Name: Eamonn Herlihy **Student ID:** 14155605

Module: IN5103 - Risk, Ethics, Governance and AI

Assignment: Week 4, Part B: Create a report that analyses and defends the design and content of your code of conduct

Note: This assignment submission is accompanied by a supporting code of conduct (Week 4, Part A).

Report

With advances in Artificial Intelligence (AI), the economy, business, healthcare, and society have and will continue to undergo unprecedented transformations. The benefits that AI derives for society can be easy to see, with a wide range of applications such as diagnosing cancer (Horie et al. 2019), translating language (Roziere et al. 2020; Ranade et al. 2018), mimicking human speech (Kim et al. 2019), playing games and defeating the most competent humans (Holcomb et al. 2018), etc. Despite these obvious benefits, the risks are much more subtle. Increasingly, people are becoming interested in AI-related rights and ethics as these staggering AI developments progress. Among the reasons for this concern are the vast quantities of personal data that AI technologies are now able to capture, the large number of analytical and statistical tools available, the sheer power of its computation, and the uncertain effects on the interests and well-being of humanity.

Many of these innovations have been introduced in the past decade and as a society, we are slowly becoming aware of the many, potentially substantial risks. Due to this recency, the governance surrounding AI to date is limited. The EU is very much leading in the way in attempting to rectify this issue. 'The AI Act' is a proposed European law on artificial intelligence (AI), the first of its kind to be proposed by a major regulator anywhere. Although this is clearly a very positive step in the right direction, many believe that regulation alone will be insufficient. Zuiderwijk et al. (2021) argue that, as AI systems are becoming more complex and less predictable, most governments have a limited understanding of the multifaceted implications and therefore are unable to implement sufficient governance. Moreover, regulation takes time and, in the paradigm-shifting, dynamic ecosystem that is the field of AI, keeping up with sufficient regulatory needs is a difficult task indeed. One can also take the cynical point of view of unsavoury motives, corporate lobbying, and party politics as a barrier to proper governance (Ron and Singer, 2020). It is due to such rationale, that effective governance in our new AI society, is likely to contain a mix of hard governance (such as the EU's AI Act) and soft governance. One such method of soft governance organizations, industries, and professions have at their disposal is the implementation of codes of conduct.

"Corporate codes of conduct are a practical corporate social responsibility (CSR) instrument commonly used to govern employee behaviour and establish a socially responsible organizational culture" (Erwin, 2011). The literature surrounding corporate codes of conduct

is not a recent phenomenon. Over the years there has been a lot of research and debate surrounding their prevalence, their effectiveness, and what makes an effective code of conduct. White and Montgomery (1986) study the prevalence of such codes of conduct in the US, the variations of such codes, and the general topics covered. The authors show that such codes were prevalent, and many had been developed, expanded, or modified following "The Foreign Corrupt Practices Act of 1977" (White and Montgomery, 1986). Somers (2001) further added to this body of literature by attempting to prove the hypothesis, "Unethical behaviour is less prevalent in organizations that have adopted corporate codes of ethics than it is in those organizations that have not formally adopted such codes" (Somers, 2001). Somewhat surprisingly to the author, a realisation that although most organisations had indeed codes of conduct, only 8% of those who participates (employees of such organisations) reported being aware of such codes of conduct. This suggests that although codes of conduct existed, there were certainly not being utilised correctly. As a result, the proposed hypothesis was rejected.

Fortunately, since the Somers paper in 2001, the literature suggests that code of conduct utilisation has dramatically increased and is achieving tangible results, especially such codes that are well written and fit for purpose. For example, Erwin (2011) revealed a positive relationship between the quality of codes of conduct and their effectiveness on corporate conduct, ethical performance, public perception, and sustainability. Erwin (2011) states that "companies that devote specific human and capital resources to developing comprehensive codes of conduct that are consistent with corporate values have a more significant impact on ethical behaviour within the organization". It is due to such findings that modern-day companies place a great emphasis on developing such codes of conduct as an effective soft governance tool.

