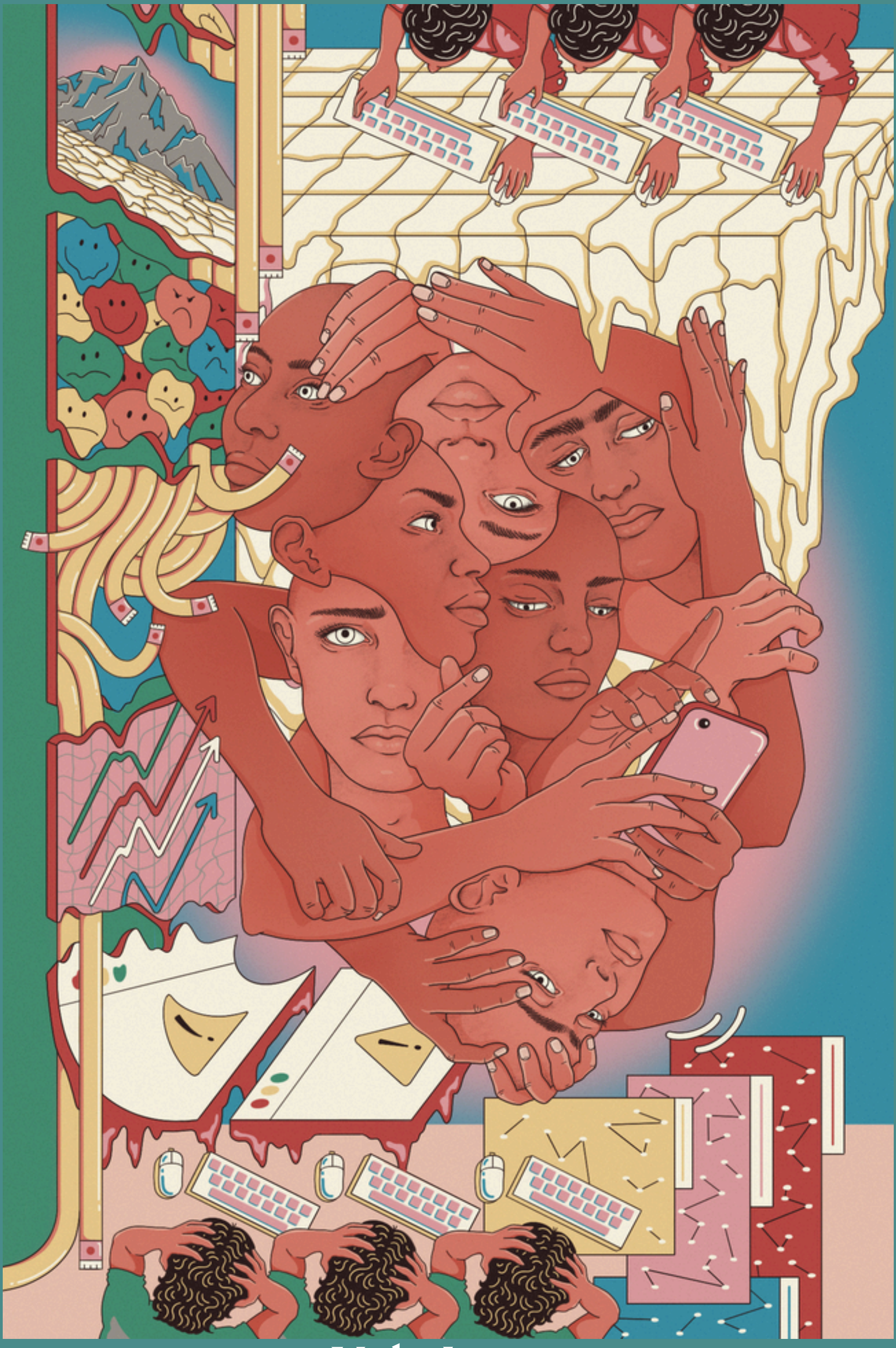
CAMBRIDGE JOURNAL OF ARTIFICIAL INTELLIGENCE



Vol. 1 Issue 1
July 2024

CJAI CAMBRIDGE JOURNAL OF ARTIFICIAL INTELLIGENCE

Published by *Cambridge Journal of Artificial Intelligence*, Cambridge, United Kingdom.

The cover artwork is *User/Chimera* by Clarote and AI4Media. The piece represents the user as a Chimera, a visual metaphor suggested by Kate Crawford, in which the "end-user" also provides valuable feedback, personal data and other invisible, unpaid labour. The illustration brings out the contrast between their multi-dimensionalities and the flat, biased categorisations of the human in inputs and outputs of AI.

Editorial Team

July 2024

Contents

Editorial

A warm welcome to the Cambridge Journal of Artificial Intelligence (CJAI).

AI has become the buzzword of the day, infiltrating every aspect of our lives, from the mundane to the profound. The rapid integration of AI technologies in both public and private sectors reflects a global race to harness its potential while simultaneously mitigating its associated risks. We have seen governments drawing up plans for autonomous vehicles as well as private companies leveraging AI to gain a competitive edge. It is the ever-changing nature and purpose of AI that has led to the creation of our organisation.

At the CJAI, our mission is to bridge the gap between theoretical research and practical application, providing a platform for dialogue and the exploration of strategies that ensure the responsible and beneficial use of AI for all. We recognise that the true value of AI lies not just in revolutionary algorithms and models but in its ability to address real-world problems and improve the human experience. By bringing together perspectives from varied and diverse disciplines, we aim to foster a holistic understanding of AI's impact and emerge as a leading forum for debate, analysis and creativity. We invite you to get involved with our organisation – whether you are a seasoned academic or a curious student, this organisation is for you.

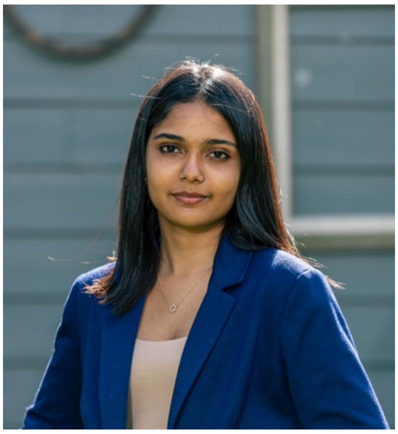
The CJAI team are committed to academic rigour and our review process reflects this. All manuscripts undergo a collaborative double-blind peer review. Each manuscript is assigned to a managing editor and two review editors, who will leave comments whilst maintaining anonymity between the author and editors. All editors that contribute to CJAI have been trained by a Nature Masterclass in scientific peer-reviewing and are registered with an Open Researcher and Contributor ID (ORCID).

I extend a huge thank you to all the editors who have helped make this journal a reality. Without their time, dedication and enthusiasm, this issue would not exist. I would also like to thank Dr. Kerry McInerney and Dr. Henry Shevlin for their support and insight.

From an interview with Dr. Claire Benn, course leader of a ground-breaking new MPhil at the University of Cambridge, to discussions on the EU AI Act, AI-generated artwork, border AI, medical AI and artificial moral agents, this issue promises to illustrate the numerous forms and potentials AI can fulfil.

We hope you enjoy reading it as much as we have enjoyed putting it together.

With best wishes,



Mahera Sarkar
Founder & Editor-in-Chief

Foreword

I am absolutely delighted to be introducing the inaugural issue of the Cambridge Journal of Artificial Intelligence (CJAI). I hope that this journal will help readers navigate the tense and complex narrative landscape around artificial intelligence (AI), machine learning (ML), and other data-driven technologies. The release of ChatGPT (GPT-4) in November 2022 sparked a widespread public reckoning with the new capabilities of AI-powered technologies. Academic debates over the accuracy and desirability of large language models (LLMs) suddenly became household, dinner-table conversations. Writers, artists and creatives grappled with the implications of these technologies for their industries and job prospects; teachers and professors were faced with new dilemmas around plagiarism and AI-assisted writing; and workers worried about the possibility of automation and job loss. Some of these conversations were, of course, part of the artificially-generated AI hype wave created and sustained by leading technology companies like Google and OpenAI to encourage the continued consumption of their products.

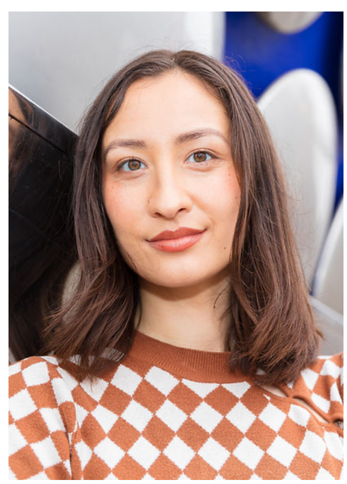
However, our debates around the development and deployment of AI applications across a range of sectors has also compelled a reassessment of these sectors themselves, sparking questions like: what is education actually for, and what do we hope to achieve through educational programmes? What is art's societal and political value? How will AI shape the future of work, and what kinds of futures do we want to strive for? How can we meaningfully address the ethical and sociopolitical impacts of AI, without feeding into false narratives about AI or locking us into a future where AI seems inevitable?

The CJAI aims to intervene in these debates. It is a richly interdisciplinary journal, drawing together insights from fields as diverse as law, philosophy, psychology, science and technology studies, computer science, politics and international relations, science communication, and sociology. This issue is no exception. From exploring explainability through the lens of the EU AI Act through to the relationship between AI generated art and the aesthetic experience, through to conceptualising consent in the medical AI context and investigating how AI applications at the border are built on systems of racial capitalism, this issue represents the wide range of crucial conversations we need to have about AI.

This interdisciplinary approach is essential for understanding not only how AI and ML systems work, but also their broader social and political implications. How technologies work cannot be divorced from their wider context, including how they are narrativised and imagined in popular culture and science fiction; how they are deployed to support or subvert existing sociopolitical systems and agendas; how they are designed to fit the needs of some users, while excluding others; who owns these technologies and profits from them; and the environmental costs of creating and using said technologies. The CJAI places ethics at the heart and soul of its research agenda, and this is borne out in the thoughtful and nuanced journals in this inaugural issue.

Furthermore, by fostering a lively student-led conversation on AI and ML, the CJAI is providing an essential and underserved forum for students to bring their innovative and interdisciplinary work to a wider audience. I am truly thrilled that the thought-provoking writing I see from students can be read and shared by a community of interlocuters who are equally interested in AI and its wider impacts.

