CS4920 | Essay 2 Final

Aayush Bajaj | z5362216

April 29, 2025

Contents

§ Question 1

What is the difference between *explainability* and *interpretability*? In what ways might XAI be helpful or unhelpful?

§§ Response

XAI is helpful in finance, health, transport, etc, but current XAI is not meeting the expectations because our approaches are too technical and not interdisciplinary enough (Malizia 2023).

XAI is only unhelpful in the sense that improving interpretability comes at the cost of accuracy (Lipton 2018). (Think decision trees vs. deep neural nets.)

Mariarosaria Taddeo affirms the importance of both explainability and interpretability in her 2022 paper, but positions interpretability as more important than explainability. She does this in her first principle (of 8) in designing (Non-Lethal) Automated Weapon Systems. She asserts that such systems should be more interpretable and less explainable.

Rosaria draws her definition of interpretability and explainability from Rudin's 2019 paper as intepretability being the designing of models whose operations are transparent and understandable to humans, whilst explainability is the ad-hoc attempt of making sense of opaque models after they have been trained (LIME, SHAP).

Finally, there is still a smudged view of both interpretability and explainability, where both are used (unfortunately) synonymously in the AI literature (Doshi-Velez and Kim).

§ Question 2

> We can remove discrimination by removing all group membership information from the dataset (for example, by removing gender data), and the model would become fair to different gender groups. Similarly, we can remove information about age or race. Do you agree or disagree with this approach of *fairness through unawareness?* Why?

§§ Response

This seems like it might work, but a paper by Cornacchia in 2023 disproved it by entertaining the counterfactual "If gender information was removed, then what would happen?".

It turns out the algorithms just learn the same features in a different latent space through proxy variables.

Yates splits the biases into three buckets of algorithmic, activity and data. Clearly removing the data bias will not affect the systemic bias that also prejudices a group.

And finally, perhaps the most informative paper on the matter: The (Im)possibility of Fairness, Friedler et al. 2016 positions the worldview of being individually fair as mutually exclusive with being group-fair. The authors mathematically juxtapose the WYSIWYG (What You See is What You Get) data approach with the WAE (We are All Equal) approach.

Thus, whilst you can remove group discrimination by removing the group information, you will not eliminate ALL discrimination, and further there is still a large chance for the group relationships to be learned nonetheless.

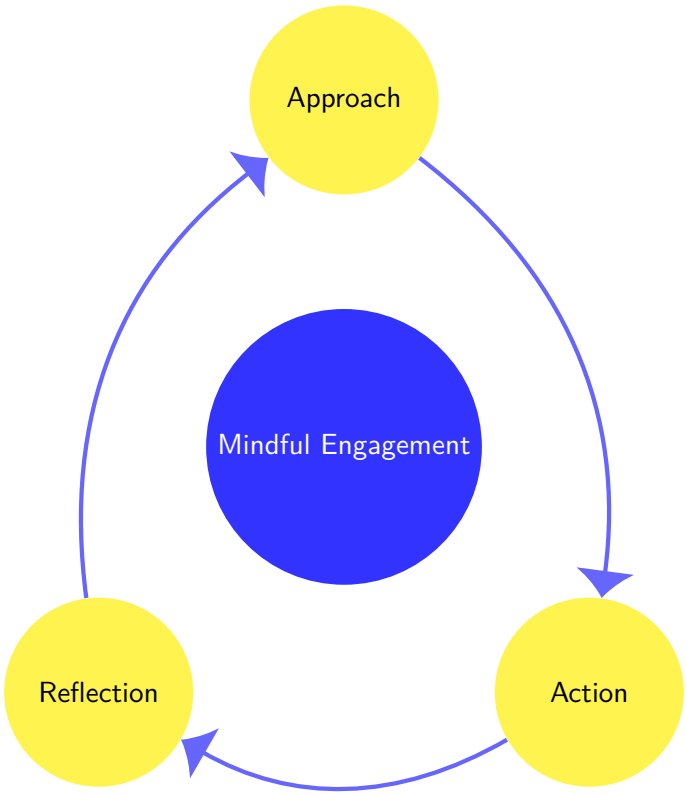
§ Question 3

> What does *everyday leadership development* look like? Give an example (either actual or counterfactual) from your Group Project in terms of Positive Organisational Scholarship. In your answer, place yourself in the role of hypothetical Group Project Manager.

