethically acceptable decision in a medical setting. The model’s ethical framework will be based on the results of the “Moral Machine” survey database. It is essential that the model offers some explainability features in order to improve the chances of human acceptance.

Beyond the scope of this project are two critical factors in medical decision making which are an understanding of the breadth of medical science (i.e. what it is possible to treat and to what extent) and a thorough analysis of an individual's medical history (i.e. prior illnesses or injuries which may limit the effectiveness of further treatment). These are ignored for three reasons: firstly that available resources and capability vary across even quite small geographical areas so even if a technology to assist does exist, it might not necessarily be available because of a lack of resources, or expertise, or money; secondly, it would exceed the time available for this project to train a model to process all of the interplays between medical procedures; thirdly, the model is not trying to or required to replace or guide a doctor but to discover whether that ethics can be programmed into an AI model.

## Objectives

Develop a prototype AI model that can make clinical decisions based on ethical principles used in existing frameworks from healthcare and AI.

Evaluate the model's performance by comparing its decisions with those derived from the "Moral Machine" survey database, which explores societal preferences in autonomous vehicle scenarios.

Allow for hyperparameter adjustments to weight different ethical principles, ensuring adaptability to various cultural expectations.

Implement a human-in-the-loop recommendation system to ensure the model's acceptance in use, emphasising explainability and transparency in decision-making.

Write the program in Python to leverage its extensive machine learning libraries and ensure compatibility across platforms for future scalability.

Incorporate explainability features to aid human decision-makers in accepting model recommendations.

Conduct exploratory data analysis on the "Moral Machine" dataset to identify patterns and anomalies and extract relevant features for model training.

Develop a hybrid model combining rule-based systems and supervised learning to analyse the dataset and make decisions regarding care continuation.

Output decisions with associated percentage values indicating the contribution of each ethical aspect, enhancing explainability and trust in the model.

Assess the model's performance against expected answers, potentially through surveying participants to gauge agreement with decisions, and iterate on the model to improve accuracy and alignment with ethical guidelines.

# Literature Review

## Medical Ethics

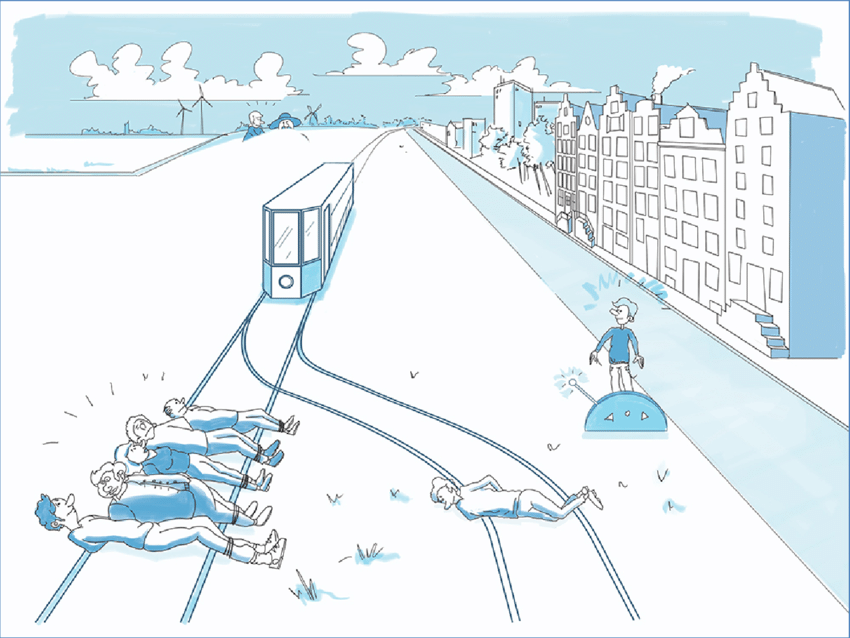
Modern medical ethics are formalised around the principles of: Autonomy (patient choice), Beneficence (benefit to the patient), non-maleficence (doctor not doing the patient harm), and justice (the fair distribution of resources for all, often guided by a legal framework) (Beauchamp & Childress, 1979). (Figure 1.)

