# Ethics Coursework

Jamie Couchman (jhc21)

March 28, 2022

## 1 Questions

### 1.1 What are the different types Artificial Moral Agents, and what makes them different?

If machine ethics are taken to concern moral agents in some significant way i.e. programs that can act or make decisions in a 'moral' way, then these agents are classified as 'Artificial Moral Agents' The three different types of Artificial Moral Agents are: Amoral Agents, Implicit Moral Agents and Explicit Moral Agents. Amoral agents have no ethical considerations and cannot be considered to be pernicious, although their actions can have consequences. An example of this would be a calculator that adds a negative sign to all calculations; this could mean a building falls down or an engine explode, though the calculator has no intent. Implicit Moral Agents have ethical considerations built in such as safety or security, but do not identify these considerations themselves. Rather, they are ethical by design; an example would be an ATM or car reversing warning. An Explicit Moral Agent identifies and acts upon ethical information about a range of situations and make sensitive decisions about how to act. When ethical principles are in conflict, these agents have the ability to find reasonable solutions.

## 1.2 What are some of the issues or challenges facing the prospect of building ethical AI?

There are two contrasting approaches to building ethical AI agents. The first is a top-down approach starting with some high level principles of morality built in. This approach relies on first formulating the necessary principles and then designing the agent with these in mind. This clearly causes a plethora of issues for the designer. What moral philosophy should be used? Philosophers have debated over the merits of various paradigms for millennia, should a robot be Kantian or Utilitarian in its view of its environment? There is also a concern about human nature itself, it could be that humans have a weak and animalistic nature that must be overcome for us to act ethically. Robots should be built to overcome this base nature and therefore they might need a new code of conduct altogether. Further, the absence of any evident algorithm for moral decision making is a barrier for ethical robots, we cannot formulate these highly nuanced principles in any obvious way. It is another challenge altogether to test and evaluate the results of this top down encoding, a simple binary cross-entropy loss calculation will not cut it.

For the second approach, the ’bottom up’ build approach, the agent begins with no moral framework, instead learning it’s ethics from it’s environment. This approach comes with its own set of problems. Reinforcement learning has made huge strides in the capabilities of machines to maximise their rewards. When designing these rewards is based on actions (e.g. Atari game score) this is possible, but it becomes much harder to evaluate a reward for being ‘good’, the definition of which is not clear. Relying only on actions does not give a true sense of ‘good’, although consequentialists might rebut this. The supervised agent needs many examples to train from, which is another difficulty arising from this inability to evaluate ‘good’. The self supervised agent suffers from a lack of explainability in how it makes decision making in its development, and this can lead to it being corrupted; Googles word2vec is a good example of such behaviour. This algorithm produced biased stereotypes in word embeddings such as doctor - man = nurse. This model clearly needs more careful consideration of the ethics involved, which must be based on some normative moral theories. The bottom up approach cannot therefore be sufficient in itself.

## 1.3 Aristotelian Virtue Ethics, what is ‘Eudaimonia’?

The idea of ‘eudaimonia’, is standardly translated as ‘happiness’, ‘flourishing’ or occasionally “well-being”. Specifically this is flourishing possibly only for rational beings, and happiness that is not subjective to any individual but rather a higher order, virtue laden concept of happiness. One could be happy in the more simplistic sense, and mistakenly so if they live well *only* in physical pleasure or luxury. However, they will not be happy in the true sense of one who’s life is ‘eudaimon’[8]. In fact virtue ethicists claim that a human life dedicated to the pursuit of physical pleasures is a life wasted. Aristotle argues we choose this for itself and not for the sake of anything else.

## 1.4 What is virtue (Areté), according to Aristotle, and how is it gained?

In Aristoteleian Virtue Ethics, arête is the concept thats translates to excellence or virtue, and is treated as necessary for ‘eudaimonia’. A virtue is an excellent trait of ones character that they possess deep down and influences how they choose to feel, desire and act in certain characteristic ways. It can be described as a mindset that takes into account a distinctive range of considerations when making an action. Arete is relative to something and can be considered the highest quality state it can reach, judged on the proper function. In the example of an honest person, they cannot be considered virtuous just because ‘honesty is the best policy’ or they do not cheat. Virtue can be achieved only if one acts honestly because to do otherwise would be dishonest, they tell the truth because it is the truth and place equal weight on the importance of not lying. In order to be truly virtuous, one takes the action they do with no internal struggle against the contrary, with no temptation to do otherwise.