§§ Response

From the paper by Ashford and DeRue (2012), we learn that the Positive Organisational Scholarship perspective enables people to become leaders at all levels of an organisation, and that one needs to take responsibility for their own leadership development.

In the context of our Group Project I was cognisant of the Leadership lecture and of constructing a Mindful Engagement Feedback loop with both actual and counterfactual hypotheticals.



One particular hypothetical was between myself and Shayyan wherein I suggested he do the fourth part of the assignment. I ran counterfactuals against Shayyan particularly because I did not want to overstep. In the end I maintained a forward but flexible stance on delegating parts, simply getting to know my Group Members and asking about their degrees / courses and interests so that I may act as the kintsugi of the group and lead by not leading (Lao Tzu, Tao Te Ching)

§ Question 4

What is ethical principlism? Is it useful, dangerous or both? Why?

§§ Response

Ethical Principlism falls within Applied Ethics and is different to both Normative and Meta-ethics. As such, due to its closeness with the physical world, principlism becomes both useful and dangerous. I believe though, that the pros outweigh the cons.

The origins date back to Biomedical Ethics and Beauchamp and Childress' (2004) work on the matter. In a historical development paper by Beauchamp and DeGrazia, they recount that "a single-principle theory has struck many as misguided as well as presumptous or dogmatic". Here, they are referring to Utilitarian and Deontological approaches. The authors then go on to highlight that adherents to most Normative Ethical camps "agree at the level of the principles of biomedical ethics". As such Principlism was born.

Giving instruction to "be alert to matters of justice", and "to think about justice", suddenly "action" became promoted to first-class citizen out of Normative Ethics.

Furthermore, any conversation of Justice alludes to Rawls. Indeed, his work on Reflective Equilibrium underpins what the Justice is interpretted as in the following 4 Principlism Principles:

1. Respect for Autonomy
2. Beneficence
3. Non-maleficence
4. Justice

Finally, by considering a tangible example of Emergency Department Triaging, we can quickly see how Ethical Principle is more useful than dangerous.

1. Listen to the patients
2. Do good
3. Do not cause harm
4. Be fair

Compare this now, to a Virtue Ethicist who would be forced to ask the ambiguous question:

What would an virtuous person do in this situation?

Overall, it is clear that Ethical Principlism is a set of percolated principles, informed by Normative and Metaethical study that can be used to derive ethical documents such as the ACM code.

As a concluding caveat Principlism is not perfect, and we shall see 2 arguments below that turn the simplicity of Principlism into its naivete.

+ Cybersecurity ethics. They just include Explicability. "A Principlist Framework for Cybersecurity". 2021 et al.

§ Question 5

> Why does Munn claim that AI Ethics principles are meaningless, isolated, and toothless? Is he correct? Why?

§§ Response

He is correct. There is a gap between high-minded principles (as enumerated in Q4) and technological practice.

Meaningless means commendable values such as "fairness" and "privacy" break down when subjected to scrutiny. Value "X" can really be bent to mean whatever you need it to mean, and furthermore developers may conduct an "ethics shopping".

Isolated means that unethical AI is the logical byproduct of an unethical industry! Munn provides case studies of Silicon Valley which highlight the misogynistic environment and of Universities that fail to teach the appropriate macro-ethical concepts.

Toothless means that AI ethical principles have failed due to the lack of consequences. Munn provides a case study of Google here, who whilst producing an ethical board (great!), do not give the board any authority to enforce actions in any meaningful way (boo!).

