Ethics in Artificial Intelligence: Should there be more regulation?

**Introduction**

With the increased utilisation of AI across businesses and consumers, there has been debate over the ethics and fair usage in AI. This can include AI models that are trained on bias data, and therefore reproduce that bias in its outcomes; ethics surrounding usage of others intellectual property without the owners permission to train AI models; and ethics for what AI is used for in its product (i.e. ‘deepfake’ technology being used to impersonate important figures).

Thus, there are four topics to explore relating to ethics and its subsequent regulation:

* Bias in the training and development of AI models
* Use of intellectual property without relevant permissions or licencing
* Usage and purpose of AI
* Current and future guidelines and regulation

**Research Question**

With consideration of the above, a pertinent question is does AI need statutory regulation, beyond mere guidelines, to overcome bias and ethical issues that are prevalent in many AI models and their utilisation? If so, to what extent can regulation be implemented without hindering development and innovation and what could stricter regulation look like?

**Significance**

The significance of proposing regulation is multifaceted due to the multifaceted nature of the question posed.

Firstly, there is a need to overcome bias in AI to prevent bias outcomes and eliminate the self-perpetuating issue emanating from usage of bias data and necessitates exploring whether more significant regulation, especially when it comes to matters of public health and crime, will help alleviate that bias.

Secondly, the use of others intellectual property to train AI models has sparked controversy and debate on the topic, but its regulation could limit performance of models and reintroduce the bias previously discussed due to having more limited resources. Thus, this research will seek, if possible, to find a balance.

Lastly, exploring the potential of restricting usage and perhaps even criminalising malicious usage of AI which may prevent unethical use of AI products.

The aim then is to explore these topics and whether regulation is possible and, where it is, to what capacity can be reasonably expected of regulation and its applicability with due consideration given to any current or upcoming regulation. Subsequently, the outcome will be a framework for regulation and supplementary demonstrative AI models based off this research.

**Methodology and Artefacts**

Most of the artefacts will be produced via secondary research and subsequently will be a collation of discussions and documents which will be critically examined in order to produce the final artefacts discussed below.

For the first step, a comprehensive literature review of the underlying issues regarding bias and ethics for each topic will be undertaken. Secondly, there will be a literature review of the ongoing discussions relating to regulation and guidelines for each of the areas including both arguments for and against more regulation. Following this, a third literature review of the currently suggested guidelines and in place regulation for each topic will be conducted. Examples of this would include (but not limited to) the EU AI Act and regulations from the World Health Organisation (WHO) (World Health Organization, 2023).

From this, a report will be generated that takes the work from each of the literature reviews and evaluates current and future regulation of each topic as well as the potential limitations resultant from increased regulation. It will aim to give insight on how far regulation can go without impeding research and AI models and subsequently provide suggestions on what a higher regulatory standard for AI would look like. This can then be compared to current and other suggested guidelines and regulations. Further, this will analyse the possibility of whether regulations can be generalised or do they need to be specific to a given industry or domain.

A framework for regulation, either specific to each topic or more generalised, will be established. This can then be used as a basis to create examples of best practice using the suggested regulations in the development of AI models with accompanying reports for each of the topics explored, and, if possible, a holistic AI model that demonstrates and complies with all suggested regulation.

Because this work cannot possibly account for all topic and cover all development and usage of AI models, there is future direction for regulation on AI that is not covered by this research.

**Key Literature**

The literature for this research proposal can be split into four topics. Each of them is briefly described below followed by a table with a non-exhaustive list of literature.

**Bias in training and development of AI models**

In the context of AI, bias can be defined as: “the difference in performance between subgroups for a predictive task” (Yang et al., 2023).

Bias introduced in the development and training of AI is prevalent in a plethora of areas, but where it has the greatest impact is in the utilisation of AI in the judicial process, in crime prevention, and in medical diagnostics. Usage in these areas has a greater impact due to the potential life changing affects it may have on the people that are subject to bias outcomes originating from AI.

In a medical context, such bias can result in underdiagnosis of underrepresented groups such as females, people of colour, and patients of lower socioeconomic backgrounds due to the AI being trained on data with little representation of these subgroups (Seyyed-Kalantari et al., 2021). There are many examples of this such as in X-ray imaging (Larrazabal et al., 2020) as well as a widely used AI algorithm in the US health system (Obermeyer et al., 2019). Literature such as this are paramount to the discussion of the greater need for regulation in this area and, whilst there is already extant literature on checklists and tools such as CLAIM (Mongan et al., 2020) and PROBAST (Wolff et al., 2019) that aim to assess AI algorithms, regulation is limited and thus bias remains prevalent. However, there are some attempts at regulation from the US Food and Drug Administration (FDA) (Muehlematter et al., 2023) and the WHO (World Health Organization, 2023) and these will be integral to further discussion on regulation.