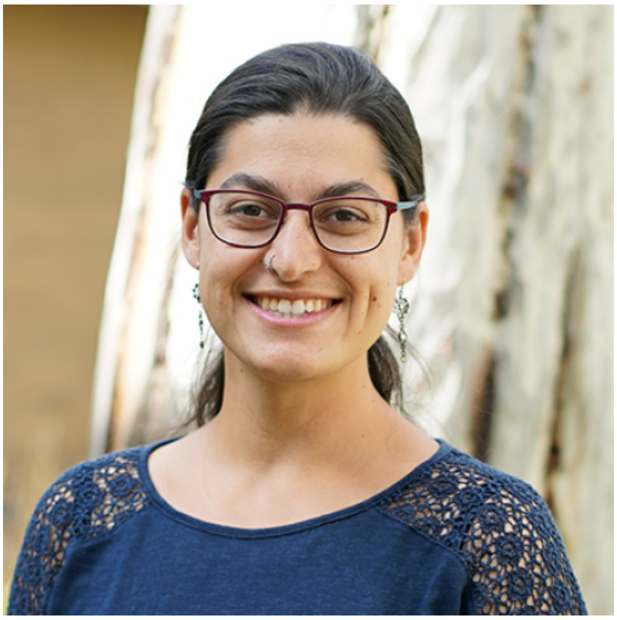
Thank you so much for your support of the CJAI, and I hope you enjoy the issue!



Dr. Kerry McInerney
Leverhulme Centre for the Future of Intelligence

An Interview with Claire Benn

Claire Benn is an Assistant Professor at the University of Cambridge and Course Leader of the MPhil in Ethics of AI, Data, and Algorithms at the Leverhulme Centre for the Future of Intelligence. In this interview, Editor-in-Chief and current MPhil student Mahera Sarkar sits down with Dr. Benn to reflect on the programme's first year.



What motivated you to create the MPhil alongside the MSt?

The primary motivation was to create a more traditional Master's experience through an intensive, in-person programme. The MSt caters to a more professionally established audience, while the MPhil aims to open this opportunity to a broader group of people who may not have as much professional experience but are still passionate about AI Ethics.

Which seminar paper was your favourite this year and why?

I made a conscious decision to attend every seminar paper. With a background in both philosophy and experimental psychology, I found it fascinating to gain new conceptual understanding, especially from John Burden's course on Evaluation of AI Systems. John did a fantastic job bridging the technical and theoretical assumptions. His seminars highlighted overlaps between his own work and my module's perspectives, and made me realise we needed a stronger technical foundation in our curriculum. As a result, John will run a technical module at the beginning of next year to introduce students to the basics of algorithms, machine learning, and AI.

As Course Leader, what challenges did you face during the MPhil's first year?

There were a few challenges during the inaugural year. As the sole leader of the course, I've always worried that I might lean too much towards my areas of expertise. Finding a balance when teaching between covering material that was not too specialised for some but also not too obvious for others was tricky. Additionally, the administrative burden was significant. Building infrastructure and anticipating how to tailor our programme while incorporating insights from other departments were crucial but challenging tasks. One of the most difficult aspects was finding a balance between being flexible and setting precedents for fairness. Ensuring individual student needs are met whilst establishing policies that are equitable is extremely important and something that I am constantly evaluating.

Was the programme's interdisciplinary approach beneficial or challenging?

Interdisciplinarity is essential. It's vital for quality research and understanding these complex issues. Exposure to different academic backgrounds only enriches discussion, especially in the course's Work in Progress seminars. It's valuable for those entering public corporations due to the breadth of perspectives. However, maintaining rigour without disciplinary bias and setting a standard for marking can be challenging. Those pursuing further study may also find it difficult to re-integrate into discipline-focused programmes. That's why we created a balance between the introductory module, which is purposefully broad, and discipline-specific electives to ensure students still get to specialise whilst gaining the skills to critique their own fields.

Did the cohort bond well?

I've been impressed and thrilled with how the cohort has bonded. Students spend time together and share their work. Support groups have organically formed, and students show genuine interest in each other's presentations.

The discussions we have had as a class have been spirited and engaging and I am so pleased that it has been driven by the students themselves.

Why is the MPhil assessed entirely through research?
There was never a debate about exams, which require a core corpus of shared knowledge. Instead, I wanted students to develop a shared vocabulary but also have the confidence and skills to produce novel specialised research. The focus on coursework allows them to build skills like project management, self-reflection, and recursive improvement. I don't want to tell people what to write about. Instead, students tell me what they're working on - the range of incredible projects have blown my mind.

What role do you see graduates of the course playing in the future development of AI?
I have high hopes for graduates of this course shaping AI discourse across multiple avenues. This could occur in public policy, private companies, or academia. I also envision a "soft influence" through a cultural shift addressing AI concerns. I often find public discourse to be pitched too high or overly sensationalised so by having those workplace or dinner-table discussions, I see this cohort as custodians and gateways for meaningful public engagement.

Looking ahead, what are your priorities and goals for the MPhil's second year?
As I said earlier, John Burden will lead a technical introductory module, and we're considering a new elective on legal, regulatory, and policy aspects. We will definitely incorporate feedback from the current cohort to refine the curriculum. I also plan to create a college family structure to connect the new cohort with past students. Finally, I am passionate about finance not being a barrier to this opportunity so hopefully we can secure more funded places for future students.

Do you have any advice for prospective applicants?
When reviewing admissions, I simply want to find people who will enjoy the course and do well. It is my responsibility as Course Leader to ensure students succeed and thrive during the course. I therefore encourage applicants to show their suitability through prior experience or enthusiasm. Given that the course is solely assessed on the basis of academic writing, applicants should demonstrate strong writing skills and thrive in an intensive, independent research environment. The level of independence that you have in this course is both an incredible opportunity and a challenge. As a result, people should be flexible but be prepared to be decisive and settle on what they want to pursue.

Any book recommendations?
I'm a huge fan of sci-fi and fantasy, and I believe in reading whatever you're passionate about – even if it is not specifically about AI. In my lectures this year, I spoke about Ursuala Le Guin's 'The Left Hand of Darkness". It has nothing to do with technology and yet it makes you reflect on what it means to be human, which is just as important as any AI-specific literature.

For further details on the MPhil in Ethics of AI, Data, and Algorithms, visit:
https://www.lcfi.ac.uk/education/mphil

AI Explainability in the EU AI Act: A Case for an NLE Approach Towards Pragmatic Explanations

Anna Nicolis and Nigel Kingsman
Hughes Hall, University of Cambridge

This paper navigates the implications of the emerging EU AI Act for artificial intelligence (AI) explainability, revealing challenges and opportunities. It reframes explainability from mere regulatory compliance with the Act to an organising principle that can drive user empowerment and compliance with broader EU regulations. The study's unique contribution lies in attempting to tackle the 'last mile' of AI explainability: conveying explanations from AI systems to users. Utilising explanatory pragmatism as the philosophical framework, it formulates pragmatic design principles for conveying "good explanations" through dialogue systems using natural language explanations. AI-powered robo-advising is used as a case study to assess the design principles, showcasing their potential benefits and limitations. The study acknowledges persisting challenges in the implementation of explainability standards and user trust, urging future researchers to empirically test the proposed principles.

Keywords: EU AI Act, Explainability, Explanatory Pragmatism, Natural Language Explanations, Robo-Advising

Introduction

This paper focuses on AI explainability under the EU AI Act, assessing its role in enhancing regulatory compliance, fostering a culture of ethical AI, and building trust with customers, thereby helping align AI practices with European societal values. Section 1 interrogates the concept of AI explainability under the EU AI Act; it is argued that the concept of explainability is tied to two key normative outcomes: user empowerment and regulatory compliance. Section 2 looks at the theoretical foundations of explanation and proposes the utility of the explanatory pragmatism framework based on a communicative view of explanation and an inferentialist view of understanding. Having laid the theoretical foundations, this section suggests five pragmatic principles of what a "good explanation" is. Section 3 introduces the concept of Natural Language Explanations (NLE) as a human-comprehensible verbalisation of the working of a model in natural language. Design principles for a dialogue system based on NLE are proposed. The final section applies the design principles for a good explanation and respective delivery mechanisms in a hypothetical Robo Advising Dialogue System (RADS) with model examples and test scenarios to evaluate the effectiveness of the design principles in practice. It is concluded that future work would comprise technical testing of LLM/NLP to help drive application-building best practices.

1. EU AI Act: Meanings of Explainability

The EU AI Act is a central component of the EU's digital single market strategy, aiming to facilitate the efficient functioning of the internal market by establishing common regulations for the development, deployment and adoption of AI-powered products and services (European Commission, 2021).

The cornerstone of the EU AI Act ("Act") is a risk-based taxonomy of AI systems; AI applications that are deemed to pose an unacceptable risk, such as state-driven social scoring, are prohibited outright. In contrast, AI systems with limited or minimal risk, such as spam filters, can enter the market with minimal new requirements. Title III constitutes the majority of the regulation, outlining prescriptive rules for so-called high-risk AI systems, which are defined as systems that pose significant threats to the health, safety or fundamental rights of natural persons.

This paper focuses on explainability, an aspect of the Act that involves not only a lack of prescriptive technical rules but also general definitional challenges, similar to its regulatory precedent constituted by the GDPR (Casey *et al.*, 2019; GDPR, 2016).

Although the Act does not contain an explicit mandate for "AI explainability" requirements, Recital 38 alludes to the concept of "explainable AI", cautioning against potential negative impacts on individuals' fundamental rights and power imbalances if AI systems lack sufficient transparency, explainability, and documentation.

The Act incorporates two essential articles that imply a degree of explanation for AI systems in use: Article 13, "Transparency and provision of information to users", and Article 14, "Human oversight."

Article 13 delineates three sub-articles with key requirements:

- Article 13(1) mandates that high-risk AI systems be designed with sufficient transparency to enable users to interpret the system's output for appropriate use. The level of "transparency" functions as a condition to ensure compliance with other obligations specified in Chapter 3 of the regulation;
- Article 13(2) prescribes that high-risk AI systems must furnish users with relevant, accessible and comprehensible instructions;
- Article 13(3) further details the characteristics of information to be included in the user instructions outlined in Article 13(2), such as the system's purpose, technical robustness, potential risks to health and safety or fundamental human rights, performance, training dataset, human oversight measures assisting users in interpreting the AI system's output and lifetime and maintenance measures.

The notion of transparency is not value-neutral and has been subject to critical scrutiny (Ananny & Crawford, 2018). The term lacks a singular definition, and in the literature on explainability, "transparency" and "explainability" are often employed somewhat interchangeably (Grady, 2022).

Hayes's (2020) notion of transparency is applied to interpret its meaning in the Act. Transparency is a condition that is conducive to knowledge acquisition about X, a phenomenon, an object or, in our case, an AI algorithm. Such a condition is defined by several properties:

- Availability: the information about X is documented;
- Accessibility: the information about X can be accessed by the agent who is looking to find it;
- Understandability: the agent is able to understand the information about X that is made available.

It is argued that, in general terms, transparency serves as an enabling principle for explainability. Without availability and access to information, generating and sharing explanations about AI systems would not be possible.