Ultimately, despite the pessimism of Munn, he provides a solution in that AI Ethics should be re-branded to AI Justice.

§ Question 6

What are the Menlo Principles? Which type of normative ethics might be used to justify each of the principles? Why?

There are 4 Principles:

1. Respect for persons → Kantian
2. Benificence → Utilitarianism
3. Justice → Virtue Ethics
4. Respect for Law → Rule-based Utilitarianism

It is worth noting the replacement of Respect for Law with Non-maleficence in the Principlism approach. The reason for this is that the Menlo Principles add legal and public accountability to their status quo.

The existence of the Menlo Principles as an entity proves what a study of all the Normative Ethics individually suggests: we need an ensemble of these models to make effective ethical decisions.

Briefly, I recount each of the stances and construct a mapping between their core values and the Menlo Principle in Question

1. **Kant:** Autonomy and rational agency are central. The Categorical Imperative demands that we treat others as ends in themselves, never merely as means → Respect for Persons.
2. **Mill:** Right actions are those that promote the greatest happiness for the greatest number → Beneficence reflects this utilitarian aim of maximizing overall good.
3. **Aristotle:** Justice is a cardinal virtue and a mean between selfishness and selflessness. It sustains social harmony and reflects moral excellence → Justice aligns with Virtue Ethics.
4. **Rule Utilitarians:** Moral rules (like laws) help ensure consistent, beneficial outcomes for society over time → Respect for Law ensures accountability through stable rule-following.

§ Question 7

> Assume Nihilistic Error Theory. How might the moral education of computer science students then proceed?

§§ Response

Error Theory posits that the truth-maker is unknowable, and additionally nihilism posits that there is no way to know the truth-value of these moral judgements.

As such there are no moral facts, and morality is a systematic error rooted in false beliefs about objective value.

With this in mind, we must construct a moral education for computer science students without appealing to objective moral truths.

Instead of collapsing into relativistic apathy or nihilistic despair, we can structure things much in the way that this course has done so:

- normative ethics are socially and practically useful despite nihilistic error theory
- professional codes (acm, ieee) can be thought as shared fictions
- bias, fairness explored in terms of stakeholder values
- transparency, explainability focusses on pragmatic necessity instead of a moral imperative
- pragmatic pluralism, where metaethical consideration is given to competing ethical frameworks to enrich utility not truth
- systems thinking as opposed to individual blame

To summarise, even under nihilistic error theory, moral education in computer science can proceed constructively — not by asserting false objectivity, but by reframing ethics as a socially negotiated toolkit that enables cooperation within an otherwise amoral universe.

§ Question 8

> What are the risks and opportunities for our understanding and practice of moral responsibility given the rise of automated weapons systems in particular, and automated decision-making systems in general?

§§ Response

Moral Responsibility is defined and distinguished by Taddeo and Blanchard in A Moral Gambit:

1. Intentionality
2. Causality
3. Consequence
4. Choice

The authors distinguish this from meaningful moral responsibility, which can only apply to Non-Lethal Automated Weapon Systems, as deployment of Lethal Autonomous Weapon Systems is "morally unacceptable".

Thus the risks posed by automated systems are manifold:

- eroding choice as AWS / ADM are by definition automated
- unsatisfiability triad: ADM cannot simultaneously satisfy *fairness, accountability and accuracy.* attempts to improve one (fairness) reduces the other (accuracy). this is bad because moral responsibility requires all three. this can also result in ethics-bashing (Bietti 2020)
- risk of moral deskilling and normative drift. over time, human capacity for complex decision making and ethical integrity could erode.
- opacity of ADM's allows institutions to exploit ethics-washing (Bietti 2020)

Yet, there are also opportunities:

- HCAI with High Control and High Automation (Shneiderman 2020)
- IEEE couples individual and institutional responsibility with authors of ADM's

It seems that ultimately automated systems — especially in high-stakes domains like warfare or sentencing — have more cons than pros. The unsatisfiability of the fairness-accountability-transparency triad underscores the need for human oversight and Taddeo's Moral Gambit plus the IEEE supports the need for a HITL (Human in the Loop).