Similar bias can be seen for judicial process regarding assessment of the risk a person will commit a crime in the future (Mayson, 2019) as well as in judicial decision making (Winter, 2022). Furthermore, crime prevention suffers from much racial bias which can also adversely affect AI (Barabas, 2019). Again, such discourse and literature on bias furthers the need for regulatory interference for each area and has led to literature discussing possible remedies which will be crucial when considering possible regulation in these areas.

**Use of intellectual property without relevant permissions or licencing**

With copyright lawsuits regarding lack of attribution and compensation for material used by generative AI ongoing (Samuelson, 2023), there is much discussion on the fair use of intellectual property in AI (Sobel, 2017). The outcomes of the lawsuits should the plaintiffs be successful could determine the future of AI and may restrict AI to only using public domain works or licenced works. Thus, examining this ongoing debate of fair use and the respective case law will help determine how regulation could and should develop.

**Usage and purpose of AI**

There are a multitude of stories that demonstrate AI is used for purposes that can be construed as unethical. For example, a shopper algorithm buying items on the darknet bought drugs and was subsequently ‘arrested’ (Eveleth, 2015). Whilst this is a more obvious example, there are also examples where AI such as deepfakes present ethical challenges where subjects of research studies are purposely deceived (Eberl et al, 2022). Further, the ethical implications of such AI have been discussed in the context of the US election (Diakopoulos et al, 2020). Again, the literature is key because each demonstrates unregulated usage of AI and necessitates the discussion of regulation in addition to discussion of current and future regulation.

**Current and future guidelines and regulation**

Because of the risks and ethical issues discussed, there is recognition from not only researchers and developers, but also organisations such as UNESCO which has produced a text named ‘Recommendation on the Ethics of Artificial Intelligence’ (*Ethics of Artificial Intelligence* 2022). Notably, this is a developing area with the European Union looking to implement the EU AI Act as the first attempt to regulate AI. The initial bill seeks to prohibit applications of AI such as emotion recognition in certain settings and social scoring (*Artificial Intelligence Act: Deal on comprehensive rules for trustworthy AI: News: European parliament* 2023). However, there remains exemptions for law enforcement, which as discussed previously, has prevalent bias. This may be a good first step at regulation and makes note that there should be measures in place to support innovation, but is it great enough in scope to resolve the ethical issues in AI beyond it’s usage? This will be discussed at length in subsequent reviews and reports resultant from this research.

|  |  |  |
| --- | --- | --- |
| **Content** | **Title** | **Category** |
| Regulations | Regulatory Considerations on Artificial Intelligence for Health - World Health Organization | Website |
| Regulations | Ethics of Artificial Intelligence - UNESCO | Website |
| Regulations | Artificial Intelligence Act: Deal on comprehensive rules for trustworthy AI: News: European parliament | Website |
| Bias - Judicial & Crime | Bias in, Bias out | Journal Article |
| Intellectual Property | Generative AI meets copyright | Journal Article |
| Intellectual Property, Regulations | Ai and IP: Theory to policy and Back again – policy and research recommendations at the intersection of Artificial Intelligence and intellectual property | Journal Article |
| Intellectual Property, Regulations | Legal reform to enhance global text and Data Mining Research | Journal Article |
| AI Usage | 'my robot bought illegal drugs' | Website |
| AI Usage | Anticipating and addressing the ethical implications of deepfakes in the context of elections | Journal Article |
| Bias - Judicial & Crime | The challenges of artificial judicial decision-making for liberal democracy | Journal Article |
| Bias - Judicial & Crime | Chapter 14 Automated Justice: Issues, Benefits and Risks in the Use of Artificial Intelligence and Its Algorithms in Access to Justice and Law Enforcement | Book |
| Bias | Bias in data‐driven Artificial Intelligence Systems—an introductory survey | Journal Article |
| Bias - Judicial & Crime | The datafication revolution in criminal justice: An empirical exploration of frames portraying data-driven technologies for crime prevention and Control | Journal Article |
| Bias - Judicial & Crime | Beyond bias: Re-imagining the terms of ‘ethical ai’ in criminal law | Journal Article |
| Intellectual Property | Can artificial intelligence infringe copyright? some reflections | Journal Article |
| Intellectual Property | Artificial Intelligence’s Fair Use Crisis | Journal Article |
| Bias - Medical, regulations | FDA-cleared Artificial Intelligence and machine learning-based medical devices and their 510(k) predicate networks | Journal Article |
| Bias - Medical | Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations | Journal Article |
| Bias - Medical | Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis | Journal Article |
| Bias - Medical | Dissecting racial bias in an algorithm used to manage the health of populations | Journal Article |
| Bias - Medical, regulations | Checklist for Artificial Intelligence in Medical Imaging (Claim): A guide for authors and reviewers | Journal Article |
| Bias - Medical, regulations | PROBAST: A tool to assess the risk of bias and applicability of Prediction model studies | Journal Article |

This is a table that shows several key pieces of literature that helped produce this initial proposal and will be the foundation of further research.