1.1 Explainability: User Empowerment and Regulatory Compliance

The rules delineated in Articles 13 and 14 of the Act offer a perspective on explainability that enhances the user's agency and establishes conditions that foster adherence to other requirements within the Act. Specifically, transparency and, subsequently, explainability function as conditions that permit the AI system user to employ the system for its intended purpose while concurrently ensuring compliance with the Act's regulatory mandates. Below, it is explained how the notion of explainability in the Act can be understood through the lenses of user empowerment and regulatory compliance.

Explainability and User Empowerment

The concept of explainability as a user-empowering condition is articulated in Article 13(1) and Recital 47 of the Act. These provisions mandate that high-risk AI systems function in a manner comprehensible to their users, providing them with sufficient information to interpret the system's operations and utilise it appropriately.

Explainability and Regulatory Compliance

The notion of explainability as a condition for compliance is articulated in Article 13(2), which prescribes information transparency requirements "with a view to achieving compliance with the relevant obligations of the user and the provider." This compliance-oriented objective is also evident in Article 11, which outlines technical documentation requirements, and Article 40, which establishes compliance as a necessary safety condition aligned with technical standards (Sovrano *et al.*, 2021).

2. Explanatory Pragmatism

This section transitions from examining the role of explainability in the EU AI Act and turns the discussion to an assessment of frameworks for implementing explainability requirements. Design principles for generating explanations and dialogue systems are proposed, drawing on the philosophical framework of "explanatory pragmatism" (Nyrup & Robinson, 2022).

Numerous theories of explanation have been developed in philosophy, which are often influenced by fields such as psychology and linguistics. Prominent theories in contemporary philosophy, including causal realism (Salmon, 1984), constructive empiricism (Van Fraassen, 1980), ordinary language philosophy (Achinstein, 1983), cognitive science (Holland *et al.*, 1989), naturalism and scientific realism (Sellars, 1962), offer distinct definitions of "explanation", sometimes in complementary ways. All theories, except for causal realism, are pragmatic, as they aim to make explanations specific and tailored for the individual recipients.

Most definitions incorporate the process of question-answering as an element of the act of explaining. This pragmatism can also be seen in the objective to adapt explanations to suit individual users, ensuring that the same explainable information is presented and reformulated in a unique manner for each user.

The features of explanation under the explanatory pragmatism framework (Nyrup & Robinson, 2022) are based on the following key aspects:

Communicative View of Explanation

Explanations are communicative acts where an explainer shares certain information with an audience to help them achieve relevant comprehension. This definition contains two key concepts. First, explanations should be regarded as speech acts (Austin, 1962) and thus evaluated based on their effectiveness in fulfilling their communicative function. Second, the primary communicative function of explanations is to facilitate the audience's understanding of the information transferred by the explainer (Franco, 2019).

Inferentialist View of Understanding

Understanding is a context-dependent concept (Kelp, 2015; Wilkenfeld, 2017). Merely acknowledging someone's ability to draw inferences is not enough to claim that the person has "understood" something. Building on this idea proposed by Nyrup and Robinson (2022), the conversation plays a crucial role in determining the class of inferences that are relevant for achieving that purpose. The role of dialogue as a means for conveying explanations is further set out in section 3.

Drawing on these two dimensions of the explanatory pragmatism framework, explainability is defined as:

> "Explainability: in the conversational context, C, a given phenomenon (model, system, prediction, ...), P, is explainable by an explainer, S, to an audience, A, to the extent S is able to convey information to A that enables A to draw inferences about P that are needed to achieve the purposes that are salient in C." (Nyrup & Robinson, 2022)

After establishing a theoretical foundation for explanation and explainability based on explanatory pragmatism, a set of design principles for implementing explanations in a practical business product context is introduced. It is attempted to translate the philosophical structure presented in the previous section into terms that product managers and business analysts can readily understand and apply when designing explainability frameworks for AI products.

2.1 Design Principles for a "Good Explanation"

An Explanation Should Be Factually Correct

The explanation should follow a robust technical framework, where the information included addresses the object of the explanation. The accuracy of the explanation will depend on the level of explanation provided. A local explanation relates to a specific output prediction from the model. A global explanation provides information relating to the higher-level workings of the system. Both need to be factually correct with respect to the operations of the AI model.

An Explanation Should Be Useful

The explanation should not only be factually accurate but also useful to the recipient of the information. The concept of utility is defined as a function of the actionable insights that are meaningful within the specific context in which the recipient of the explanation operates.

An Explanation Should Be Context-Specific

The concept of utility cannot be defined in absolute terms. Utility is always bound by the meaning that information has in a specific context. The context is described by the normative principles of a particular setting, the goals of that system and the constraints within which the user operates.

An Explanation Should Be User-Specific

The degree of utility and the context-bound nature of an explanation are always experienced from the positionality of a defined user. In the context of explainability, the main features that define user specificity are the user's degree of technical knowledge and role in relation to the working of the model (e.g. creator, user, regulator, auditor).

An Explanation Should Provide Pluralism

While an explanation should be adapted for the context and the audience, that does not mean that it should be limited to what the model is programmed to interpret as the intentions and desires of the audience. The explanation framework should begin with the objectives that are important within the given context and offer inferences that agents can utilise to achieve their objectives (Nyrup & Robinson, 2022, p. 6). This implies that the explanation allows for a diversity of normative perspectives rather than imposing a single normative stance on the intended recipient.

Table 1*: User Empowerment and Regulatory Compliance Outcomes Mapped to "Good Explanation" Principles.*

Principles	User Empowerment	Regulatory Compliance
1. An explanation should be factually correct	The information provided in the explanation can be empowering to the user only if it is correct and relevant to the product or service concerned. Incorrect information will be misleading to the user and may lead to detrimental outcomes	Information about the workings of the system must be correct in order to meet external audit and record-keeping requirements.
2. An explanation should be useful; 3. An explanation should be context-specific; 4. An explanation should be user-specific.[1]	Providing explanations that are relevant to the user and the context of use will be useful to the recipient of the explanation, allowing the user to act on the information provided and make decisions in an empowering way.	The concept of utility in relation to regulatory compliance can be described as a meta-outcome. If the information is presented in a way that is not meaningful, for example, as a disorganised collection of code and training data, it will not be deemed suitable for the purpose it is meant to serve. This principle regarding information clarity and utility is already widely adopted in financial services regulation for retail customers (ESMA, 2014).

[1] The concept of utility (principle 2) is strictly linked to the context (principle 3) and user specificity (principle 4). The utility of something is measured as a function of the outcomes delivered to a particular user in a specific context. For example, a technical explanation using scientific language and formulae will be of little utility to a layperson. Similarly, an explanation using plain English with a simplified version of the information will not benefit a technical auditor or specialist, but it will be highly relevant to a layperson.

5. An explanation should provide pluralism	By starting with the purposes that are important in the given context, the explanation can be tailored to the user's needs and preferences, empowering the user to make better decisions. Allowing for a range of normative views can also help users understand different perspectives and make more informed decisions.	The diversity of explanations is not tied to a specific regulatory outcome. However, it aligns with the mandate that the information provided to the user should be clear and not misleading, allowing users to make decisions that best suit their individual circumstances rather than prioritising the interests of the business entities supporting the AI system.

3. NLE and Dialogue Systems

After introducing the conceptual framework for a good explanation based on explanatory pragmatism, a pragmatic solution for delivering explanations from AI systems to users is set out.

The current literature on explainable AI (XAI) primarily focuses on technical solutions and challenges related to interpreting AI models and creating explanations (Cambria *et al.*, 2023). However, it is also important to consider how to deliver the explanation to non-technical users after it is generated by the model.

In particular, the paper focuses on the Natural Language Explanations (NLE) framework. This explanatory framework provides a human-comprehensible verbalisation of the working of the model in natural language. This type of explanation can be generated in different ways, ranging from deep generative models to simple template-based frameworks. It is argued that NLE has the potential to enhance the user experience and foster trust in AI systems by using familiar language and presenting information in a natural way (Paek & Horvitz, 2000).

By making information more accessible and user-friendly, NLE can improve user understanding and trust in the system. The role of NLE is examined due to the growing prominence of large language models (LLMs) as the default interface for users interacting with AI systems.

As academic resources and definitions of this approach are scarce, a generalised understanding of a "dialogue system" is used, encompassing conversational agents and chatbots (Lakkaraju *et al.*, 2022). The precedent for using dialogue systems for NLE can be traced back to expert systems, a category of symbolic AI that emerged around the mid-1960s. These expert systems were based on the principle of transferring specific human expertise into a computer. This transferred knowledge enabled the computer to offer advice as needed, similar to a human advisor, and, if necessary, to clarify the reasoning behind its suggestions. However, the application of expert systems was restricted due to various overarching AI challenges, including issues related to knowledge representation, generalisation and learning (Liao, 2005).

A dialogue system can help implement the design guidelines articulated as part of the explanatory pragmatism framework by emulating the way humans typically convey explanations and share knowledge. Rather than offering a rigid, one-directional output that the user must merely accept, a dialogue system allows users to interact with the model using their own language. A dialogue system can improve contextual comprehension and promote user trust in the system.

3.1 Design Principles for Dialogue Systems

After establishing the design principles for a good explanation, the list below sets out the design principles and components for a dialogue system focussed on delivering user explanations in a business product context. (Lakkaraju *et al.*, 2022, p. 7).

Proposed design principles for a dialogue system

1. The user should be able to prompt the dialogue system using natural language, eliminating the requirement to write any code to interact with the system.

2. The system should be able to understand ongoing user requests and associate

them with appropriate explanations to be generated and presented to the user.

3. The system should understand the context of the questions and adapt the explanations accordingly.

4. The system should build on previous inputs and clarify/rephrase anything generated by a previous prompt or prompts.

5. The system should state when it does not have an answer due to a lack of information or because it does not understand the question instead of trying to persuade the user.

6. The system should provide a confidence level of its explanation.

7. The system should be near real-time to simulate a human verbal dialogue.

3.2 Key Benefits of Dialogue Systems

By enabling users to interact with the system using natural language (1), the dialogue system reduces the need for technical expertise and is accessible to a wider range of users, reducing audience comprehension issues. The ability of the system to understand ongoing user requests and associate them with appropriate explanations (2) aids in the delivery of relevant and coherent information, addressing the challenge of audience comprehension, particularly with complex models. By understanding the context of questions and adapting explanations accordingly (3), the dialogue system helps improve domain knowledge, allowing users to make meaningful inferences in their specific fields of interest. This implies that the user is being educated with new knowledge, or at least guided towards it, which adds an innovative dimension to XAI. Instead of treating the audience as a static receiver that the XAI system needs to accommodate, the system could actively improve their comprehension and use of explanations.

By building on previous inputs and offering clarification or rephrasing when needed (4), the system promotes better understanding, addressing both semantic mapping and audience comprehension issues. The system's capability to state when it lacks information or does not understand a question (5) rather than misleading users addresses the truthfulness of information, thereby promoting trust between users and the system.