Having outlined the benefits for organisations having a well-written, fit for purpose, and properly utilised code of conduct, I will introduce the submitted code of conduct produced for the company NewBank. NewBank is a fintech start-up that gains a competitive advantage from its extensive use of AAAI (Advanced Analytics & Artificial Intelligence) solutions throughout its business. The code of conduct was produced for all NewBank employees working within the AAAI division as well as employees that interact in any way with the AAAI suite of products and services. The document produced is defined using 5 cornerstone principles that align with NewBank's core values and serve to ensure NewBank and its employees uphold the highest moral and ethical standards. The following paragraphs will serve as a defence of the design and content of this code of conduct, beginning with the justification for inclusion of the 5 cornerstone principles.

Explainability & Transparency

As described in this report, the progress made in AAAI over the past years has been remarkable. Much of this success has come in the form of deep learning - a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data. In comparison to traditional machine learning methods such as decision trees and support vector machines, deep learning methods have achieved substantial improvement in various prediction tasks. "However, deep neural networks (DNNs) are comparably weak in explaining their inference

processes and final results, and they are typically treated as a black-box by both developers and users" (Xu et al. 2019). And so, a trade-off emerges. Unfortunately, there have been countless examples of AI that have gone terribly wrong where corporate bodies develop complex deep learning-based solutions where proper explainable was not a priority. For example, in 2014 Amazon's AI recruiting system was found to be sexist (Dastin, 2018) and Leslie et al. (2021) argue that the lack of explainability in the AI systems used during the covid-19 pandemic had adverse effects on vulnerable groups, exacerbating health inequity. Of course, the importance of transparency depends on the potential of a system to harm stakeholder interests or rights and the importance of redress.

In the case of NewBank, a company where AI products and services directly impact customers and customer outcomes, transparency is of the utmost importance. This is an obligation NewBank takes seriously, and the code of conduct sets a strict criterion that is required to be met before launching any product or service. In essence, NewBank has committed to only launching products such that all decisions made via their AAAI can be clearly explained to customers. In fact, during development, an XAI (Explainable AI) solution that produces an automated document to clearly explains each decision made is produced. This is an exceptionally high standard to set, and I believe, NewBank is paving the way for proper AI corporate governance with this approach.

Integrity of AAAI

The first principle set out the requirement for transparency in the AAAI and algorithms NewBank uses. However, equally important is the entire data infrastructure required for these algorithms to function. As a company that utilises AAAI, it, therefore, has a large reliance on data (and big data). With data at the heart of all things at NewBank, maintaining a big-data infrastructure that is secure, maintains privacy, fit-for-purpose, efficient and accurate is a top priority. The code of conduct states:

'It is important that AAAI systems are robust, secure, and safe throughout their entire lifecycle, so that, whether they are used normally, misused, or otherwise adversely affected, they continue to function properly and do not pose unreasonable safety risks'

The code of conduct highlights that as NewBank utilises information or data that is private or sensitive, it is important to make sure that the system does not violate or infringe upon the right to privacy and that private and sensitive data is well-protected. Maintaining this standard is a regulatory requirement as per the GDPR - General Data Protection Regulation (Albrecht, 2016) and one that NewBank has embraced.

The code outlines general practices that should undertake regularly in order to maintain these high standards NewBank has a regulatory requirement to meet as well as those that are self-imposed.

Fairness & Alignment to Human Rights

The two previous principles are intended to ensure that outcomes are transparent in terms of how decisions are made and that NewBank's data infrastructure is kept to the highest

standard in terms of security, privacy, accuracy, etc. The 'Fairness & Alignment to Human Rights' principle is the next layer and is added to ensure that Newbank's AAAI solutions produce fair and equitable outcomes. This principle requires employees to critically evaluate any AAAI, proposed or otherwise, and sets out an appropriate framework to do so. This framework was borrowed from the work of Beauchamp and Childress (1979) who wrote about the principles of biomedical ethics. Even today and in the field of AI, this landmark paper remains exceptionally relevant.

The code of conduct defines the four principles and for clarity utilises NewBank-specific examples. The requirement for a technology assessment at each technology decision point is required as per this document whereby the employee should examine the short and long-term consequences of the technology, Beit societal, economic, ethical, legal, etc. using the framework defined by Beauchamp and Childress (1979).