## 1.5 What is the 'Greatest Happiness Principle' and what is its role in Classical Utilitarianism?

The 'Greatest Happiness Principle' is the utilitarian concept that the world's utility is maximised by considering the 'greatest happiness of the greatest number' devised by Jeremy Bentham in 1863 [6]. By happiness, pleasure and the absence of pain is intended, and by unhappiness, pain and the privation of pleasure are intended. Actions in a broader sense are right in proportion if they tend towards promoting happiness of the greatest number and wrong if they promote the reverse, unhappiness. The definitions of pleasure and pain are nuanced, however this theory of life is grounded in the sense that pleasure and the absence of pain are desirable and that these desirable things are the *only* things desirable as ends. All desirable things are desirable simply because they promote pleasure and remove pain. This theory is compatible with the idea that pleasure is a wide ranging spectrum with some pleasures being more desirable and valuable than others. No rational human being would trade their life for that of a pampered Wagyu cow despite the physical comfort of said cow's environment; a being of higher faculties requires more than this to take true pleasure from life. The utilitarian standard is not the happiness of the agent but the greatest amount of happiness combined. The existence of a world that can be secured for mankind, exempt as far from possible from pain and as rich as possible with regards to pleasure, is the true measure of utilitarian greatest happiness.

## 1.6 How are 'act' and 'rule' Utilitarianism different?

There is a clear divide between two branches of utilitarianism. 'Act' utilitarians consider only the consequences of the single act while rule utilitarians consider consequences of the result of following a certain code of conduct. The 'act' utilitarian might consider the best action to maximise happiness in a certain situation may be to lie thereby minimising pain. An example is a doctor telling a patient they have 2 months to live is useful whereby the 'act' utilitarian chooses to lie. The family experiences less pain, the doctor experiences less pain and the utilitarian goal is maximised.

The ‘rule’ utilitarian might take a very different approach, considering the long term implications of telling this lie. More patients may lose trust in this doctor or doctors in general; there is far more harm in losing this confidence when aiming for the overall goal of minimising pain. ‘Act’ utilitarians believe the right action is the one that yields more utility while ‘rule’ utilitarians place more stress on moral rules. A specific action is justified if it conforms to an ethical rule, this rule can be included in the code of conduct if it creates more utility than other possible rules. In other words, both types adhere to maximising utility with the key difference that the ‘act’ utilitarians apply the principle of maximal happiness directly to evaluating individual actions. The ‘rule’ utilitarians apply it to a set of rules.

‘Act’ utilitarianism can be considered utilitarianism in the purest sense, if we maximise utility of every action that we take, then the total utility of all those actions combined is the highest possible yield of utility. Taking this approach has its drawbacks as it often leads to answers that are obviously morally wrong. Take this example:

> *‘If a judge can prevent riots that will cause many deaths only by convicting an innocent person of a crime and imposing a severe punishment on that person, act utilitarianism implies that the judge should convict and punish the innocent person.’*[12]

This is clearly not the ‘right’ approach. The ‘rule’ utilitarian goes some way to answer the flaw of the act utilitarianism by insisting that the role of the criminal justice official is to have authority and trust over the public; this is undermined if they punish the innocent.

## 1.7 In Kant’s view, why does the Categorical Imperative apply to all agents?

The fundamental principle of the categorical imperative (CI) is the law of autonomous will, that a rational agent must be regarded as autonomous, free and the author of any law that binds it [9]. Kant argued that the heteronomy (influence from outside the subject) can yield only hypothetical imperatives conditional on the end goal and whoever sets these goals. This cannot hold for everyone by design, therefore there must be a self law-giving autonomy that binds us. The CI manifests in four formulations that can apply to all agents unconditionally. One of the formulations will always be able to tell you what is wrong or right. One can take a maxim and check its compatibility with the formulations to determine whether it contradicts or not. The main strand is that other rational beings should never be treated as tools but rather that they also have reason, grounded in something universal. The CI is the unconditional, objective and rationally essential principle that we must follow despite our base animalistic urges. The CI is essentially a principle of practical rationality, through which all moral requirements are satisfied. On the reverse, any immoral actions are also irrational as these violate this CI. In a sense it is the principle that *means* matter as much as *ends*. One should do their best to their ability, not acting immorally to achieve goals by cheating or cutting corners. Every agent has present in them this self governing reason and therefore they demand an equal level of respect. This philosophy applies to all.