By providing a confidence level for its explanations (6), the dialogue system empowers users to assess the reliability of the information and make informed decisions, which further addresses the audience's ability to comprehend the information and its relevance to their domain knowledge. Lastly, a near-real-time dialogue system (7) simulates human verbal interactions, creating a seamless user experience.

Key conceptual components required to build the dialogue system are laid out below (Lakkaraju *et al.*, 2022, p. 2).

3.3 Core Components of a Dialogue System

1. **Graphical User Interface (GUI)**: The GUI is the visual interface designed to facilitate user interaction with the dialogue system. It should be user-friendly and intuitive, making it easy for users to enter their input via text and receive the system's output in a clear and understandable format. The GUI may incorporate elements such as text boxes, buttons or visual cues to guide users through the conversation.

2. **Natural Language Understanding (NLU)**: This component is responsible for processing and interpreting user input in natural language. NLU involves parsing the input text, identifying its structure and extracting relevant information, such as keywords, intents and entities. By doing so, the dialogue system can comprehend the user's query or request and determine the appropriate response or action.

3. **Explanation Mapping**: This component is responsible for connecting the user's input to the appropriate explanation. It may involve querying a knowledge base, database or another AI model to retrieve the necessary information to generate the explanation.

4. **Explanation Generation**: The explanation generation component is

responsible for creating human-readable explanations based on the retrieved information from the mapping component. Using NLE, it produces explanations that are both context-specific and comprehensible to the user based on the language of the initial query. The generated explanation is then returned to the user through the GUI.

In real-world applications, these capabilities can be implemented as separate components or as an end-to-end model. For instance, components 2 and 4 could be integrated within a single model, while component 3 might be part of the same model or a separate component. ChatGPT's plug-ins exemplify this approach, where a text model interacts with another system or model through model prompting (OpenAI, 2023).

4. Robo-Advising Case Study

This section evaluates the significance of explainability requirements by focussing on a specific case study within the financial services domain: robo-advising, which involves automated, algorithm-driven financial planning services in contrast to a human wealth manager. At the time of conducting this research project, robo-advising is not explicitly mentioned in Annex III of high-risk AI products, which outlines mandatory compliance with the Act's requirements for high-risk products, such as Article 13 and Article 14 (European Commission, 2021). A narrow interpretation of the regulation would imply that financial services firms offering robo-advising products are not required to address the explainability requirements described in Section 1.

However, the focus on robo-advising underscores the crux of the paper's argument: addressing explainability is not an end goal in itself tied to minimal regulatory compliance with the EU AI Act but rather an organising principle that can help achieve other normative outcomes.

Despite a boom in Fintech services, robo-advising accounts for less than 1% of the overall assets under management in the EU (ESMA, 2023). The ESMA TRV risk analysis (2023) highlights that the lack of explainability provisions offered as part of the robo-advising products impacts consumer trust. Furthermore, the report suggests that fund management companies may be discouraged from advertising their businesses' use of AI in the investment process to avoid attracting regulatory attention and potential reputational risk. Firms may also be reluctant due to their use of AI in more limited contexts (e.g. not directly affecting investment strategy) and the murky definitions surrounding it.

By going beyond minimal regulatory compliance, robo-advising firms can develop tools that future-proof their products and services as regulations continue to evolve.

Furthermore, there are systemic benefits to consider. Without consumer trust enabled by improved explainability, financial services providers may hesitate to invest in innovation related to robo-advising capabilities. Consequently, the democratising potential of such products, which aim to make financial advice more accessible and affordable, may not be fully realised (Bianchi & Briere, 2021, p. 20).

In the final section it is sought to set the foundations for a plan to empirically test AI explainability via dialogue systems using LLMs and an explanation generation mode within robo-advising.

Hypothetical dialogues were created between users and the Robo Advising Dialogue System (RADS), each demonstrating how the system could fulfil the design principles for effective explanations. These can be found in the appendix to this paper.

4.1 RADS User Vignettes

Model user vignettes have been crafted into hypothetical dialogues, assessing user empowerment and regulatory compliance to define what constitutes a good explanation.

Three distinct user profiles—a retail consumer, a data scientist, and an ESMA regulator—represent stakeholders typically interfacing with AI explanation systems. Their diverse expertise ranges from non-technical to highly specialised, and their expectations from AI systems differ correspondingly, from simple usability to stringent compliance with technical and regulatory benchmarks.

The dialogues exemplify how RADS might handle inquiries, varying from basic consumer questions about credit scores to a data scientist's technical scrutiny of prediction models and an ESMA regulator's audit on privacy metrics.

The retail consumer dialogue illustrates how RADS provides factually correct, context-specific, and user-tailored explanations without requiring technical knowledge from the user. This approach not only empowers users by demystifying financial recommendations but also aligns with regulatory standards by ensuring clarity and utility of information as mandated in financial regulations.

For the data scientist user vignette, RADS offers technical data on forecast errors, demonstrating its capacity to communicate complex information clearly and transparently, thus facilitating expert users to assess the model's performance critically. Such interactions are key to ensuring that AI systems remain under informed human supervision, a core requirement of regulatory compliance.

The dialogue with the ESMA regulator presents a contrasting scenario where RADS fails to meet the design principles. When asked to translate technical details into plain English, RADS responds with an oversimplified answer that neither empowers the regulator with actionable information nor complies with the necessary regulatory transparency and insight.

4.2 Analysis of RADS Vignettes

Each dialogue scenario is assessed against the dual criteria of user empowerment and regulatory compliance, with a scoring mechanism evaluating the "goodness" of the explanation provided. The assessment gauges whether the explanations enable users to make informed decisions (user empowerment) and whether they adhere to the transparency and information quality required by regulators (regulatory compliance).

Table 2: *Model User Vignettes Mapped Against Good Explanation Principles, User Empowerment and Regulatory Compliance Outcomes.*

Vignette	Adherence to Good Explanation Principles 1-5	Adherence to User Empowerment	Adherence to Regulatory Compliance
1. Retail consumer	• RADS converses in natural language (Principle 1) • Provides targeted • responses to each user's context (Principles 2 and 3) • Ensures a coherent, ongoing dialogue (Principle 4) • Presents pluralistic, open-ended explanations, recognising its limitations and fostering user autonomy (Principle 5)	• RADS enhances user experience with natural language responses and personalised insights, like advising on investment changes after a pay rise • It is transparent about its limits, encouraging informed user decisions • Unlike static Terms and Conditions, RADS supports financial literacy with dynamic, contextual dialogue,	• RADS meets Article 13(1) and 13(2) of the Act by providing transparent, tailored, and easily understandable information, enabling effective use of system outputs • It also aligns with the ESMA Final Report guidelines by clarifying portfolio allocation processes, exemplifying how enhanced explainability can exceed basic regulatory expectations

		fostering trust through transparency and avoiding false certainty	
2. Data scientist dialogue	• Data scientist converses with RADS in technical language • RADS effectively addresses inquiries about error forecasting (Principles 2 and 3), and • Offers options like MSE, MAE, and MAPE metrics (Principle 1) • Once a metric is chosen, it details the forecast error percentage (Principle 4), and • Provides thorough global and local explanations for troubleshooting (Principle 5)	• RADS facilitates a transparent QA process by providing the data scientist with pertinent information about forecasting error rates • It adapts to the technical language used by the user • Empowers the user to supervise the model's performance and address any shortcomings	• Dialogue demonstrates how the system can help meet the requirements under Article 14 of the Act concerning human oversight • For instance, if the average deviation rate of 12.75% from the actual value, as determined by the MAPE for the specified period, exceeds the targeted deviation rate, the user may further investigate the relevant information and the model components that may impact this calculation
3. ESMA regulatory dialogue	• An ESMA regulator audits model compliance, but RADS inadequately addresses their plain English request for privacy and fairness metrics, violating principles • Oversimplifies, lacking detail and context (violations of Principles 2 and 3) • Fails to build on previous non-technical queries (Principle 4) • Provides general explanations without meaningful insights, not admitting its own limitations (Principle 5)	• RADS dialogue lacks adaptability and detail, failing to provide information tailored to the regulator's needs • The inefficiency hampers the audit process, forcing the representative to seek further clarification elsewhere • RADS does not meet the requirements for providing accessible and specific information necessary for effective regulatory oversight.	• RADS aims to provide detailed explanations of its operations and metrics to regulators, aligning with regulatory compliance goals • However, its failure to understand the regulator's knowledge level hinders this aim • RADS does not offer the necessary access and transparency, impairing effective oversight as mandated by Article 14 of the Act

For practical application, these hypothetical dialogues serve as benchmarks for developing and refining dialogue systems. They help identify where the system performs well and where improvements are needed to align with the stated principles of effective explanation. By including these vignettes as part of a training dataset, researchers and developers can enhance the dialogue system's ability to provide more nuanced and contextually appropriate explanations.

5. Limitations

Inherent risks and limitations of the proposed approach are acknowledged. Technical challenges include LLMs' handling of complex queries, generating coherent and relevant explanations, discerning context, retaining information, ensuring truthfulness, assessing confidence levels, and delivering timely responses.

To address these, multi-disciplinary approaches should be considered. Confidence scoring mechanisms can help monitor performance; reinforcement learning from human feedback (RLHF) can adjust model accuracy; active learning could refine data handling. Regular expert reviews are necessary to evaluate and improve the system.

Beyond technical challenges, dialogue systems risk fostering overreliance and a false sense of assurance in users. There's a delicate balance between building trust in AI and preventing uncritical acceptance of AI decisions, especially in sensitive fields like financial services. Privacy remains a critical ethical concern, with the need for stringent protocols for data handling.

6. Conclusion and Future Work

This research evaluates the EU AI Act's implications for AI explainability, underscoring explainability as not only a regulatory need but as an ethical AI guide fostering user empowerment and broad regulatory compliance outcomes. This study particularly focuses on the 'last mile' problem of explainability, presenting design principles derived from explanatory pragmatism for effectively communicating AI explanations to users. These principles aim to translate academic insights into actionable strategies for businesses and product developers.

A case study on AI-driven robo-advisory services exemplifies applying these principles, suggesting benefits beyond compliance, like competitive edge and consumer trust.

Acknowledging the challenges in implementing explainability through dialogue systems and potential user overreliance, the paper calls for empirical testing of the proposed principles in practice.

Whilst the research presented in this paper proposed the seven design principles to create an NLE Dialogue System and showcase the conformance of case study examples against those design principles, it stopped short of investigating the feasibility of applying the principles in a scalable, practicable manner. Future research would comprise building on top of the rich research in the Large Language Model and Natural Language Processing space to understand whether such technologies could deliver robust real-time scoring of dialogue systems' conformance to the principles, which in turn can inform the real-time assessment of dialogue systems' ability to suitably empower their users and to adhere to regulations. Such research, if successful, could help drive best practices across a range of applications, including, but not limited to, robo-advisory dialogue systems.

