EKIP: designing a practical framework for embedding ethics in AI software development

Rebecca Raper[1] and Mark Coeckelbergh[1]

[1]The University of Vienna, Vienna, Austria

rebecca.raper@univie.ac.at

Abstract. As Artificial Intelligence (AI) systems are used more within society, there is the pressing need to ensure that they are developed in an ethical way. Several notable attempts to ensure that AI is developed ethically have been advanced. Though each approach has its merit in terms of offering guidance for ethical AI development, none, yet offer a rigorous, practical approach to ensure that AI systems are developed in the right way. Furthermore, these approaches seem to be out of sync with current actual software development methodologies. EKIP is a three-year-long FFG-sponsored project which aims to address this issue by offering a *practical* framework for software developers. Inspired by *Ethical by Design* (Mulvenna et al., 2017) – that ethics should be placed into the design phase of software development – EKIP's purpose is to explore more practical approaches for AI software developers to embed ethics into their systems. The aim of this paper is to outline the motivation behind EKIP, by giving an overview of the attempts that have been made to embed ethics in AI so far, and by introducing a 'methodological gap'. It is argued that if we are to effectively embed ethics in AI systems, we need to consider new frameworks that integrate with the traditional software development techniques, such as 'waterfall' and 'Agile'. An *ethical requirements*-based approach is described which presents a twist on traditional software development methodologies. To conclude, suggestions are made for future work, and there is a plea for further research to advance this topic.

Keywords: AI Ethics, AI Frameworks, Ethical AI, EKIP, Software Development Lifecycle, Software Development, Waterfall, ML-Ops, Kanban, Requirements Engineering

1 Background and Introduction: Ethical AI

Defining *Ethical Artificial Intelligence* (AI) as the pursuit to design, create and implement AI systems that adhere to ethical standards (Coeckelbergh, 2020), the discipline has gained an increasing amount of traction in recent years. This is partially owing to various high-profile incidents concerning the use (or misuse) of AI systems. For instance, in March 2020, it was reported that Cambridge Analytica was collecting and processing vast quantities of data to form decisions about the types of

advertisements an individual might be targeted with, or what articles they see online (BBC News, 2020). The controversy and objections concerned the fact that AI profiling was being used to target and manipulate individuals with material relating to a political referendum campaign, highlighting the vast amount of influence AI decision-making can have. Also in 2020, A-Level result allocation in the UK had a similar story, with algorithms used to assign grades to students showing to be prejudiced against those from certain demographics or poorer backgrounds (The Guardian, 2020) – the implication being that if you were from a poorer background, you were less likely to get into the university of your choice – ultimately reducing social mobility and career fulfilment. There have been similar cases in the United States, where AI profiling used by the police to forecast criminal behaviour, has been shown to be prejudiced against those from minority groups (Wired, 2020). Similarly, facial recognition technology used in procedures such as interviews, or for identity verification, has been shown to be less effective for users with darker skin (The Independent, 2021). As AI is being used more and more in society, to deal with economic, social, and technological problems, there is the concern that these identified types of incidents are widespread and that they are having a significant effect on individual lives and communities.

Other implications of AI systems concern issues such as *privacy* – if an AI system is making inferences about our personality characteristics, to what extent do we still have privacy? Another issue is *trust* – with algorithms having such bad press (whether this be qualified or not), how can (and should) public perception of AI be developed so that they can realise the true benefits that AI can bring? *Autonomy* and *control* are other issues – who gets to decide how AI algorithms are applied, and then decide the eventual fate of individuals based upon the algorithmic decisions? This then leads to questions of delegated *power*, and *political influence* (Coeckelbergh, 2022). There is also a piece for *education*. With AI systems affecting everybody's lives (in some way), and with Artificial Intelligence itself due to become increasingly powerful (Bostrom, 1998) everybody should be equipped to understand how it works and the impact it has on their day-to-day lives, so they can appropriately interrogate, ask questions and engage in social and political debates.

It is these such issues that have led to calls for more Ethical AI and have led to increased academic attention to the area. Institutes such as 'The Ethical AI Institute' in the UK, and 'The Institute for Ethical AI and Machine Learning' in Germany, have been established in recognition of this problem - the idea being that interdisciplinary collaboration (i.e. between science and the humanities (Tasioulas, 2021)) will help to resolve some of these issues. There are calls for immediate action to be taken.