In conclusion, the risks outweigh the benefits and whilst we have systems in place to facilitate a constrained optimisation of this Wicked Problem, we are still stuck in a local minima for now.

§ References

Ashford, Susan J. (2012). "Developing as a Leader: The Power of Mindful Engagement". In: *Organizational Dynamics* 41.2, pp. 146–154.

Baeza-Yates, R. (2018). "Bias on the web". In: *Communications of the ACM* 61.6, pp. 54–61.

Beauchamp, Tom L. and David DeGrazia (2004). "Principles and Principlism". In: *Principles of Health Care Ethics*. Ed. by Raanan Gillon. John Wiley & Sons, pp. 55–66.

Bietti, Elettra (2020). "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy". In: *Philosophy & Technology* 33.4, pp. 541–559. DOI: 10.1007/s13347-020-00405-1. URL: https://doi.org/10.1007/s13347-020-00405-1.

Computing Machinery, Association for (June 2018). *ACM Code of Ethics and professional conduct*. Association for Computing Machinery. URL: https://www.acm.org/code-of-ethics (visited on 03/30/2025).

Cornacchia, Giandomenico et al. (2023). *Counterfactual Reasoning for Bias Evaluation and Detection in a Fairness under Unawareness setting*. arXiv: 2302.08204 [cs.LG]. URL: https://arxiv.org/abs/2302.08204.

Daniels, Norman (2020). *Reflective Equilibrium*. https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/. Stanford Encyclopedia of Philosophy, Summer 2020 Edition, edited by Edward N. Zalta.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*. URL: https://arxiv.org/abs/1702.08608.

Formosa, Paul, Michael Wilson, and Deborah Richards (2021). "A Principlist Framework for Cybersecurity Ethics". In: *Computers & Security* 105, p. 102226. DOI: 10.1016/j.cose.2021.102226.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian (Mar. 2021). "The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making". In: *Commun. ACM* 64.4, pp. 136–143. ISSN: 0001-0782. DOI: 10.1145/3433949. URL: https://doi.org/10.1145/3433949.

Friedman, B. (1996). "Value-sensitive design". In: *interactions* 3.6, pp. 16–23.

Gabriel, Iason and Vafa Ghazavi (2021). "The Challenge of Value Alignment: From Fairer Algorithms to AI Safety". In: *Minds and Machines* 31.4, pp. 629–653. DOI: 10.1007/s11023-021-09563-0.

Macnish, Kevin and Jeroen van der Ham (2020). "Ethics in Cybersecurity Research and Practice". In: *Technology in Society* 63, p. 101382. DOI: 10.1016/j.techsoc.2020.101382.

Malizia, Alessio and Fabio Paternò (2023). "Why Is the Current XAI Not Meeting the Expectations?" In: *Communications of the ACM* 66.12, pp. 20–22. DOI: 10.1145/3588313.

Munn, Luke (2023). "The Uselessness of AI Ethics". In: *AI and Society*. DOI: 10.1007/s00146-023-01673-z.

Sayre-McCord, Geoffrey (n.d.). *Metaethics*. https://iep.utm.edu/metaethi/. Internet Encyclopedia of Philosophy.

Sequoiah-Grayson, Sebastian (2025). "The Unsatisfiable Triad: A Problem for Automated Decision Making". Manuscript in preparation. Unpublished manuscript. URL: https://logicalrockpools.com/.

Shneiderman, B. (2020). "Human-Centered Artificial Intelligence: Three Fresh Ideas". In: *AIS Transactions on Human-Computer Interaction*. Vol. 12. 3, pp. 109–124.

Taddeo, Mariarosaria and Alexander Blanchard (2022). "Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—a Moral Gambit". In: *Philosophy & Technology* 35.3, pp. 1–24. DOI: 10.1007/s13347-022-00571-x.