Appendix

Vignette 1: Retail Customer to RADS – User Clarifications about Credit Score, Weightings and Portfolio Allocations

Retail Customer: *I have a credit score of 350, which I know is considered a high-risk profile. Does this impact my portfolio allocation?*

RADS: *Yes, your credit score has an impact on your portfolio allocation. A low credit score, such as 350, may indicate higher financial risk, and my recommendations would likely be adjusted to account for this factor. I would typically recommend a more conservative investment strategy, focussing on lower-risk assets and diversification to help address the financial risks associated with a lower credit score. It is important to note that other factors, like your financial goals, risk tolerance and investment horizon, will also play a role in creating your portfolio allocation.*

Retail Customer: *What weighting does my credit score have in relation to my annual income, £87k per year, in my portfolio allocation?*
RADS: *In determining your portfolio allocation, the system accounts for multiple factors, including your credit score, annual income, risk tolerance, financial goals and investment horizon. While I cannot provide an exact weighting for each factor, as the model dynamically adjusts based on the information provided, I can assure you that both your credit score and annual income are considered when generating recommendations.*

Your credit score may influence the risk profile of your portfolio, while your annual income can help determine your investment capacity and how aggressively or conservatively you might want to invest. The combination of these factors, along with other personal details, enables me to create a personalised and well-rounded investment strategy tailored to your specific needs and financial situation.

Retail Customer: *Next year I am expecting a 15% raise in my base salary. Can you show me how my current portfolio allocation would change on that basis?*

RADS: *Congratulations on your expected raise! I can certainly help you understand how a 15% increase in your base salary might impact your portfolio allocation.*

With a higher annual income, you may have an increased capacity for investment, which could lead to several adjustments in your portfolio allocation:

- *Increased diversification: With more funds available for investment, you might consider diversifying your portfolio across more asset classes, sectors or geographic regions.*
- *Risk tolerance adjustment: Depending on your financial goals and risk tolerance, a higher salary might enable you to take on more risk in your investments, potentially leading to a higher allocation in equities and other higher-risk assets.*
- *Rebalance based on goals: If your financial goals have changed due to your increased income, your portfolio allocation might need to be adjusted accordingly. This could involve revisiting your target allocations and ensuring they align with your updated financial objectives.*

Vignette 2: Data Scientist to RADS Error Metrics and Technical Robustness Check
Data Scientist: *Can you provide information on the forecast errors for the past seven days? I want to know how well your predictions have performed.*

RADS: *I can provide you with forecast error metrics for the past seven days. However, I need to know which specific metric you're interested in, such as mean squared error (MSE), mean absolute error (MAE) or mean absolute percentage error (MAPE).*

Data Scientist: *I want to start with MAPE for the past seven days. Can you also explain how this metric helps evaluate forecast accuracy?*

RADS: MAPE is an accuracy metric that measures the average percentage error between the actual values and the forecasted values. It is calculated by taking the average of the absolute percentage errors across all predictions. Lower MAPE values indicate better forecasting accuracy.

For the past seven days, the MAPE of our predictions is 12.75%. This means that, on average, our forecasts deviated by 12.75% from the actual values.

Data Scientist: *How reliable are these forecast error metrics in determining the overall accuracy and robustness of your system?*

RADS: *Forecast error metrics like MAPE provide valuable insights into the accuracy of predictions. However, it's important to acknowledge their limitations. For example, MAPE may be less informative when dealing with small values, as the percentage error can be disproportionately large. It's best to consider multiple error metrics to gain a comprehensive understanding of the system's overall accuracy and robustness.*

Data Scientist: *Which prediction was the largest contributor to this error?*

RADS: *The largest residual error occurred with security X. The model predicted a 5% increase, but in reality, security X saw an 11% decrease.*
Data Scientist: *Interesting. Could you provide some context as to why the model forecasted this increase for security X?*

RADS: *The model's forecast for security X was largely influenced by the lagged behaviour of security Y. Historically, there has been a tendency for increases in security X to follow increases in security Y with a one-day lag. The feature 'security_y_lag_1_day' had the highest SHAP value for this prediction, indicating its significant contribution to the model's forecast.*

Vignette 3: ESMA Official to RADS – Audit on Privacy Metrics
ESMA Representative: *How do you ensure that your AI model does not inadvertently learn any patterns or biases that could lead to privacy breaches or unfair treatment of certain clients?*

RADS: *The model is designed with fairness and privacy as key considerations. We use techniques such as differential privacy, which introduces a controlled amount of noise into the training data, to prevent the model from overfitting or memorising specific client details. Additionally, we perform regular audits of the model's outputs to identify and mitigate any potential biases or privacy concerns.*

ESMA Representative: *Can you provide some specific metric examples or parameters that you use for differential privacy to ensure the protection of client data during the model training phase?*

RADS:
import numpy as np

from diffprivlib.models import LogisticRegression

 # Set privacy budget

epsilon = np.log(2)

 # Initialise a differentially private Logistic Regression model

model = LogisticRegression(epsilon=epsilon, data_norm=1)
ESMA Representative: *I am not a technical user. Can you explain this in plain English?*

RADS: *We use complex mathematical algorithms to balance privacy and model effectiveness.*

References

Achinstein, P. (1983). *The Nature of Explanation* (1st ed.). Oxford University Press.

Austin, J. L. (1962). *How to do things with words.* Clarendon Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), pp. 973–989. https://journals.sagepub.com/doi/10.1177/1461444816676645

Bianchi, M., & Briere, M. (2021). *Augmenting investment decisions with robo-advice (Report No. 3751620).* Université Paris-Dauphine. https://doi.org/10.2139/ssrn.3751620

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1), 103111. https://www.sciencedirect.com/science/article/pii/S0306457322002126?via%3Dihu b

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines: The GDPR's right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, 34(1), p. 143. https://doi.org/10.15779/Z38M32N986

Council of the EU. (2022, December 6). Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights. https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental rights/

European Commission. (2021). Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on artificial intelligence and amending certain

Union legislative acts. https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=celex%3A52021PC0206

European Parliament. (2023). AI Act: A step closer to the first rules on Artificial Intelligence [Press release]. https://www.europarl.europa.eu/news/en/press room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial intelligence#:~:text=The%20new%20law%20promotes%20regulatory,that%20significantly%20impact%20their%20rights.

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1. https://eur lex.europa.eu/eli/reg/2016/679/oj

European Parliament and Council of the European Union. (2022). Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU (recast). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0065

European Securities and Markets Authority (ESMA). (2014). Enhanced protection for retail investors: MiFID II and MiFIR. https://www.esma.europa.eu/sites/default/files/library/2015/11/2014-726_enhanced_protection_for_retail_investors_-_mifid_ii_and_mifir.pdf

European Securities and Markets Authority (ESMA). (2022). Final report. https://www.esma.europa.eu/sites/default/files/library/esma35-43-3172_final_report_on_mifid_ii_guidelines_on_suitability.pdf

European Securities and Markets Authority (ESMA). (2023). TRV risk analysis 1 February 2023: Financial innovation – Artificial intelligence in EU securities markets. https://www.esma.europa.eu/sites/default/files/library/ESMA50-164-6247-AI_in_securities_markets.pdf

Franco, P. L. (2019). Speech act theory and the multiple aims of science. *Philosophy of Science*, 86, pp. 1005–1015. https://doi.org/10.1086/705452

Grady, P. (2022). The EU should clarify the distinction between explainability and interpretability in the AI Act, https://datainnovation.org/2022/08/the-eu-should clarify-the-distinction-between-explainability-and-interpretability-in-the-ai-act/

Hayes, P. (2020). An ethical intuitionist account of transparency of algorithms and its gradations. *Business Research*, 13(3), pp. 849–874. https://ideas.repec.org/a/spr/busres/v13y2020i3d10.1007_s40685-020-00138-6.html#refs

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P.R. (1989). *Induction: Processes of Inference, Learning, and Discovery* (1st ed.). MIT Press.

Kelp, C. (2015). Understanding phenomena. *Synthese*, 192, pp. 3799– 3816. https://doi.org/10.1007/s11229-014-0616-x

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking explainability as a dialogue: A practitioner's perspective. https://arxiv.org/abs/2202.01875

Liao, S. (2005). Expert system methodologies and applications – A decade review from 1995 to 2004. *Expert Systems with Applications*, 28, pp. 93–103. https://www.sciencedirect.com/science/article/abs/pii/S0957417404000934

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(13). https://doi.org/10.1007/s10676-022-09632-3

OpenAI. (2023). Language models can explain neurons in language models. https://openai.com/research/language-models-can-explain-neurons-in-language models

Paek, T., & Horvitz, E. (2000). Conversation as action under uncertainty. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI'00),* pp. 455- 464. San Francisco, CA.

Saeed W., & Omlin C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, 110273 https://www.sciencedirect.com/science/article/pii/S0950705123000230#section cited-by
Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World* (1st ed.). Princeton University Press.

Sellars, W. S. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Science, Perception, and Reality*, pp. 35–78. Humanities Press/Ridgeview.

Sovrano, F., Vitali, F., & Palmirani, M. (2021). Making things explainable vs explaining: Requirements and challenges under the GDPR. https://arxiv.org/abs/2110.00758

Van Fraassen, B. C. (1980). *The Scientific Image* (1st ed.). Oxford University Press.

Wilkenfeld, D. (2017). MUDdy understanding. Synthese, 194, pp. 1273–1293. https://doi.org/10.1007/s11229-015-0992-x

Artificial Creativity and the Incomplete Aesthetic Experience

Dvija Mehta
Kings College, University of Cambridge

This paper assesses the aesthetic experience provided by AI-generated visual art to assign aesthetic values to the same. Following from an experiential theory of aesthetic value, the notion of value remains inseparable from that of the experience of the aesthetic object in question. By conducting a detailed exploration of a complete, unified, and correct aesthetic experience through aesthetic judgement, I argue that AI-generated works lack the intentional relation thus providing an incomplete aesthetic experience – resulting in aesthetic values lower that than of anthropogenic works valued for a unified aesthetic experience. In doing so, the paper additionally answers recent allegations of human bias in perceiving AI art. The findings of this paper contribute to the field of computational creativity by treading on novel ground and providing a qualitative unravelling of the aesthetic values such works hold and their credibility as a tool to aid human creativity.

Keywords: Aesthetic Experience, Computational Creativity, Artificial Intelligence, Conscious Experience, Aesthetic Value, Art Bias

Introduction

With the phenomenal traction gained by new technologies such as generative models and LLMs, especially in the field of creativity,[2] it becomes increasingly important to have a nuanced understanding of their outputs. Recent quantitative (Franceschelli & Musolesi, 2022) and qualitative (Shanahan & Clarke, 2023) evaluations of AI creativity reveal valuable insights into computational creativity's abilities. Although sophisticated, these studies leave out the crucial concept of the type of value they assign to AI creativity. The notion of value remains a multi-faceted concept yet to be deeply explored in its distinctive aspects of AI creativity. Most studies (Magni, 2023) and consensus discussions highlight the social, economic, and decorative value whilst referring to the creative capacities of generative models.