2 Attempts to deal with the issues so far

Various attempts have been made to address the ethical issues pertaining to AI, without the need to call a moratorium on AI use/development in general. One of the first attempts made was to create a set of guidelines or principles for AI developers to follow. Different bodies have different principles (see Floridi and Cowls, 2021) but

broadly speaking, the principles can be reduced to themes such as *Transparency*, *Trust*, *Autonomy* and *Explainability*, namely, an AI is only deemed ethical if it matches these principles - so is Trustworthy, Grants Human Autonomy etc. In line with the principles, Ethical Standards have been developed to highlight what needs to be done to ensure that the principles are adhered to. For instance, a system might only be regarded as 'explainable' if it can be explained to stakeholders in a certain way. Under this approach, only systems that meet these standards can be said to be ethical, and therefore worthy of approval. The problem is: with such standards in place, it's difficult for developers to understand (without extensive training) (a) what the specific standards mean or (b) what *actually* needs to be done to meet them. Requests such as 'explain to stakeholders' can be ambiguous when notions such as 'explainability' are still very much open to philosophical debate (Arrieta et al, 2020). With the very pressing need to ensure that AI systems are developed in the right way now (rather than in 10 years' time after the terms have gone through rigorous philosophical discussion) another methodology is required that prescribes to developers how they can create their AI systems to meet such standards.

'Ethical by Design' (Mulvenna et. al, 2017) is one such methodology that has been proposed to ensure this, with the suggestion being that ethics be integrated into the AI software development process from the very start. This gained a significant amount of traction, with the IEEE adopting an 'Ethically Aligned Design' (How, 2018) manifesto to aim to achieve this. However, how to create AI systems that meet the various list of standards is still left very much open. Still – developers know they must do something to make their systems more ethical, but it is not obvious what.

One proposal - 'Value-sensitive design' (see Umbrello, 2018, Friedman, 2004 and Van den Hoven, 2007) - aims at determining a set of values at the beginning of the development process from which to work upon. Stakeholders (and end users) undergo a series of interviews to assess *what is important to them*, and then this information is used to inform the design of the AI system.

However, there are problems with such an approach. Namely, no end of user interaction will be able to fully capture every individual value, and even if they were, how do we then deal with the fact that individual values may be conflicting? An additional mechanism is required on top of this approach to determine which values take priority, and without strict oversight, this then faces the possibility of becoming nothing more than a series of prejudices. More fundamentally, however, the Value by Design approach neglects the foresight and academic rigor that has gone into the development of principles and standards such as those developed by the IEEE in the first place. There needs to be some way to bridge the gap between AI system design and the need for adherence to principles such as transparency, fairness etc.

More recently, AI Auditing (Mokander and Floridi, 2021) has been offered as a verification method to ensure that AI systems meet ethical standards previously set out. So, AI systems can be assessed against a checklist of criteria for satisfaction of the standards. However, though a good mechanism to assess that AI systems are doing what they should, there is still a *methodological gap* insofar as something is required to outline to developers how to make their AI systems ethical.

3 A 'methodological gap' in Ethical AI

In practice, we can see that we have *top-down* principles or standards that prescribe how AI systems ought to be developed (i.e. with *transparency* etc). We also have *horizontal* methodologies that advise that such principles should be incorporated into the AI system at the start of AI system creation (i.e. Ethical by design). We also have a technique to assess that AI systems have been created at the appropriate ethical level (AI auditing), and some techniques to show us that stakeholder values are important in the design of AI systems. Schematically, we can represent the interplay of all the different approaches by a diagram as per Figure 1:

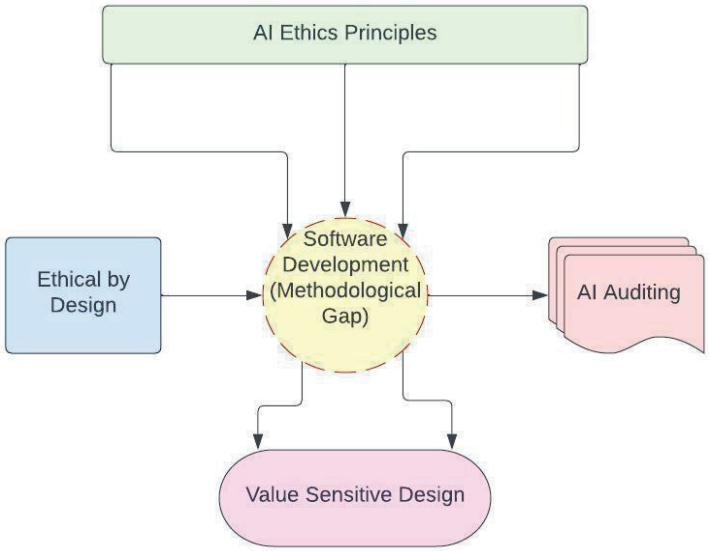

Fig. 1. The current state of AI Ethics

We can see that although we have mechanisms to provide oversight of AI systems (namely, AI Ethical Principles), and a technique to prescribe when we ought to be incorporating ethics into our systems (Ethical by Design), and how to assess them (AI Auditing), there is still space to understand how we ought to be developing the AI systems to ensure that both the principles are adhered to, and that the relevant appropriate standards are met. **This** is the methodological gap.

Describing this problem in terms of a 'methodological gap' allows us to begin to envisage what needs to be done to rectify the problems associated in the first chapter of this paper: namely, AI bias, discrimination, privacy and broad impacts to humans. What is required is a methodology that takes note of the already determined AI principles (i.e.

that it's important not to infringe upon autonomy), assess these against real-life potential impacts to stakeholders (i.e. my autonomy might be infringed if AI decides what I'm watching on TV tonight), and then puts mechanisms in place to protect against these. Schematically again, this need can be represented as per Figure 2.

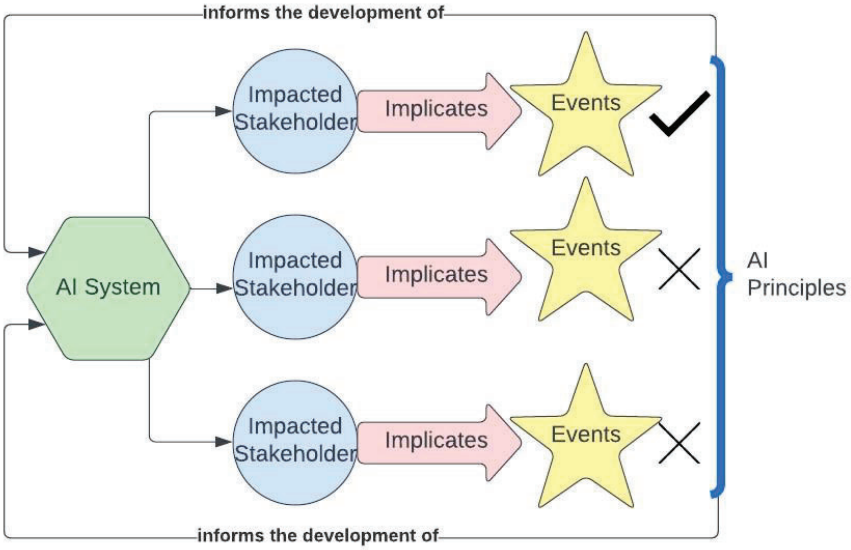

Fig. 2. A need for integration of current Ethical AI techniques.

As can be seen from the diagram, this approach still appeals to the 'Ethical by Design' methodology, since what ultimately informs the development of the AI system, is either an infringement or adherence to AI Ethical Principles (indicated by the cross or tick on the diagram): the system being ethical insofar as it adheres to Ethical Principles (such as those previously discussed). However, the way in which this approach differs is that the principles act as a *guide* as opposed to enforcement for how AI systems ought to be developed. In other words, the AI developer's autonomy is preserved, since they are not forced to develop their systems in a certain way (See Raper et al, 2022, for some benefits of preserving developer autonomy). Up until recently, a top-down regulationstyle approach has been used (see Madiega, 2019), but it has been criticised for potentially inhibiting innovation. The new approach allows AI systems to be developed ethically, whilst preserving AI developer autonomy.

In understanding this approach, it's important to realise that it is not an impact-based model. There have been recent discussions about using impact assessments to determine how to develop an AI system (Kazim, 2021) (the ethical system being that which does not lead to negative consequences). Ultimately, however, it is not possible to know the full impact a system might have until it has been developed and put into place. We might reflect afterwards (based upon impact) that an AI system should not

be the way it is (i.e. the negative consequences of the Cambridge Analytica situation), but this approach would be *reactive* as opposed to *proactive*. We want to ensure that there are no negative consequences for the AI systems that are developed, and therefore, that they are ethical before they are integrated into the world.