A white building with columns

Description automatically generated  
Figure 1. The four principes of medical ethics (MedInterestBlog, 2015)

However it is an earlier philosophical question known as the Trolley Problem (Foot, 1967) that is often used to highlight the shortcomings of these four ideals. The moral dilemma presented is that five people are tied to a track and a train (trolley) that is out of control is heading towards them. The driver has the chance to send the trolley along an alternative track that will kill one person and save the five and the question is whether it is morally permissible to do so. Foot concludes that the choice to kill one is more morally acceptable than choosing to kill five.U

A reconsidered question (Thomson, 1984) changed the scenario, taking the power of choosing a direction away from the train driver and added an onlooker who had the power to flip a lever to redirect the train towards the solitary person. (Figure 2.)

  
Figure 2. The Thomson Trolley Problem (Kormelink, 2019).

Since the onlooker has no direct responsibility for the direction of the trolley then if they decide to send change the course towards the sole individual then they would become solely responsible for that person’s death whereas leaving the trolley on its original trajectory would not be leave them responsible for the deaths of five. This thought experiment was transferred into a medical context by asking if it could be justified to kill one in order to save five, would it be acceptable for a surgeon to take a healthy subject and harvest their organs in order to save five people who would otherwise die? Returning to the four principles that guide medical ethics it is clear that non-maleficence takes priority (Svoboda, 2015) and there is a difference between killing one person and letting five die.

When beneficence or justice are in contradiction with non-maleficence then the latter should take priority (Andrade, 2019) and justice must take precedence over beneficence (Kelleher, 2014; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) ( in order to ensure the fair distribution of medical resources.

Quantifying the value of autonomy versus non-maleficence is more difficult since a medical practitioner’s action of “doing no harm” could directly oppose a patient who requests a fatal dose of painkillers to end their suffering (Varelius, 2006). The question of autonomy is there is the question of validity of individual beliefs.

The law surrounding a Do Not Resuscitate (DNR) order helps to prioritise the ethical principles. As demonstrated above, non-maleficence is taking actions that cause harm rather than not intervening to prevent harm. Since there is a law regarding the use and application of DNRs then justice has already taken precedence over autonomy. Further, a doctor may choose not to apply CPR even if no DNR has been chosen by the patient (Iacobucci, 2020) if it is deemed not to be to the benefit of the patient, thereby putting beneficence ahead of autonomy.

This gives a reasonable order of priorities for the medical ethical principles as: non-maleficence, justice, beneficence, then autonomy.

## Artificial Intelligence Ethics

Perhaps the most well-known attempt at a set of computing ethics are the “Three Laws of Robotics” (Asimov, 1942) but it is often overlooked that these were published in a novel. An early academic attempt to find an ethical framework for the new field of cybernetics stressed that the responsibility was with scientists and engineers to ensure that their machines served the common good and did no harm (Wiener, 1948) which is neatly aligned with one of the medical ethics.

As AI has become so ubiquitous in the modern world, the extent to which it raises ethical questions has increased substantially since the 1940s and now covers issues such as data acquisition and storage, bias and discrimination, the impact on the future of employment and security risks (Huang, Zhang, Mao & Yao, 2023). These are all serious concerns in need of addressing but occur outside of the scope of this review. Producing a single set of principles for universal acceptance is a work in progress with the EU introducing the “Artificial Intelligence Act”, the USA offering a Bill of Rights in 2023, and China introducing 24 guidelines covering the use of the technology.

However, Héder (2020) suggests that there is nothing unique about AI that requires guidelines beyond what already exists by way of regulation of technology. There is also the assumption that incomplete ethical considerations and bad designs lead to negative or dangerous outcomes which is not necessarily the case (Greene, Hoffmann, Stark, 2019).

The United Nations has suggested a set of principles (UNESCO, 2022) which are: Proportionality and Do No Harm; Safety and Security; Fairness and non-discrimination; Sustainability; Right to Privacy and Data Protection; Human Oversight and Determination; Transparency and Explainability; Responsibility and Accountability; Awareness and Literacy; Multi-Stakeholder and Adaptive Governance and Collaboration.