## 1.8 In Kantian Deontology, how are 'Perfect' and 'Imperfect' duties different?

According to Kant if you take a maxim you are able to determine that there is no contradiction in conception or contradiction in will, this was expanded upon in section 1.7. For example if your maxim is that 'it is permissible to steal', logically speaking this cannot make sense. In a world where everyone would take this action, everyone would take from everyone else and order breaks down; there is a contradiction in conception. Therefore the reverse case 'one should never steal' can be defined as a Perfect Duty, a duty that should never be violated as its consequences are inherently morally wrong contradicting the categorical imperative of all agents (you shouldn't treat other beings as means but as ends themselves). These duties take the form 'one must never (or always) X to the fullest extent possible in Y' and there is no exception in favour of inclinations of this kind.

In contrast, Imperfect Duties are a direct consequence of a contradiction in will. If you take the maxim 'one can never help others', in a world where nobody helps anyone, order would not completely break down but a rational being would not want to live in such a world; we know it would not be a good one. Therefore an example of an Imperfect Duty would be the reverse ‘one must help others’, this may be subordinated to perfect duties but not to any inclinations.

# 2 Essay - Descriptive vs Normative Moral AI

*13. Descriptive vs. Normative approaches to building moral AI - should an artificial agent reflect the actions that human agents would actually perform in the relevant situation (Descriptive) or should it reflect the actions that ought to be performed, according to a set of rules, norms, or otherwise agreed upon course of action (Normative)?*

Concerns about how machines make moral decisions are not merely in the realm of science fiction but a major challenge facing current artificial intelligence (AI) research. Autonomous vehicles are sophisticated enough to confront the moral dilemma of dividing up risk between different stakeholders in the road in the event of an accident. MIT's Moral Machine Experiment [3] documents the wide-ranging individual variations in preferences that humans show when faced with these ethical conundrums; often they cannot be solved by simple normative ethical principles alone. 'Inverse reinforcement learning' (IRL) [13] has pushed performance of AI to new frontiers through observing human behaviour and learning what reward signal is being optimised; parallels with a descriptive moral AI agent can be drawn. Instead of setting a clear set of rules that philosophers have debated over for millennia, perhaps inferring human behaviour is safer than exploring the inevitable loopholes that come built into a normative ethical system. This essay aims to evaluate the benefits of both the normative and descriptive approach and propose a hybrid that may go some way to build safe, moral AI.

The normative approach to building moral AI is fraught with issues in defining a coherent set of goals that ensure the outcomes of machines actions can be considered 'good' ones. The difficulty we face is that if we put the wrong objective into this machine then we create a conflict that we are likely to lose. Take, for example, Asimov's laws of robotics [2]:

> *The first law is that a robot shall not harm a human, or by inaction allow a human to come to harm. The second law is that a robot shall obey any instruction given to it by*

> *a human, and the third law is that a robot shall avoid actions or situations that could cause it to come to harm itself.*

The first rule takes preference over those below it and the second takes priority over the third. These laws are a literary device designed to exploit the difficulty in designing morality into AI, however, they act as a useful thought experiment on the flaws associated with such a design. Almost all AI systems have an objective of maximising human preferences, but as soon as the agent's actions affect more than one person, the waters become muddied. The agent cannot simply adhere to the preferences of one human by, for example, crashing the university submission site to buy their human extra time on a coursework deadline. This does not violate any of Asimov's laws but is clearly morally dubious.

Does then our agent become a utilitarian or a virtue ethicist? The prospect of an AI with any serious power abiding by the 'Greatest Happiness Principle'[6] conjures up dystopian images of HAL 9000 [10] killing the astronauts onboard the ship in order to complete the mission. The motives behind taking these actions could be to maximise the happiness of humans further down the line; these are the issues that political theorists and philosophers have been grappling with for thousands of years. In the context of AI, many researchers working on topics in relation to rational agency consider the concept of utility maximisation a perfectly suitable characteristic for evaluating rationality.