The framework aims to ensure that NewBank maintains a great respect for the autonomy of each customer, each solution causes no harm (intended or otherwise, to customers or otherwise), employees aim to do as much good as they can for customers and those outcomes are distributed fairly – free from bias.

Contestability & Human Empowerment

The fourth principle is defined to support the empowerment of all stakeholders through open lines of communication. This principle is defined broadly to include all stakeholders but places particular emphasis on two groups in particular, customers and employees.

The code of conduct aims to foster great respect for customers. The first three principles defined help in this goal. However, this principle, in particular, empowers customers to challenge decisions made by NewBank's AAAI systems and set in place processes that allow for effective human review, oversight, and control of AAAI decisions. NewBank instils a sense of humility with this principle, the acceptance that NewBank may not be operating optimally at all times, and that customer input can be an invaluable tool to improve the suite of products and services, the customer experience, and build greater trust and transparency. Kunz et al. (2017) support this approach as they emphases the benefits of customer engagement in big-data products and suggests that this benefit can be optimally realised via a value-generating process that is dynamic and continuously changing through an iterative process.

The second key group this principle empowers is its employees. In an effort to ensure NewBank is operating as ethically as possible, the code of conduct encourages employees to speak up in the event that they have any ethical or legal concerns. The literature suggests that fostering a culture where open communication is encouraged, and retaliation is prohibited can ethical positive results (Watts and Ronald Buckley, 2017).

Responsibility & Accountability

The first four principles are set out to ensure that NewBank acts with the utmost moral and ethical integrity in all that they do. However, the final piece of this puzzle is the principle of

accountability and responsibility. It can be argued that the previous four principles would be insufficient without the addition of this final principle. In order to achieve the goals defined by Newbanks's code of conduct, accountability and personal responsibility must be a priority. To achieve this, the code of conduct sets out several requirements for its employees including clarity in defining the responsibilities of each individual role and the requirement for these parameters to be updated periodically as the technological landscape shifts.

Although the key pillars/principles defined in a code of conduct are of great importance to the success of this soft governance mechanism, it is not the only consideration. The literature suggests that as well as defining such principles, other factors such as clear and concise format, astatically appealing design, and formal tone can yield preferable outcomes (Bischof and Eppler, 2011; McCloskey 1987; Gino et al. 2020). For example, Gino et al. (2020) examined the relationship between the language used (personal or impersonal) in such codes of conduct and corporate illegality. Surprisingly, the author found that firms that used personal language in their codes of conduct were more likely to be found guilty of illegal behaviours. Using the literature as a guide in the design of NewBank's code of conduct, much emphasis was placed on ensuring clear and concise language, creating an aesthetically pleasing document, and ensuring the language remained impersonal i.e., not using the term "we" or "our".

The final consideration in the design of this document was around enforcement. In the introduction, this code makes very clear that these are not just guidelines but requirements and that failure to comply would lead to disciplinary actions and potential termination. Berglöf and Claessens (2004) suggest that corporate governance and enforcement mechanisms are intimately linked as they affect firms' ability to govern effectively. This suggests that such documents tend to lack effectiveness unless those intended to utilise such codes fear repercussions from deviating from the code.

Creating a code of conduct as the one defines is not a simple task and in creating this document, I encountered several difficulties. Examples include defining the appropriate scope, information overload, creating an aesthetic design, and writing style. In the end, I limited my scope to those utilising AAAI within NewBank, identified appropriate literature as it related to AI ethics, favoured research with many citations, leaned on a subset of codes of conduct as a guide, utilised a simple but aesthetically pleasing template and choose to use a formal writing style. I would argue that the produced code of conduct is in fact a professional style document that is fit for purpose and ready to be deployed.

Having conducted this assignment, I appreciate the considerable value that codes of conduct can derive for organisations, professions, and industries. Moreover, I realise that it is important for a code of conduct to be written carefully, fit for purpose, updated as appropriate, and appropriately enforced to ensure optimal results. This being said, I do not believe this to be the sole answer to the complex problem surrounding AI governance. Instead, I believe codes of conduct form an element in the multifaceted approach that will be required to reach an equilibrium where ethical AI and social good are a top priority in all organisations.