**Ethical Considerations and Risk Assessment**

It is crucial that research on ethics considers and complies with best practices. For this, I have considered four areas that may impact this research.

**Plagiarism**

For plagiarism, it is vital to acknowledge the use of other people’s work, especially in a project that leans heavily on secondary research. Wherever necessary, work will be cited and referenced correctly.

**Intellectual Property**

With intellectual property already discussed in this paper, it’s important to acknowledge that where any texts, datasets, or algorithms used that are propriety, permissions and licences will need to be obtained for fair usage.

**Data Privacy**

Regarding data privacy, in any instance where data could contain personally identifiable information, it is important to acknowledge this and either remove or anonymise the data to avoid infringements on data privacy. The data used in this research will come from publicly available sources.

**Outdated Information**

Due to the ongoing changes regarding AI regulation, it is important to keep up with the latest developments and keep them in consideration up until the publish date of the research. Additionally, should there be anticipated developments, such as new legislation and case law regarding the regulation of AI, this will also be taken into account and noted in any relevant artefacts.

**Timeline**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Activity** | **Description** | **Estimated Start Date** | **Estimated End Date** | **Number of Days** |
| Supervisor Consultation | Guidance and discussion with supervisor | 07/03/2024 | 21/03/2024 | 14 |
| Literature Review I | On ethics and bias in AI | 22/03/2024 | 11/04/2024 | 20 |
| Literature Review II | On debate on regulation of each topic | 12/04/2024 | 02/05/2024 | 20 |
| Literature Review III | On current and future planned regulations | 03/05/2024 | 24/05/2024 | 21 |
| Report | Consolidation of the literature reviews into one report | 25/05/2024 | 08/06/2024 | 14 |
| Regulation Framework | A framework of regulations to alleviate ethical issues in AI | 09/06/2024 | 09/07/2024 | 30 |
| AI Model Creation | AI models following suggested regulations from framework | 10/07/2024 | 10/08/2024 | 31 |
| Feedback | Feedback from supervisor. Changes may be implemented to prior activities. | 11/08/2024 | 25/08/2024 | 14 |
| Documentation | Final documents of framework and demonstrative AI model | 26/08/2024 | 07/09/2024 | 12 |

This is a table that provides an overview of an estimated timeline for the activities to be undertaken for the research project. The literature reviews combined will take the most time, as this will be most of the research conducted to provide later artefacts. The final artefacts, ‘Regulation Framework’ and ‘AI Model Creation', are also considerable undertakings, as this will be the culmination of the research to produce guidelines and demonstrative AI models. Of course, feedback will be taken into consideration, and amendments and changes will be made wherever necessary.

**Conclusion**

To summarise, the purpose of the research is to examine regulation in AI in terms of why regulation is needed and how regulations are implemented in three of the discussed topics. In doing this, a framework for improvements and suggestions on how best to regulate AI can be created with due consideration given to the ethics and risks discussed. The reason for conducting this research is to find a solution (whilst giving fair consideration to balancing restrictiveness and leniency of regulation) to the ongoing bias and ethical issues surrounding unfair usage of intellectual property, discrimination stemming from biassed AI, and malicious or improper usage of AI.

**References**

*Artificial Intelligence Act: Deal on comprehensive rules for trustworthy AI: News: European parliament* (2023) *Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI | News | European Parliament*. Available at: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai (Accessed: 28 January 2024).

Barabas, C. (2019) ‘Beyond bias: Re-imagining the terms of “ethical ai” in criminal law’, *SSRN Electronic Journal* [Preprint]. doi:10.2139/ssrn.3377921.

Bonadio, E., Dinev, P. and McDonagh, L. (2022) ‘Can artificial intelligence infringe copyright? some reflections’, *The human cause*, pp. 245–257. doi:10.4337/9781800881907.00019.

Diakopoulos, N. and Johnson, D. (2020) ‘Anticipating and addressing the ethical implications of deepfakes in the context of elections’, *New Media &amp; Society*, 23(7), pp. 2072–2098. doi:10.1177/1461444820925811.