This leads to the familiar pitfalls of the utilitarian school of thought, but never before have the stakes been so high as we approach allowing machines to choose who should live and who should die without real-time supervision by humans. The 'Moral Machine' experiment drew stark attention to cultural differences in preferences. This was achieved via a thought experiment whereby a car has the opportunity to redirect into a barrier to avoid three elderly road crossers, killing the driver and two young passengers. Preferences based on gender or social status vary considerably across 200 countries, reflecting underlying societal bias and a lack of consensus. However, there were three strong preferences to consider when approaching universal machine ethics: sparing human lives, sparing more lives and sparing young lives. Experiments such as this one, based on evaluating broad points of agreement, would be a good place to start in attempting to reach a consensus on the normative rules of a moral agent. However, in the real world, the environment an agent finds itself in is dynamic, stochastic, continuous and partially observable. Thus strict codes of conduct will never be sufficient in themselves.

The descriptive approach to building AI systems has been a critical advancement and many consider it may hold the key to the 'alignment problem'. Imitation may at first seem inferior when we have the chance to instil profound moral values in machines. However, by pursuing human objectives, rather than their own programmed ones, an agent is able to learn the consensus on what is considered 'good'. Through this we can hope for safe AI without any sinister exploitation of its hardwired goals. IRL demonstrates the power of imitation allowing a remote control helicopter to perform a manoeuvre thought impossible for a human to achieve and 'Cooperative Inverse Reinforcement Learning' [7] ensures that the objective of the human is adhered to. Any moral agent would benefit from learning a human's moral code descriptively in this way, creating a rational and ethical AI. This approach, however, is not without its drawbacks. Firstly, actually mimicking this behaviour is difficult in itself as our actions very frequently do not reflect our intentions. Humans also suffer from various weaknesses from person to person; we do not want a moral agent to reflect all of our foibles. Further, our preferences change as we get older or experience new things; the example of the 'Moral Machine' still holds - which human is the best one to imitate? The principles of IRL rely on an 'expert'. This is someone who is sure of their objective and knows what they are doing. We do not have such a moral expert to train all AI on and this concept in itself collapses the distinction between the normative and descriptive approach.

There is also the far more sinister implication for the imitation of human behaviour in machines being the mechanism for laundering human bias through machine learning. High profile investigations into established software using machine learning such as COMPAS [4] demonstrate a worrying tendency to expose the worst prejudices in society. Aguera y Arcas takes this further [1] comparing the facial recognition service of an Israeli startup called Faceception to 'recognise people with high IQ, academic researchers, professional poker players, terrorists and pedophiles' with the eugenic scientists of Nazi Germany measuring people's noses to determine whether they are Jewish. Whilst this may seem tangential, algorithms based on large databases are in effect descriptive of human behaviour. Clearly a simple consensus of human behaviour is not sufficient in building a moral AI.

In comparison with the evaluation of the normative approach, sometimes the lines between the two are more blurred that we might imagine. Besold and Uckelman [5] argue that the normative role of rationality in AI is built on a notion of human rationality which is descriptive and not normative at all; AI aims at building machines which reason as humans do. This is a more abstract philosophical idea that any normative theoretical framework is and has to be descriptive of human rationality, but it emphasises the fact that the two approaches are not entirely separate.

It is necessary to build on both ideas by introducing uncertainty into the decision making process of moral agents [14]. We must ensure they do not fall too easily into either the moral dilemmas of the normative approach nor the pitfalls of the descriptive approach.

It is essential that moral agents have some degree of moral uncertainty built into them, removing the idea of a fixed morality from normative or descriptive approaches. Stuart Russell, one of the leading lights in the field of ethical AI, [15] outlines three normative principles for AI in a nod to Asimovs laws of robotics:

1. *The machine's only objective is to maximise the realisation of human preferences*

2. *The machine is initially uncertain about what those preferences are*

3. *The ultimate source of information about human preferences is in human behaviour*

The second principle is the one that makes these tenets so powerful. In other words, machines should follow a descriptive model of human behaviour, but ensure they do not follow them too seriously. Agents should reason as if they are incomplete and potentially flawed in dangerous ways. We can once again draw from centuries of philosophical study when asking how certain we should be of our ethics. Catholic theologians have grappled with this very question. When trying to live your life by the rules of your faith, how can you ensure you do not sin if there are significant disagreements in the church over exactly what constitutes good and evil? This question forms the basis of the 'effective altruism' movement which has become arguably the most significant social ethical movement in this century. MacAskill, Ord and Bykvist [11] have formalised the definition of how we weigh the importance of the impacts of actions taken now against future repercussions. The movement has cultivated a community that has a strong sense of moral uncertainty in decision making, frequently in regards to the impending climate crisis. To take a simple example, a government wants to tax a population for carbon emissions; this would presently make people far worse off but increase the welfare of the future generation by slowing down climate change. MacAskill asks the policy maker to consider the *choiceworthiness* of a range of different normative moral theories. Ord takes this further by asking the reader to imagine sitting in a kind of 'moral parliament' with various different ensembles and coalitions coming to a consensus about the best way to proceed. This is a new concept in philosophy never mind in the AI domain. However the empirical representation of the *choiceworthiness* function presented has the potential to provide a rationality to machines when they lack a single, certain standard against which any action will be judged.