As opposed to assessing an AI system based upon its (previously unknown) impacts, the AI system is assessed according to its (previously known) implications. The implications can be determined through case studies, interviews and workshops with impacted stakeholders. For example, I might assess how a new medical diagnostic device will implicate the end stakeholders (i.e. patients) by asking a series of questions to determine how the system (in principle) will affect their lives. A conversation such as the following might be used in a user workshop to ascertain this:

AI Analyst: We are developing an AI tool to assess how likely somebody is to develop diabetes based upon their current diet and lifestyle.

Patient: That sounds good, but some days I am healthy, other days I eat unhealthy because it depends upon how busy I am…

This is a very short interview, but it highlights the types of conversations that need to be held to truly ascertain the human implications an AI system might have. As this example demonstrates, after discussion with a patient we see that such a tool would need to take account of variable diets and lifestyles (since some patients eat mixed diets), therefore, we are able to determine that there is a requirement to ensure that any AI system used to predict diabetes risk, considers diets that change. After significant further analysis, we begin to develop a requirements-based model for Ethical AI Development of this specific system.

There are many ways to satisfy the requirements specified during the analysis phase, but what is important to note is that doing so should be an exercise in *requirementsbased design*, rather than harm/bias etc. mitigation. In this instance, the new system has the need to accommodate for changing diets. It then becomes the AI developer's creative agency to design a system which satisfies this.

4 EKIP: A practical methodology for Ethical AI Software Development

The requirements-based approach to Ethical AI software development is one that has been proposed previously (Guizzardi et al, 2020), however there has not yet been a thorough articulation for why this approach is needed, how it significantly differs to current methodologies, and – most importantly - what an implementation of the solution might look like.

EKIP (shorthand for "Ethische KI Plattform" = "Ethical AI Platform") - an Austrian, FFG-sponsored project - was established in an attempt to provide a very practical way to integrate ethics into software development by proposing that ethics be integrated into an already-established AI development platform created by the company Gradient Zero, called 'DQ0'.

DQ0 offers a privacy-preserving platform for data scientists to develop Machine Learning algorithms to run queries against sets of (sensible) data. The aim of EKIP then is to extend the platform so that when AI developers create their systems, it facilitates creation of them in an ethical way.

Referring to the Ethical Requirements Framework approach detailed in the previous section, therefore, DQ0 best enables this facilitation (and in turn, preserves developer creativity and agency) by providing a platform to integrate the elicited requirements. Therefore, the most effective way to practically integrate ethics that resolves the 'methodological gap' is to create a new software development process that allows for *Ethical Requirements Elicitation*. Figure 3 shows how this should be done in respect to the DQ0 platform.

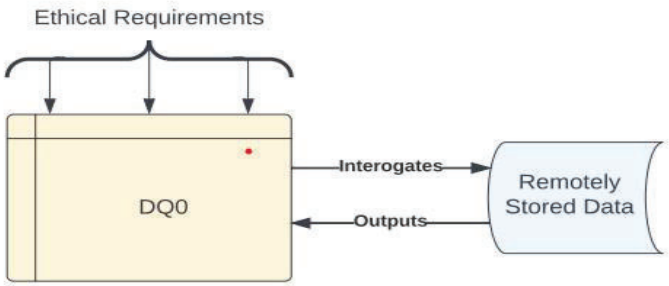

Fig. 3. EKIP - A practical approach to embedding ethics in AI software development

As mentioned previously, the ethical requirements would be ascertained through case studies, interviews, or general analysis techniques. For instance, in the (brief) conversation scenario outlined above, we determine that the ethical requirement is that: 'The AI system must accommodate for variable diets', with the development *problem to be solved* being left to the freedom of the AI developer. It may be that a new role is required within a project team to satisfy this new type of requirement – an individual with the capacity to facilitate such workshops and map implications for affected stakeholders. The key to this phase is in ensuring that the analysis assesses stakeholder implications rather than just looking toward impacts.

5 Integrating ethics into pre-existing software development methodologies (i.e. Waterfall and Kanban)

It is important in understanding how to execute this framework, that it integrates well into the pre-existing software development processes, and that it can also be distinguished from these. Where we are able to know the outcomes of a traditional piece of software, when it comes to AI software development, because AI acts autonomously, the outcomes are not always known.