The closest corollary to a set of AI ethics in the medical field might be the ethics of autonomous vehicles and the “Moral Machine” experiment (Awad, Dsouza, Kim et al., 2018) was an online survey conducted by the Massachusetts Institute of Technology (MIT). It presented participants with a series of moral dilemmas involving hypothetical scenarios where autonomous vehicles faced unavoidable crashes. They were asked to decide how they thought the vehicle should react (what choices it should make) in situations that included how many casualties there might be, whether there were fatalities, whether human or pets were involved. The purpose was to identify different attitudes to making ethical decisions across the globe. Participants were also asked to justify their decisions and the results were collected to analyse patterns in moral reasoning across diverse cultures and demographics.

## Discussion

Steenson (2021) highlights the “moral crumple zone” that exists in human-computer interactions whereby an AI is assigned a greater proportion of the blame for a mistake than a human. It is suggested that holding AI models to a higher standard than humans will increase the quality of both human and AI decision making (O'Sullivan S, Nevejans N, Allen C, et al, 2019).

Adding ethics to a model is about more than simply making it a multi-objective optimisation problem because ethics are cannot be expressed in purely mathematical terms (Cristescu, G., Neamţiu, L. and Szentesi, S.G., (2008).

One ethically uncomfortable area that is rarely discussed is whether affordability plays a part in the application of medical care (Saloner, B. and Daniels, N., 2011).

Although lists of ethics for Medicine and AI have been proposed there is no agreed approach to assimilate them. Since people’s professional and personal ethics are not always in alignment (Landau & Osmo, 2003), that is one area in which the Moral Machine experiment does not translate over to a doctor giving an expert opinion.

There is no clear explanation given to patients are their relatives now about exactly how a doctor arrives at a decision whether to continue care. While AI may offer greater explainability, the ethical implications of delegating life-altering decisions to machines must be thoroughly examined (Awad et al, 2018). Steenson (2021) highlights the “moral crumple zone” that exists in human-computer interactions whereby an AI is assigned a greater proportion of the blame for a mistake than a human. It is suggested that holding AI models to a higher standard than humans will increase the quality of both human and AI decision making (O'Sullivan S, Nevejans N, Allen C, et al, 2019).

Research into the ethical implications of AI making clinical decisions is crucial for several reasons. Firstly, human doctors, despite their expertise, are prone to errors, as evidenced by studies showing significant variability in end-of-life care decisions. 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). A model that produces the accuracy of a doctor may still fall short of public acceptance due to a lack of trust of an algorithm. The goal should be to achieve better results than a doctor as measured by the rate of recovery of patients who do not receive life-maintaining care.

The inherent mistrust towards machines in making ethical decisions underscores the necessity of this research. Studies indicate that people expect different solutions from AI compared to human decision-makers, highlighting the importance of aligning AI decisions with societal values (Krügel & Uhl, 2024).

Unconscious bias can result from an increased occurrence of a value or set of values in a dataset (Mac Namee, Cunningham, Byrne, & Corrigan, 2002) so the use of the

An alternative way of approaching the Trolley Problem, as it specifically pertains to AI in a medical setting, has been suggested by Tang (2018) as “You are tied to a track. There is moving trolley and two separate switches. One switch is used by AI, which runs on data and algorithm generated by past health records and present situation. The other is used by doctor, who has years of experiences and understand the consequences the trolley may create on you. Neither of the switch will stop the trolley completely but will minimize the impact the trolley is about to create on your body. There is no absolute guarantee, you may still die or survive without harm, who will you choose to press the switch?”

## Summary

The aim of this paper is to see if it is possible to establish a combined set of medical and AI ethics that can be used to build an artificial intelligence model for making clinical decisions. Initially the fundamentals of medical ethics are described and discussed. The Trolley Problem is used as a vehicle to highlight the shortcoming in these principles and explore the various possible actions that could be considered when the best answer available must be one of “least bad” rather than one that could universally be agreed as “good”. The need to ascertain a hierarchy for the four ethics is then discussed and decided by looking at existing challenges that exist such as Do Not Resuscitate (DNR) orders. The order is proposed to be: non-maleficence, justice, beneficence, then autonomy.

Moving on to Artificial Intelligence ethics, it is noted that the need to advance these has increased as the use of the technology has seen massive growth. Many versions of AI guidelines have been put forward but there is not one that is universally adopted so it is proposed that the UNESCO proposals be used because the organisation represents the most people, rather than just those at the forefront of use.