Bibliography

Albrecht, J.P., 2016. How the GDPR will change the world. Eur. Data Prot. L. Rev., 2, p.287.

Beauchamp, T. L. and Childress, J. F. (1979). Principles of biomedical ethics. New York, Oxford University Press.

Berglöf, E. and Claessens, S., 2004, September. Corporate governance and enforcement. World Bank, Corporate Governance Department, Global Corporate Governance Forum.

Bischof, N. and Eppler, M.J., 2011. Caring for Clarity in Knowledge Communication. J. Univers. Comput. Sci., 17(10), pp.1455-1473.

Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of Data and Analytics (pp. 296-299). Auerbach Publications.

Erwin, P.M., 2011. Corporate codes of conduct: The effects of code content and quality on ethical performance. Journal of Business Ethics, 99(4), pp.535-548.

Floridi, L., 2018. Soft ethics, the governance of the digital and the General Data Protection Regulation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), p.20180081.

Gayle, H.D. and Childress, J.F., 2021. Race, racism, and structural injustice: equitable allocation and distribution of vaccines for the COVID-19. The American Journal of Bioethics, 21(3), pp.4-7.

Gino, F., Kouchaki, M. and Feldman, Y., 2020. Your Code of Conduct May Be Sending the Wrong Message. Harvard Business Review.

Gunning, D. and Aha, D., 2019. DARPA's explainable artificial intelligence (XAI) program. AI magazine, 40(2), pp.44-58.

Holcomb, S.D., Porter, W.K., Ault, S.V., Mao, G. and Wang, J., 2018, March. Overview on deepmind and its alphago zero ai. In Proceedings of the 2018 international conference on big data and education (pp. 67-71).

Horie, Y., Yoshio, T., Aoyama, K., Yoshimizu, S., Horiuchi, Y., Ishiyama, A., Hirasawa, T., Tsuchida, T., Ozawa, T., Ishihara, S. and Kumagai, Y., 2019. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointestinal endoscopy, 89(1), pp.25-32.

Kim, N.Y., Cha, Y. and Kim, H.S., 2019. Future English learning: Chatbots and artificial intelligence. Multimedia-Assisted Language Learning, 22(3), pp.32-53.

Kunz, W., Aksoy, L., Bart, Y., Heinonen, K., Kabadayi, S., Ordenes, F.V., Sigala, M., Diaz, D. and Theodoulidis, B., 2017. Customer engagement in a big data world. Journal of Services Marketing.

Leslie, D., Mazumder, A., Peppin, A., Wolters, M.K. and Hagerty, A., 2021. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?. bmj, 372.

McCloskey, M.A., 1987. The Importance of Aesthetics. In: Kant's Aesthetic. Palgrave Macmillan, London.

Ranade, P., Mittal, S., Joshi, A. and Joshi, K., 2018, November. Using deep neural networks to translate multi-lingual threat intelligence. In 2018 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 238-243). IEEE.

Ron, A. and Singer, A.A., 2020. Democracy, corruption, and the ethics of business lobbying. Interest Groups & Advocacy, 9(1), pp.38-56.

Roziere, B., Lachaux, M.A., Chanussot, L. and Lample, G., 2020. Unsupervised translation of programming languages. Advances in Neural Information Processing Systems, 33, pp.20601-20611.

Solomon, J., 2020. Corporate governance and accountability. John Wiley & Sons.

Veale, M. and Borgesius, F.Z., 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), pp.97-112.

Watts, L.L. and Ronald Buckley, M., 2017. A dual-processing model of moral whistleblowing in organizations. Journal of Business Ethics, 146(3), pp.669-683.

White, B.J. and Montgomery, B.R., 1980. Corporate codes of conduct. California management review, 23(2), pp.80-87.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D. and Zhu, J., 2019, October. Explainable AI: A brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing (pp. 563-574). Springer, Cham.

Zuiderwijk, A., Chen, Y.C. and Salem, F., 2021. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, 38(3), p.101577.