Previously, before AI software development became more commonplace, methodologies such as 'The Waterfall Method' (see Figure 4) were traditionally used to ensure that software systems were designed according to the needs of the customer (or business). Traditionally, this included several phases, the first being 'project scoping' and 'requirements elicitation', where it was determined (1) how useful the new system might be and then (2) what design requirements the system should have to satisfy the overarching business objectives. Though The Waterfall Method has evolved since its initial inception (i.e. more agile software development approaches such as Kanban (Huang, 1996) are now frequently used), fundamentally, the same procedural steps remain (albeit in a more dynamic way). However, though still useful for broad software development, these are insufficient for AI software development. Because an AI system acts on its own, it is not possible to design it to meet specific conditions (i.e. have a blue screen) - since these will be variable. Instead, mechanisms are required to ensure that the AI system acts in an *ethical way* - hence it must meet ethical design requirements.

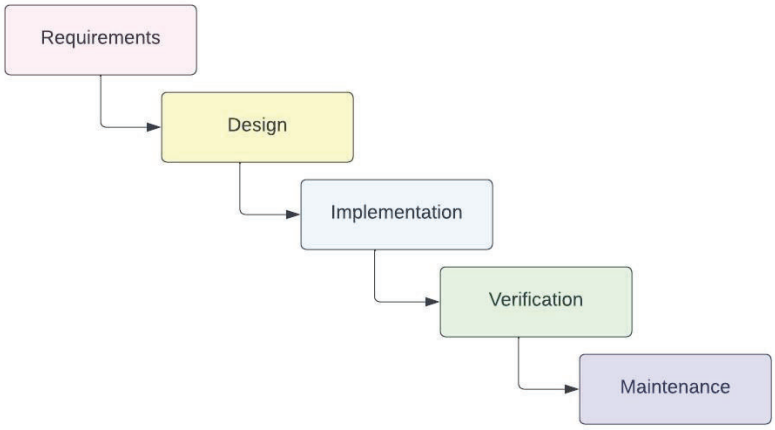

Fig.4. The 'Waterfall' software development methodology

Considering ethical requirements elicitation alongside traditional software development approaches, therefore the new process for Ethical AI Software Development should look as per Figure 5.

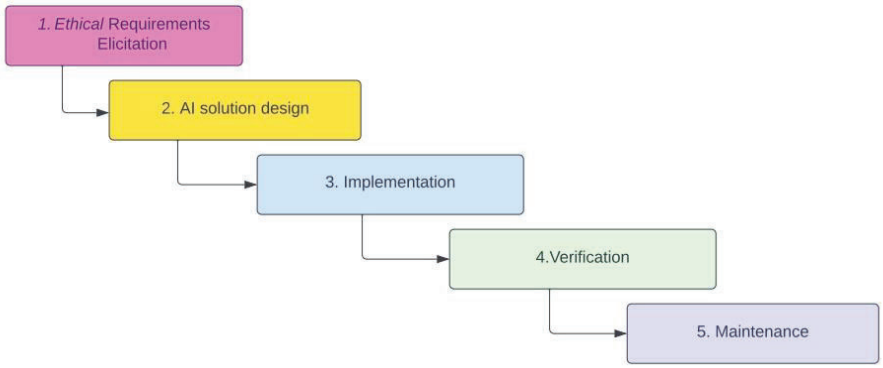

Fig.5. The Ethical AI Software development cycle.

Though the models are very similar, as already mentioned, the type (and methodology) for the requirements elicitation (and thereby solutions design) are very different. In Ethical AI case, workshops are used to find out not just what needs to be done to meet business needs, but what needs to be done to ensure that human beings have a fulfilling and flourishing life. As already mentioned, this might require a new role within the project team – a psychologist or social scientist? – but it is a twist on the original requirements elicitation framework.

As with traditional software development, it is important that there is still verification of the AI system in the end (i.e. it should match the earlier prescribed requirements) along with constant re-evaluation and maintenance of the AI system to ensure issues such as 'AI drift' (the susceptibility to later become biased) don't creep in. In the case of AI systems this means periodic checking to ensure that the system still meets its initial ethical requirements (or requirements for the current stakeholders) and an iterative design process to constantly improve an evaluate the AI system.

As a final point, it is worth mentioning processes that have already been established to try to address the differences between traditional software and AI software development. For example, Microsoft's 'ML-Ops' (Alla and Adari, 2021) was designed to accommodate for these very differences. However, though offering an agile environment for AI development, they neglect the need for 'human-centricity' in AI software development. This can only be achieved through thorough a proper analysis of humanity and human lives.

6 Conclusion

Much current academic literature and industry conversations talk about integrating ethics into software, but few practical methodologies are available to do this in specific software development and business contexts in a way that is useful to developers and respects the way they work and solve problems. In this paper we have described a motivation for, and given a broad description of the EKIP project, which

enables us to work on precisely this problem through exploring how ethics can be integrated into software development methodologies in the context of an AI platform.