On one hand, value derived from a socio-economic standpoint is appraised to AI-generated artworks; for instance: in the case of MidJourney's prize-winning artwork Théâtre D'opéra Spatial.[3] But on the other hand, it appears that these frameworks pose a threat to artists. Studies by Horton Jr *et al.* (2023), Magni (2023), and Hong (2020) indicate a prevailing bias that places higher value on anthropogenic artworks as opposed to AI generated works. When addressed as artworks, it is crucial to define the function of art in this context. The function of art encompasses the amusement, novelty, experience, and most of all aesthetic recognition, which in turn accounts for the importance of aesthetic value. Upon adopting Beardsley's theory to account for this: art with aesthetic value allows for a distinctive and complete experience through *affects* of feelings in a spectator (Dickie, 1974).[4] In referring to AI works as *art*works in the *art world*, it is imperative that they are subject to the aesthetic principles (Kant, 1914) of the art world (Danto, 1964) and subject to an analysis of their aesthetic values over and above social and economic values.[5] Aesthetic value here refers to the properties and qualities of an experience

[2] Creativity as defined by Sternberg & Lubart (1999) is "the ability to produce work that is both novel (i.e. original, unexpected) and appropriate (i.e. useful, adaptive concerning task constraints)". For comparative psychology theories of creativity see Cushen & Wiley (2012) and Weisberg (2015). Also refer to Dietrich (2004) for the neuroscience of creativity and divergent thinking. And, Berys Gaut (2010) for the importance of psychological capacities in creativity. Lastly, see Boden (2003) for historical and psychological creativity and its types that I further refer to in section 2.

[3] Awarded at the Colorado State Fair 2022; Jason Allen x MidJourney.

[4] Affects are phenomenally conscious states involving emotions and feelings. There can also be un-conscious affects, but when speaking of affects with context to an aesthetic experience, I refer to phenomenally conscious states with affects.

[5] These refer to guidelines often used to assess or create art. These principles allow for judgement, analysis, and evaluation of artworks.

that is considered valuable from an aesthetic standpoint. (Beardsley, 1969)

Whilst quantitative evaluations are crucial to machine learning frameworks, LLMs and generative models, it is equally important to have an intricate evaluation grounded in the philosophical literature of aesthetics to form a critical aesthetic judgement (Kant, 1914) of AI art.[6] In this respect, I advocate for an enquiry into the *aesthetic value* of AI-generated visual works over and above pre-existing studies on the social and economic values of AI tools. Adopting Goldman's experiential theory of aesthetic value and Beardsley's notion of a unified aesthetic experience, I conduct a critical evaluation of the relations in aesthetic properties underlying AI-generated artwork and thus reflect upon the aesthetic value it may hold.

Section 1 presents the motivations of this paper by clarifying the misunderstood nature of value when it comes to art and provides a deeper understanding of how an aesthetic experience is shaped through the aesthetic properties of an artwork. It explores the need to assign aesthetic value to AI art to answer the question of human bias in judging AI art. By providing insights into the properties and relations underlying aesthetically valuable anthropogenic artworks (Cole, 1833), argument 1 shapes the groundwork for a complete, unified, and correct aesthetic experience (AE) (Beardsley, 1969) that results in a high aesthetic value.

Section 2 highlights the properties underlying AI-generated works and presents argument 2, which posits a lack of certain cognitive properties and a particular relation that I term "the intentional relation." Grounding the evaluation in Beardsley's theory of aesthetic experience and Goldman's (2006) experiential theory of aesthetic value, it is evident that the aesthetic experience provided by AI art remains incomplete, leading to a lower aesthetic value when compared to anthropogenic works holding all necessary properties and relations of a unified AE.

Section 3 tackles the applications of such an evaluation in answering questions about using AI as a creative tool, the changes in aesthetic values when parameters like temperature are increased, and the place of AI art in the art world.

1. The Complete Aesthetic Experience and Aesthetic Value

The value an object holds depends upon the evaluation and aspect of value one scrutinises. Moreover, whilst placing some sort of value on art, one does so based on what they take the function of art to be. Upon an experiential theory, the function of art stands on the pillars of aesthetic recognition and a unified experience. When it comes to works like Théâtre D'opéra Spatial, the value seems to be derived from spectator reception: a response of people's surprise, awe, and astonishment at the rising capabilities of such systems.

1.1. The Value Problem

When I speak of "the value problem", I refer to the uncertainties in the definition of the broader term value. The values placed in AI works are heavily derived from their function as a tool to aid human users, thus resulting in social, economic, and even decorative value upon their creations. It is important to factor in the aspect of aesthetic recognition as an underlying necessity of the function of all art over and above the function of aiding human users in the case of generative models.

Recent studies (Shanahan & Clarke, 2023; Franceschelli & Musolesi, 2022) treat the concept of value as unified, without taking apart and defining the many facets of value. But when it comes to the evaluation of AI creativity in terms of the artworks it generates, it is crucial to account for aesthetic value. This area remains under-explored in the domain of AI art and I encourage this exploration for two reasons. First, with AI artworks receiving awards, it becomes wholly accepted that AI-generated works have stepped into the art world. And second, the developing field of computational creativity would benefit from a deep dive into the aesthetic principles of the art world as it

[6] Aesthetic judgement involves critically analysing the fundamental properties and qualities of an artwork.

enters an existing domain with pre-existing judgements.[7]

The question of value is further mystified by the bias that exists in perceiving AI art. Horton Jr *et al.* (2023) suggest that spectators prefer to assign a higher value to anthropogenic works as opposed to AI-created works. It is rational to ask whether this bias is warranted and grounded in existing theories or simply a social consensus. A deep dive into the aesthetic nature of AI visual arts will bring us closer to understanding *why* the bias problem exists from exploring a judgement of AI art based on its qualities rather than personal taste.

Such a question can be solved only through understanding the value of the quality of the said works. A feat possible to uncover through understanding the *aesthetic experience* of the artworks. To solve the value problem, one must single out the several distinctive types of value and conduct evaluations on each of these separately. From here on, I will focus on the *aesthetic value* of AI artworks. Art can be defined as an object with perceptual, intentional, and representational properties when regarded from an aesthetic point of view. Thus, such a point of view as the aesthetic must be adopted in valuing any such object that may be considered art, in this case AI-generated works. In evaluating the aesthetic value of AI-generated works, it is clear one can form a balanced understanding of its place in the art world, its interactions with spectators and the future of AI creativity alongside human artists.

1.2. Range of Aesthetic Value: The Intensity of an Aesthetic Experience

Assigning aesthetic values to any form of art depends upon the lens a critic chooses to look at a said work of art. Since the notion of aesthetic experience is subjective, and personal tastes may vary it is important at this stage for me to explain a conceptual difference in placing value based on taste and judgement. The concept of taste (Hume, 1757) stems from an individual's subjective preferences for what makes an artwork beautiful. Whereas aesthetic judgement is an analytical feat wherein the spectator engages in a critical analysis based on objective and subjective elements including evaluative and non-evaluative properties of an artwork. In this paper, I place aesthetic judgement on artificial creativity to thus assign aesthetic value upon the said objects. To avoid misinterpretations from several theories, I will focus on the experiential theory of aesthetic value (Goldman, 2006) combined with an analytical judgement of formal and representational properties and their relations to unify an aesthetic experience.

The aesthetic properties of an artwork underlying an aesthetic experience include the internal: formal (shape, contrast, balance), evaluative (beauty, sublimity), non-evaluative (colour, texture), representational (style: impressionism, realism), cognitive (conceptual depth, intentionality, symbolism, complexity) and phenomenal (perceptual and feeling) properties. Moreover, external properties of an artwork also play a role in the experience of it, for example, the spectator's state of mind, existing social constructs of value, beauty, and appreciation. *But when there exists, amongst all properties, certain relations that allow the properties to marry one another to give rise to an experience, such an experience would be unified.*

When looked at certain pleasing artworks, one realizes the gravity of their experience in so far as one simultaneously places aesthetic judgement on the overall experience. This placing of judgement on internal and external aesthetic properties of the artwork coupled with the overall objective and subjective aesthetic experience of the artwork leads to assigning value to the said piece of work. Some evaluative properties such as beauty and sublimity should not be confused with the aesthetic experience of the work - simply because these intrinsic properties symbolise the *quality* of the experience. There are many more properties and relations that account for a unified experience like the artwork's phenomenological and cognitive properties as mentioned above.

For instance, when prompted to list the qualities I expect in an artwork I most value, the paramount qualities encompass an evocation of

[7] The aesthetic principles include an aesthetic judgement in order to place aesthetic value upon an object.

emotion, a moving ordeal, overwhelming even, in some cases. Not merely an artwork that signifies genuine beauty, but something inexplicable that allows for contemplation, interpretation and feeling. These qualities over and above the formal properties allow for an aesthetic experience unparalleled.

This brings me to Thomas Cole's series: The Course of Empire (1883-1836). Cole's paintings marry form, structure, and colour beautifully with the symbolism of transience. Through an intricate depiction carrying realism, it is evident that his paintings hold value in terms of evaluative properties like sublimity. But the strength of communication that allows for contemplation and meaning is what transcends its formal beauty. I believe Cole combines elements of imagination, originality, and intentionality to create, what I believe to be, an extremely valuable series of art. These are the cognitive properties of artworks that additionally allow for a certain relation that I refer to as "the intentional relation", wherein the creative agent's intention to create an intense aesthetic experience in the spectator through drawing intentional links into various internal properties of his artwork. In the case of Cole's series, his intention is reflected in the poem he advertised his paintings with:

"There is the moral of all human tales;
'Tis but the same rehearsal of the past.
First freedom and then Glory – when that fails,
Wealth, vice, corruption – barbarism at last.
And History, with all her volumes vast,
Hath but one page..."

(Canto IV of Byron's Childe Harold's Pilgrimage)

The creative agent's control in intending a certain AE is as important as any other formal or representational properties of his artwork. Without the cognitive properties and the intentional relation between formal, representational, cognitive properties, the AE remains incomplete. If one were unaware of Cole's intent, one's experience of his series would be of a lesser magnitude. Although Cole does a splendid job of conveying his intention and imagination through form and colour, it is the intentional relation he develops to see the formal properties of his paintings flourish. And this in turn reflects in a higher aesthetic value derived from a rich and unified aesthetic experience. Following from this, "the notion of aesthetic value cannot come apart from the value of the experience of it" (Goldman, 2006).