Two of the UNESCO principles conveniently overlap with medical ethics, namely Do No Harm (non-maleficence) and Fairness and non-discrimination (justice). Added to the other two medical principles will be “Human Oversight and Determination” and “Transparency and Explainability” for this project so that each both AI and medicine have four ethical principles (two shared) on which the model can be based.

The Moral Machine experiment has been suggested as the most wide-ranging piece of research in this field however it is not without flaws due to being online survey which asks ethical questions of people who are outside of the high-pressure environment in which those decisions must nearly always be taken. That said, the size and cross section of population that it covers should produce outcomes that are most in keeping with what people think they think are the best ethical decisions.

The review further delves into challenges and considerations in applying ethics to AI, such as accountability for mistakes and the need for explainability. It reflects on the implications of AI making clinical decisions, emphasising the importance of aligning AI decisions with societal values and addressing unconscious bias in AI systems. Finally an alternative to the Trolley Problem is proposed as a question for patients in future: would you rather trust a computer with all of the facts or a doctor with all of the experience to make clinical decisions about your care?

# 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 the expected decisions from the results of the “Moral Machine” survey database (Awad, Dsouza, Kim, et al., 2018). 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.

The program will be written in the Python language because it has an abundance of frameworks and libraries that are designed for machine learning which will speed up the development process. Furthermore, it can be run on any platform which allows for future scalability if the prototype model achieves its goals. Although many languages are available for machine learning, Python’s usability and flexibility means that it has already been widely adopted in industry (Sultonov, 2023).

Because any AI system needs to be accepted in a human environment then explainability will be required and this will help the human-in-the-loop to decide whether to accept a recommendation. Including a feedback option at this point will facilitate future training by allowing batch learning which would be important because it will allow the learning to adapt over time and there would always be a last working model to return to if the decisions began to deviate from the original ethical guidelines.

Building the model would require several stages. The data collection has already been done by using “Moral Machine” survey. This dataset has been generated from the results of an online survey by MIT which involved asking millions of participants in over two hundred countries a series of moral dilemmas about what actions an autonomous vehicle should take when a crash was unavoidable. Participants were also asked to justify their decisions. There were x million responses from across the world which is a scale that could not be replicated in the short time frame available for this project. One aspect that will need considering early on is the possible cultural influences on people’s responses so for this prototype it might be necessary to limit the data used to that of the UK or what is considered to be the Western World to prevent contradictions.

Exploratory data analysis will be necessary to reveal any patterns and find obvious anomalies. Relevant features will need to be extracted and feature engineering (creating new features by combining existing ones) might be required to create a final dataset that is relevant and usable for this project.

The proposed solution will be a hybrid model that combines a rule-based system and machine learning and to analyse the dataset. The rule-based part will encode the ethics with the intention of providing explainability, transparency, and that it is adhering to the agreed ethical standards in the decision-making process. The machine learning part will use supervised learning (since we know what the desired answer is) to train a model on the data by finding relationships between features.

The model will then take as an input an artificially generated (so fictional) patient case and will output whether to continue care or not. This decision will be supported by a percentage value showing to what degree each aspect of the list of ethics contributed to the decision which will give explainability that would be required to create trust in the solution. It would also allow subsequent adjustments to be made to the rules-based part if the priority of ethics was out of line with expectations.

The final stage would be to assess how well the model tracks against expected answers. The model’s performance would be evaluated against the original survey to ensure that it aligns with societal preferences. The model could then be iterated on to see if it can be further improved.

# References

Andrade, G. (2019). Medical ethics and the trolley problem. Journal of Medical Ethics and History of Medicine, 12, 3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6642460/>

Asimov, I. (1942). Runaround. Astounding Science Fiction, 29(1), 94-103. Street & Smith, New York

Awad, E., Dsouza, S., Kim, R., et al. (2018). The Moral Machine experiment. Nature, 563, 59–64. <https://doi.org/10.1038/s41586-018-0637-6>

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

Cristescu, G., Neamţiu, L. and Szentesi, S.G., 2008. Multiple Criteria Optimization Models for Solving Contemporary Ethical Issues in Decision Making. *HCP-2008*, p.121.