Eberl, A., Kühn, J. and Wolbring, T. (2022) ‘Using deepfakes for experiments in the Social Sciences - a pilot study’, *Frontiers in Sociology*, 7. doi:10.3389/fsoc.2022.907199.

*Ethics of Artificial Intelligence* (2022) *UNESCO.org*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 20 January 2023).

Eveleth, R. (2022) *‘my robot bought illegal drugs’*, *BBC News*. Available at: https://www.bbc.com/future/article/20150721-my-robot-bought-illegal-drugs (Accessed: 20 January 2024).

Fiil-Flynn, S.M. *et al.* (2022) ‘Legal reform to enhance global text and Data Mining Research’, *Science*, 378(6623), pp. 951–953. doi:10.1126/science.add6124.

Gans-Combe, C. (2022) ‘Chapter 14 Automated Justice: Issues, Benefits and Risks in the Use of Artificial Intelligence and Its Algorithms in Access to Justice and Law Enforcement’, in *Ethics, Integrity and Policymaking: The Value of the Case Study*. Springer.

Lavorgna, A. and Ugwudike, P. (2021) ‘The datafication revolution in criminal justice: An empirical exploration of frames portraying data-driven technologies for crime prevention and Control’, *Big Data &amp; Society*, 8(2), p. 205395172110496. doi:10.1177/20539517211049670.

Larrazabal, A.J. *et al.* (2020) ‘Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis’, *Proceedings of the National Academy of Sciences*, 117(23), pp. 12592–12594. doi:10.1073/pnas.1919012117.

Mayson, S.G. (2019) *Bias in, Bias out*, *The Yale Law Journal - Home*. Available at: https://www.yalelawjournal.org/article/bias-in-bias-out (Accessed: 20 January 2024).

Mongan, J., Moy, L. and Kahn, C.E. (2020) ‘Checklist for Artificial Intelligence in Medical Imaging (Claim): A guide for authors and reviewers’, *Radiology: Artificial Intelligence*, 2(2). doi:10.1148/ryai.2020200029.

Muehlematter, U.J., Bluethgen, C. and Vokinger, K.N. (2023) ‘FDA-cleared Artificial Intelligence and machine learning-based medical devices and their 510(k) predicate networks’, *The Lancet Digital Health*, 5(9). doi:10.1016/s2589-7500(23)00126-7.

Ntoutsi, E. *et al.* (2020) ‘Bias in data‐driven Artificial Intelligence Systems—an introductory survey’, *WIREs Data Mining and Knowledge Discovery*, 10(3). doi:10.1002/widm.1356.

O’Mathúna, D. and Iphofen, R. (2022) *Ethics, integrity and policymaking: The value of the case study*. Cham, Switzerland: Springer.

Obermeyer, Z. *et al.* (2019) ‘Dissecting racial bias in an algorithm used to manage the health of populations’, *Science*, 366(6464), pp. 447–453. doi:10.1126/science.aax2342.

Picht, P.G. and Thouvenin, F. (2023) ‘Ai and IP: Theory to policy and Back again – policy and research recommendations at the intersection of Artificial Intelligence and intellectual property’, *IIC - International Review of Intellectual Property and Competition Law*, 54(6), pp. 916–940. doi:10.1007/s40319-023-01344-5.

*Regulatory Considerations on Artificial Intelligence for Health* (1970) *World Health Organization*. Available at: https://iris.who.int/handle/10665/373421 (Accessed: 21 January 2024).

Samuelson, P. (2023) ‘Generative AI meets copyright’, *Science*, 381(6654), pp. 158–161. doi:10.1126/science.adi0656.

Seyyed-Kalantari, L. *et al.* (2021) ‘Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations’, *Nature Medicine*, 27(12), pp. 2176–2182. doi:10.1038/s41591-021-01595-0.

Sobel, B.L.W. (2017) ‘Artificial Intelligence’s Fair Use Crisis’, *The Columbia Journal of Law & The Arts*, 41(1), pp. 45–97. doi:https://doi.org/10.7916/jla.v41i1.2036.

Winter, C.K. (2022) ‘The challenges of artificial judicial decision-making for liberal democracy’, *Economic Analysis of Law in European Legal Scholarship*, pp. 179–204. doi:10.1007/978-3-031-11744-2\_9.

Wolff, R.F. *et al.* (2019) ‘PROBAST: A tool to assess the risk of bias and applicability of Prediction model studies’, *Annals of Internal Medicine*, 170(1), p. 51. doi:10.7326/m18-1376.