Here I speak of an anthropogenic work to explain the notion of aesthetic value. This comparison with anthropogenic artworks to understand AI art is inevitable as the aesthetic principles of the art world are based on human created artworks and experiential properties.

The greatest works of art those that allow for an experience unparalleled, involving not only the perceptual faculties but also the cognitive faculties. Allowing for contemplation, appreciation, and effects of feelings.[8] Thus, giving rise to a pleasurable unified experience encompassing the entirety of the spectator's mind for that moment in time. Few, if not many, artworks have such a quality to give rise to an intense experience in the spectator, and following from the experiential theory of aesthetic value, it is then the case that such an artwork also holds a higher magnitude of aesthetic value. Beardsley provides an account of such a value as "*the aesthetic value of [an object] is the value [it] possesses in virtue of its capacity to provide aesthetic gratification when correctly and completely experienced"* (Beardsley, 1982, 27).

In Beardsley's formalist theory of aesthetic experience, he dwells upon the importance of the formal properties of artworks like colour, shape, form, texture, and their interactions with one another to give to affects of feelings that form a unified experience. His internalist theory has a few misgivings as the internal properties of an artwork are not the sole matter of judgement (Goldman, 2006). I adopt his later work (Beardsley, 1969; Dickie, 1974) that encapsulates both the internal and external properties of an artwork in creating a complete and correct AE, whilst still maintaining the notion that aesthetic value is proportional to the

[8] Beardsley's theory places value on emotional responses from art, i.e., the affects of feelings from an experience of an artwork.

intensity of the experience the aesthetic object projects. This strengthens a qualitative evaluation such as the one I conduct by critically grounding an aesthetic judgement in properties (internal and external) and their relations. Below, I provide premises to account for an artwork having a higher aesthetic value as opposed to one having a diminished value based on the AE of the artwork and its underlying properties and relations. This argument follows from Isenberg's (1949) interpretation of aesthetic judgements and his premises on how a critic's argument must be developed.

Argument 1:
P1: Artworks having F, R, r, C, I are better for having F, R, r, C, I.[9]

P2: T is such an aesthetic object having all F, R, r, C, I.[10]

C1: Thus, T is so much better for having F, R, r, C, I.

P3: W is an artwork having F, R and r but lacks C and I.

C2: Thus, W is **not** as good as T.[11]

Wherein:
F = formal, evaluative, non-evaluative properties

R = representational and phenomenological properties

r = relations between such properties

C = cognitive properties

I = intentional relations resulting from the intertwining of C with F, R and r

Applying the premises with context to a unified aesthetic experience:
P1: F, R, r, C, I are necessary and sufficient for a unified aesthetic experience (AE-u)

P2: T fulfils the necessary and sufficient criteria.

C1: Therefore, T allows for AE-u.

P3: W only fulfils F, R, r but not C and I

C2: Therefore, W does not allow for AE-u.

From the premises I present, one can say that a certain work (T) that includes an entirety of the formal, evaluative, non-evaluative, representational, phenomenological, and cognitive properties as well as relations between them, in particular the intentional relation, qualifies for a work with a higher aesthetic value as it allows for AE-u. It follows from fulfilling the necessary and sufficient criteria for such a result. All artworks do not necessitate the combination of all these criterions. But for an artwork to provide a unified aesthetic experience, the criteria above are necessary and sufficient.

Thus, Artwork T is assigned a higher aesthetic value following from a complete and unified AE, whereas artwork W is assigned a comparatively lower aesthetic value in accordance with an incomplete AE lacking C and I. The criteria stated above is necessary and sufficient in this sense for AE-u because aesthetic judgement involves critically analysing the fundamental properties and qualities of an artwork.

2. Properties and Relations in AI-generated Art

Having understood the motivations for this enquiry along with a clear image of the path to undertake it, we move forward to gauge the aesthetic experience of AI-generated works

[9] Isenberg's (1949) original argument is:
1. Artworks having p are better for having *p*.
2. *W* is an artwork having p.
3. Therefore, *W* is so much the better for having *p*.

I do not follow from his original argument as Davies (1990) and Bender (1995) present objections to it. Their objection revolves around how there cannot be one property that is good-making in all artworks. Therefore, I suggest the criteria of several necessary properties for a unified aesthetic experience, rather than one property that equates to a good piece of work. I also develop on Isenberg's original argument and present it as an argument that allows the comparison between two works T and W, as something can only be better or worse when in comparison with another.

[10] F, R, r, C, I are necessary and sufficient for a unified aesthetic experience, therefore T is so much the better for having F, R, r, C, I.

[11] W is not as good as T as it doesn't fulfil the necessary and sufficient criteria for a unified AE. Although, W does fulfil the F, R and r criteria, therefore it stills holds aesthetic value, but not as high as T.

through laying out the properties and relations present in such works.

2.1. The Aesthetic Experience of AI Artworks

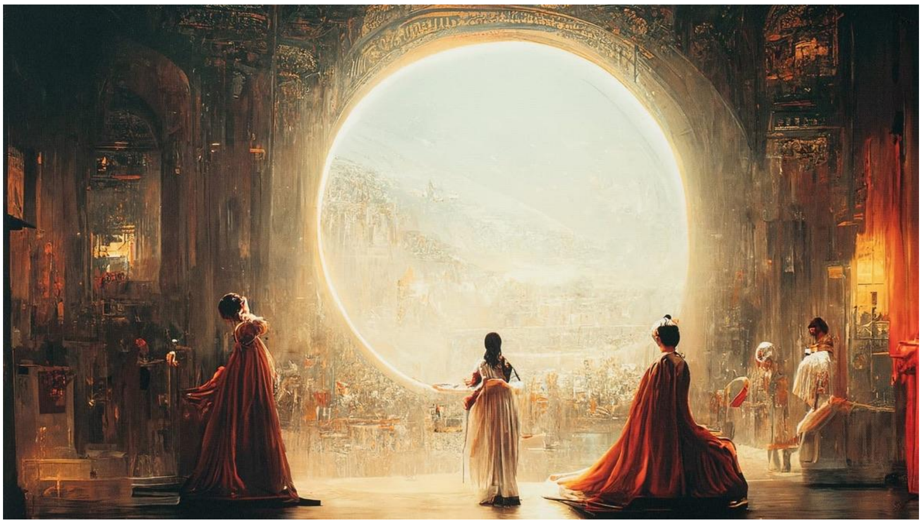


***Fig 1**: Théâtre D'opéra Spatial (TDS), MidJourney x Jason Allen*

TDS's enticing scape has caused reactions of surprise in various spectators. Its evaluative properties of beauty seem to account for a level of value. But this value raises questions of whether it should reside in the object, in this case, the digital photograph, or in the input prompt posed by Jason Allen. Some may argue, the image would not exist without the prompt, thus the value resides within Allen's will and imagination to create such an image. But this paper does not question *where* the value lies, it rather focuses on *whether* the artwork itself holds aesthetic value.

The formal perceptual properties of colour, form, and contrast reflect upon the evaluative quality of beauty and even sublimity in the experience of this work. The feelings of overwhelmingness in appreciating the beauty of this work accounts for an aesthetic experience in a spectator. Although, one cannot help but wonder what prompt Jason Allen assembled as an input for MidJourney: a facet unknown to most people. It is often the case with artworks with a high aesthetic value that the artist's intention behind creating a said aesthetic experience reflects in the spectator's reception of it. In this paper, I defend the necessity for the existence of cognitive properties reflecting a creative agent's intention to create a certain aesthetic experience and the intentional relations a creative agent supposes to exist within the properties of the artwork. Leaving that out may still account for an artwork to have aesthetic value, although the aesthetic experience of it remains incomplete. As mentioned above in Section 1.2, the range of aesthetic value of an artwork is linearly proportional to the intensity of the aesthetic experience that it allows for.

Below, I further explain why certain cognitive properties "z" and certain relations "I", i.e., intentional relations, remain unattainable in AI-generated works and how this affects the overall aesthetic experience of W.

2.2. Cognitive properties and intentional relations: the lack thereof in AI Creativity

The intensity of experience is not only dependent on the intrinsic formal properties of the artwork but also the intention of the creative agent to envision a certain aesthetic experience for the viewer. There are certain artworks that lack the artist's intention. But in the experiential view, for one to have a unified experience, one must consider all the properties attached to the work of art. And the properties associated with the creative agent are as important as any other. These properties allow for the relations between all other properties to blossom. The addition of these properties allows for a high magnitude of the AE, thus allowing for a higher aesthetic value overall. Below I highlight 3 such cognitive properties that are associated with the creative agent in artworks with high aesthetic values.

Intention

Although a contested notion that intention is crucial for valuable artworks, if one assumes the experiential theory of aesthetic value as I do here, it becomes imperative to address the cognitive property of intention. For two main reasons: one being that the creative agent's intention to create a unified AE in the spectator allows for the birth of the intentional relation between the formal, representational, phenomenal, and cognitive properties of the artwork. And second, the inclusion of intention as an appealing property for a unified AE does not mean that we engage in the intentional fallacy because the view I defend places aesthetic judgement based on several properties over and above intention. Going forward from there, as Cole's example in section 1.2 shows, the intention of the creative agent does contribute to the overall aesthetic experience of the spectator. In many cases, like intentionalism in aesthetics (Livingston, 1998), an artist's intent plays a heavy role in the aesthetic judgement placed upon the work.[12]

Imagination

To arrive at something potentially novel with value – upon the definition of creativity – one must imagine. One must visualise, feel or think of their creation. Take for example scientists and philosophers that imagine their theory to understand whether it is cohesive. The process of imagination may occur before the creative output is produced by the creative agent. And it is an elementary step. But also note that the scientist additionally has an intention to imagine the theory in question. And to imagine, one also has mental representations corresponding to intentionality. Imagination allows one to create intentionally or spontaneously. Therefore, the cognitive properties of intention, imagination and novelty all intertwine with unique relations and amount to value in creativity. Some have gone so far to argue that imagination is the epitome of higher levels of creativity. Birds and Hills (2019) suggest that "imagination is the cognitive source of genuine creativity" and suggest that imagination is an ability to produce a particular type of mental representation.

iii. Novelty and Originality

Amongst various theories of creativity, the criteria that remain prevalent are novelty and value. Although recent work by Bird and Hills (2019) suggests the importance of originality and rejects value as a crucial criterion. When one talks of novelty and originality, one refers to the production of something authentic. One may argue that nothing can ever be novel as most ideas may pre-exist, especially words and art and poetry. This argument also goes so far as to say that most art is inspired from pre-existing art, thus leaving out novelty. To this I respond that the criteria here is *potentially* novel and original insofar as the creative agent in question recognises its originality or potential novelty. Artworks are then valued for the accompaniment of the criteria I list in argument 1 wherein novelty is displayed in the way the existing properties within the work are married.