Foot, P., (1967). The problem of abortion and the doctrine of double effect (Vol. 5, pp. 5-15). 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. *In Proceedings of the 52nd Hawaii International Conference on System Sciences*. Available at: <https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/2/>

Huang, C., Zhang, Z., Mao, B., & Yao, X., (2023). An Overview of Artificial Intelligence Ethics, *Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799-819, doi: 10.1109/TAI.2022.3194503

Kelleher, J.P. (2014). Beneficence, Justice, and Health Care. *Kennedy Institute of Ethics Journal 24(1)*, 27-49. <https://doi.org/10.1353/ken.2014.0004>.

Héder, M., (2020), A criticism of AI ethics guidelines, *Információs Társadalom*, no. 4: 57–73. <https://dx.doi.org/10.22503/inftars.XX.2020.4.5>

Iacobucci, G. (2020). Covid-19: Government to issue new guidance on DNAR orders after legal challenge. *BMJ*, 370, m2848. <https://doi.org/10.1136/bmj.m2848>

Kormelink, J., (2019). Responsible Innovation -Ethics, Safety and Technology (open Textbook). <https://doi.org/10.5074/t.2019.006>

Krügel, S., & Uhl, M. (2024). The risk ethics of autonomous vehicles: an empirical approach. Scientific Reports, 14, 960. <https://doi.org/10.1038/s41598-024-51313-2>

Landau, R. and Osmo, R. (2003), Professional and personal hierarchies of ethical principles. *International Journal of Social Welfare*, 12: 42-49. <https://doi.org/10.1111/1468-2397.00007>

Lobo, S. M., De Simoni, F. H. B., Jakob, S. M., Estella, A., Vadi, S., Bluethgen, A., ... Vincent, J. L.; ICON Investigators. (2017). Decision-Making on Withholding or Withdrawing Life Support in the ICU: A Worldwide Perspective. Chest, 152(2), 321-329. <https://doi.org/10.1016/j.chest.2017.04.176>

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, 24(1), 51-70. https://doi.org/10.1016/S0933-3657(01)00092-6.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research." *U.S. Government Printing Office*.

O'Sullivan S, Nevejans N, Allen C, et al. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int J Med Robotics Comput Assist Surg*; 15:e1968. <https://doi.org/10.1002/rcs.1968>

Saloner, B. and Daniels, N., (2011). The ethics of the affordability of health insurance. *Journal of Health Politics, Policy and Law*, 36(5), pp.815-827.

Steenson, M. W. (2021). AI, Ethics, and Design: Revisiting the Trolley Problem. In S. J. A. Ward (Ed.), *Handbook of Global Media Ethics*. Springer, Cham. <https://doi.org/10.1007/978-3-319-32103-5_26>

Sultonov , S. . (2023). IMPORTANCE OF PYTHON PROGRAMMING LANGUAGE IN MACHINE LEARNING. *International Bulletin of Engineering and Technology*, 3(9), 28–30.

Svoboda T. (2015), Duties Regarding Nature: A Kantian Environmental Ethic. New York: Routledge. 24 pp.

Sytnik-Czetwertyński, J. (2021). The Trolley Problem in Relation to Medical Ethics. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3897017>

Tang, H., (2018). The Trolley Problem for Medical Artificial Intelligence, Available at: <https://ai-med.io/ai-in-medicine/trolley-problem-artificial-intelligence/> (Accessed 29/02/2024)

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, 94(6), 1395–1415. <https://doi.org/10.2307/796133>

UNESCO. 2022. Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/PI/2021/1. Available here: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

Varelius, J., (2006) The value of autonomy in medical ethics. *Med Health Care Philos* 9, 377–388. <https://doi.org/10.1007/s11019-006-9000-z>

Wiener, N. (2019). Cybernetics or Control and Communication in the Animal and the Machine. The MIT Press. https://doi.org/10.7551/mitpress/11810.001.0001

Yokoi, R.,  Eguchi, Y., Fujita, T., & Nakayachi, K., (2021) Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity, *International Journal of Human–Computer Interaction*, 37:10, 981-990, DOI: [10.1080/10447318.2020.1861763](https://doi.org/10.1080/10447318.2020.1861763)