We propose that the best approach to achieving the Ethical AI aims set out by bodies such as the IEEE and The Ethical AI Institute, is by integration of an *ethical requirements gathering* technique into pre-existing software development processes.

In terms of next steps, to progress this research further work is required to develop the finer details of the ethical requirements elicitation process. There need to be tools to facilitate this methodology as well as specific roles and responsibilities. There is also space for further discussion about what constitutes a fulfilling, flourishing life, in the context of Technology Ethics and Ethical AI more broadly, and a plea to work more towards this *practical* type of approach.

Acknowledgements

As this is an FFG-sponsored project, we would like to thank our funders for giving us the opportunity to develop this programme of work. We would also like to thank the team at Gradient Zero who have been keen and enthusiastic partners in this pursuit, in particular Jona Boeddinghaus who gave guidance and insightful comments on a late version of this paper.

References

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges to ward responsible AI. Information fusion, 58, pp.82-115.

Alla, S. and Adari, S.K., 2021. What is mlops?. In Beginning MLOps with MLFlow (pp. 79124). Apress, Berkeley, CA.

BBC News, 2020, 'Your data and how it is used to gain your vote': https://www.bbc.co.uk/news/technology-54915779 (Last Accessed: 27/04/2022)

Bostrom, N., 1998. 'How long before superintelligence?'. International Journal of Futures Studies.

Coeckelbergh, M., 2020. AI ethics. MIT Press.

Floridi, L. and Cowls, J., 2021. A unified framework of five principles for AI in society. In Ethics, Governance, and Policies in Artificial Intelligence (pp. 5-17). Springer, Cham.

Friedman B, 2004. Value sensitive design. In: Bainbridge WS (ed) Berkshire encyclopedia ofhuman-computer interaction. Berkshire Publishing Group, Great Barrington

Guizzardi, R., Amaral, G., Guizzardi, G. and Mylopoulos, J., 2020, May. Ethical requirements for ai systems. In Canadian Conference on Artificial Intelligence (pp. 251-256). Springer, Cham.

How, J.P., 2018. Ethically aligned design [From the Editor]. IEEE Control Systems Magazine, 38(3), pp.3-4.

Huang, C.C. and Kusiak, A., 1996. Overview of Kanban systems.

Kazim, E., Denny, D.M.T. and Koshiyama, A., 2021. AI auditing and impact assessment: according to the UK information commissioner's office. AI and Ethics, 1(3), pp.301310.

Madiega, T.A., 2019. EU guidelines on ethics in artificial intelligence: Context and implementation.

Mökander, J. and Floridi, L., 2021. Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), pp.323-327.

Mulvenna, M., Boger, J. and Bond, R., 2017, September. Ethical by design: A manifesto. In Proceedings of the European Conference on Cognitive Ergonomics 2017 (pp. 51-54).

Tasioloas, J. (2021) The role of the arts and humanities in thinking about artificial intelligence (AI) | Ada Lovelace Institute

The Independent, 2021, 'How racist robots are being used in recruitment': https://www.independent.co.uk/news/world/americas/robots-racism-algorithms-jobhiring-b1860835.html (Last Accessed: 27/04/2022)

The Guardian, 2020, 'The Guardian view on A-Level algorithms: failing the test of fairness.': https://www.theguardian.com/commentisfree/2020/aug/11/the-guardian-view-on-alevel-algorithms-failing-the-test-of-fairness (Last Accessed: 27/04/2022)

Raper, R., Boeddinghaus, J., Coeckelbergh, M., Gross, W., Campigotto, P. and Lincoln, C.N., 2022. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. Sustainability, 14(7), p.4019.

Shahriari, K. and Shahriari, M., 2017, July. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In 2017 IEEE Canada International Humanitarian Technology Conference (IHTC) (pp. 197-201). IEEE.

Umbrello, S. and De Bellis, A.F., 2018. A value-sensitive design approach to intelligent agents. Artificial Intelligence Safety and Security (2018) CRC Press (.ed) Roman Yampolskiy.

Van den Hoven MJ, 2007. ICT and value sensitive design. In: Goujon P et al (eds) The informationsociety: innovation, legitimacy, ethics and democracy. Springer, Dordrecht, pp 67–73

Wired, 2020, 'Police built an AI to predict violent crime: it was seriously flawed': https://www.wired.co.uk/article/police-violence-prediction-ndas (Last Accessed: 27/04/2022)