Moral agents need to know *what* to do when they are unsure of the *right* thing to do. To use the analogy of an autonomous vehicle heading straight into a tree, the machine needs a way of evaluating which way to swerve to avoid the crash.

There is no clear 'best' way to build moral agents when considering a normative or descriptive approach alone. The majority of people would agree on the fact that AI needs to align with human values but defining what these values should be is a monumental challenge. This is exacerbated when considering the wide range of different cultures, socioeconomic backgrounds and human requirements across the globe. Agreeing on these values through prescribing normative rules is challenging as experiments such as 'The Moral Machine' demonstrate. Mimicking human values descriptively has enormous potential in the research community from an objective based point of view, but the problem is more difficult when considering ethical boundaries. The most likely solution to this is to introduce built-in uncertainty into the machines, however, this needs careful study and astute comprehension of different moral theories. The imperfection of any proposed approach to building moral agents reflects an imperfection in humans themselves; humanity does not agree on common values. We are myopic, emotional and computationally deficient, frequently making choices we regret. AI provides us a unique opportunity to interrogate our own paradoxes and ensure machines make fewer decisions they, and their creators, might regret.

# References

[1] Blaise Aguera y Arcas. *The Better Angels of our Nature*. URL: `https://www3.centro.edu.mx/PDF/MemoriasVOR2017Ingles.pdf`. 2017.

[2] Isaac Asimov. *I, Robot*. New York: Bantam Books, 1950.

[3] E. Awad et al. "The Moral Machine Experiment". In: *Nature* 563.7729 (2018), pp. 59–64. URL: `https://www.nature.com/articles/s41586-018-0637-6`.

[4] Matias Barenstein. "ProPublica's COMPAS Data Revisited". In: *ArXiv* abs/1906.04711 (2019).

[5] Tarek R. Besold and Sara L. Uckelman. "Normative and descriptive rationality : from nature to artifice and back." In: *Journal of experimental and theoretical artificial intelligence.* 30.2 (Mar. 2018), pp. 331–344. URL: `http://dro.dur.ac.uk/23844/`.

[6] Dalia Eidukiene. "Jeremy Bentham: The Ideal of the Greatest Happiness for the Greatest Number of People as a Modus Vivendi". In: *Filosofija Sociologija* 28 (Jan. 2017), pp. 29–37.

[7] Dylan Hadfield-Menell et al. "Cooperative Inverse Reinforcement Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3916–3924. ISBN: 9781510838819.

[8] Rosalind Hursthouse and Glen Pettigrove. "Virtue Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University, 2018.

[9] Robert Johnson and Adam Cureton. "Kant's Moral Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University, 2022.

[10] Stanley Kubrick. *2001: A Space Odyssey*. Metro-Goldwyn-Mayer, 1968.

[11] William MacAskill, Krister Bykvist, and Toby Ord. *Moral Uncertainty*. Oxford University Press, 2020.

[12] Stephen Nathanson. “Kant’s Moral Philosophy”. In: *The Internet Encyclopedia of Philosophy*. Ed. by James Fieser. Northeastern University, 2022.

[13] Andrew Y. Ng and Stuart Russell. “Algorithms for Inverse Reinforcement Learning”. In: *in Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, 2000, pp. 663–670.

[14] Stuart Russell. “Learning Agents for Uncertain Environments (Extended Abstract)”. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT’ 98. Madison, Wisconsin, USA: Association for Computing Machinery, 1998, pp. 101–103. ISBN: 1581130570. DOI: 10.1145/279943.279964. URL: https://doi.org/10.1145/279943.279964.

[15] Stuart Russell. *Reith Lectures 2021: Lecture 4; Beneficial AI and a Future for Humans*. URL: https://downloads.bbc.co.uk/radio4/reith2021/BBC_Reith_Lectures_2021_4.pdf. 2021.