To achieve an intense and unified aesthetic experience yielding the highest of aesthetic values, an artwork must carry (i), (ii) and (iii). To carry these 3 crucial cognitive properties, the creative agent (Ca) must have:

a) An understanding of existing domains and the means to explore the same.
b) A deep imagination that allows for transformational insight.[13]
c) A recognition of originality and value from the Ca herself.

a, b and c can require several other prerequisites including experience, value, and novelty but they are set in cognitive processing and thus also form the cognitive properties of an artwork.

When a creative agent exercises such properties to bring to life their intended aesthetic experience in the spectator, they in turn birth "*intentional relations"* amongst formal and representational properties internal to the work. In so far as these intentional relations

[12] Intentionalism in aesthetics encompasses a broad range of views including types of intentionalism: fictionalist, moderate, extreme, hypothetical and anti-intentionalism. See Livingston (1998).

[13] Boden's transformational creativity occurs when the constraints of an existing domain are altered, or completely redefined. (Boden, 1998).

persist between properties, they unify an aesthetic experience to allow for a complete and correct aesthetic experience that is intense. Thus, leading to a higher range of aesthetic value placed on the said aesthetic object.

But can such properties as cognitive and intentional relations exist in artificial systems? Below, I present my second argument that allows one to draw the conclusion that such cognitive properties and intentional relations cannot be produced by AI frameworks as these tasks are set in cognitive processing and consciousness.

Argument 2:
P1: Artificial creative intelligence in current robust generative models displays *x* and *y*, but lacks *z*.

Wherein:
x = ability to coalesce existing information in a conceptual space to create potentially novel outputs.[14]

y = ability to engage and explore structured conceptual domains to generate expected and unexpected outputs.[15]

z = transforming a conceptual space to create a new one entirely, i.e., creating of its own volition and intentionality; self-awareness; creative cognition; the ability to draw from its inner life of subjective experiences to create novel outputs.

P2: Generative Artificial Intelligence (GAI) can display x and y because these tasks are set in intelligence and computation, which requires the right computation and syntax to allow the system to create a desirable output based on its training data set.

C1: Thus, GAI displays creative intelligence in this sense as associated with *x* and *y.*

P3*: But to have the ability to perform a task related to *z* would require *understanding, imagination* as well as subjective experience over and above intelligence and computation.

C2: Thus, GAI cannot perform tasks associated with *z as they are set in cognitive processing and consciousness.*[16]

Creative intelligence refers to a specific type of intelligence that contributes to tasks involving creativity, like problem solving, generating new ideas, creating art and so on.[17] In this paper the central enquiry remains art and creativity, wherein creative intelligence is central. And this also takes centre stage in the field of computational creativity. Therefore, going forward from there, I list down the capacities of generative models in context with creative intelligence.

In premise 1, I highlight the creative capacities of generative models. I then suggest that these models can perform tasks associated with x and y; the distinctions I make in x and y rise from Boden's types of creativity wherein x corresponds to combinational p-creativity and y suggests an overlap in combinational and exploratory p- creativity. So, tasks in x, for instance, could be providing answers to certain enquiry-based questions which entails combining existing information from datasets to form the right output. Tasks in x could also be instances of writing potentially novel sentences, engaging in conversation or even writing an

[14] Also, Boden's combinational p-creativity; Tasks like forming sentences, e.g.: ChatGPT engaging in potentially novel conversations.
[15] Also, Boden's exploratory p-creativity; Tasks like creating potentially novel images from prompts; e.g.: DALLE 3/Midjourney creating art.
[16] Also note that there exist viable theories of creativity that focus on unconscious states rather than conscious. See Jung (1966), Freud (1908), Wallas (1926) and Fromm (1959). Such theories could have potential for applicability in cases of generative models if one does not want to engage in the AI consciousness debate. But it remains crucial for me to engage in the importance of consciousness and cognition in the cognitive properties of an artwork, especially in context with the intentional relation. Also note that intentionality is a fundamental aspect of consciousness, and refers to the capacity of mental states to represent things, properties, and states of affairs. Therefore, if an individual has intentional states, then it is to say that the individual has mental representations.
[17] Sternberg's (1985) Triarchic theory of intelligence outlines creative, practical and analytical intelligence wherein he defines creative intelligence as "the capacity to deal with novel situations and to generate new ideas" According to him this aspect of intelligence involves "using existing knowledge to handle new problems and cope in new situations, and it is integral to innovation and problem-solving in complex, dynamic environments."

analysis essay. Now coming to y, these tasks include the exploring and combining of information to produce expected or unexpected outputs, like creating images or artwork based on input prompts, writing poetry and short stories, or even creating marketing campaigns. These tasks involve exploring in an existing domain and combining relevant elements to produce a desired result that is potentially novel. Jason Allen and Midjourney's Théâtre D'opéra Spatial is an example of y. When it comes to aesthetically evaluating it, this artwork has beauty due to the existence of formal, evaluative, non-evaluative and representational properties. But does it have high aesthetic value? In what follows, I will explain this through the conclusion of argument 2.

Coming to z, which mainly highlights Boden's transformational creativity. But, for any such task, I believe, it would be a culmination of all three types that Boden talks about. When it comes to transformational creativity, this process encompasses creating an entire new domain that does not exist. In order to exercise such creativity, one requires first: a *true* understanding of existing domains, second: an imagination to envision abstractions and third: the ability to self-represent certain subjective experiences to then birth them into artworks.

Therefore, premises 1 and 2 lead to conclude that generative models can perform tasks x and y as these tasks are set in access of information and in particular, they are associated with the computation of a goal-oriented tasks that involve intelligence and pre-existing algorithmic syntax. But, tasks in **z** seem to be over and above the computation of information. This leads me to the conclusion at hand: the said criteria for z tasks are set in cognitive processes and conscious experiences. Because z requires the ability to draw on one's inner life of subjective experiences, create self-referential representations and understanding. Therefore, there must be *something it is like* for the creative agent to be themselves in order to have self-referential representations, as well as in inner-life of subjective experiences in order to exercise transformational creativity.[18]

It is now clear that the properties thus associated with cognitive processes and conscious experiences that allow an aesthetic object T to have 'intentional relations' is missing in AI-generated works W. Circling back to argument 1, the lack of C and I in W would make it less appreciative due to an incomplete aesthetic experience. Leading to a diminished aesthetic value. Thus, AI artworks seem to have lower aesthetic value derived from an incomplete aesthetic experience due to the lack of certain properties (cognitive) and relations that contribute to a unified aesthetic experience.

3. AI Art, The Incomplete Aesthetic Experience and More

Below I address the applications of this evaluation in understanding how aesthetic values affect the use of AI as a tool for creation, and its place in the art world. Moreover, I highlight further objections and expand on the bias problem.

3.1. AI as a Tool for Creation

Since AI models are not capable of creating through volition and only aid human artists as a tool to create, it is important to factor in the cognitive properties coming from the human agent that provides the prompt. This inclusion of certain cognitive properties enhances the aesthetic experience. An important objection to my arguments rises here when one uses AI as a tool for creation in such a way that the human agent provides his envisioned image through a prompt allowing for cognitive properties congruent to his imagination and intention to come to fruition in the artwork. My response here still maintains my initial premise that AI-generated works provide an incomplete aesthetic experience due to the lack of certain properties or relations. Let's take for example that we, as spectators, did have knowledge of the prompt Jason Allen used and thus, to an

[18] Phenomenal consciousness refers to there being something it is like (Nagel, 1974) to be the subject in question. Phenomenal consciousness is subjective experience and certain cognitive tasks are intertwined with phenomenal consciousness in the sense that there is something it is like for the subject to understand their own creation, i.e., have a subjective experience of understanding their artwork. This also adheres to intentions as a cognitive property. There is also something it is like for the creative agent to have a certain intention they aim to fulfil by means of creating relations between formal, representational, and cognitive properties of their artwork. Therefore, the cognitive properties of imagination, intention and ability to recognise the novelty of one's work is intertwined in conscious experience.

extent, the aesthetic experience included cognitive properties of Allen's intention to create his envisioned image. But what remains lacking here is the *intentional relation*: a relation birthed from the artist's intention to marry each property to another in a specific way to give rise to a unified experience. Whilst using AI as a tool to create, the autonomy and control to exercise the intentional relation cannot still reside with the human prompter. This is because the authority to marry properties then resides with the tool. The system unintentionally allows various formal and representational properties to interact to form an artwork, no longer allowing the intentional relation to exist.

How this does affect the overall aesthetic value of the artwork still follows from Argument 1. Even if W happens to have C, without "I" the aesthetic experience would remain incomplete and not as unified as T which encompasses the intentional relation.

Another important question that presents itself here is would it be worth for an artist to lose the intentional relation in order to use AI as a tool? In adopting the experiential view of aesthetic value, the loss of the intentional relation would only lead to a diminished aesthetic experience, further questioning the value of collaborations between human artists and AI and the resultant works of art. Another interesting caveat here is that of how an incomplete aesthetic experience (therefore, a lower aesthetic value) would inform other dimensions of value of a said artwork, like the socio-economic aspects. This allows me to highlight two important research domains: one, the reliance of different dimensions of value on each other, for example, how may aesthetic value inform socio-economic or decorative value. I would voice here that aesthetic value can in turn inform spectator reception merely based on the aesthetic experience the spectator undergoes, thus indirectly influencing the other dimensions of value. But this is also dependent upon societal outlook and society reception of certain works. Take the case of generative models like DALL E 2, the surprise and awe such a system inspired in users set aside the evaluation of the aesthetic simply because of its socio-economic value. Alas, it is situations like these that inspire the arguments I present here, especially due to the lack of a deep-dive into the aesthetics of AI-generated works. And second, further research studies in computational creativity, especially in the domain of how a system marries certain properties in the visual works it generates is warranted here.

3.2. The Spectator: Existing Bias in Perceiving and Valuing AI Art

In this paper, an important problem I avoid is the intentional fallacy (Wimsatt & Beardsley, 1946) which follows from placing sole value on an artwork based on the artist's intention to create. The arguments in this paper evade this by the inclusion of several other formal, internal, and external properties whilst accounting for the overall aesthetic experience. Because it is important to note that in a unified aesthetic experience, there also is a presence of external properties over and above the internal properties we have so far spoken of. One such external property crucial to the evaluation of AI-generated works is the bias Horton Jr *et al.* (2023) present that is a source of resistance in placing equal value to AI works as anthropogenic works.

Whilst having an aesthetic experience of an aesthetic object, or artwork, the spectator is also subjected to external properties including phenomenological feelings of a subjective nature not related to the aesthetic object that plays into the experience. As well as pre-existing social, environmental, and political constructs and consensus that form beliefs. One such external property that is pivotal during an aesthetic experience of an AI-generated work is bias. (Schmitt, 2020) suggests that preferential treatment towards humankind contributes to the anthropocentric view that places more value on human-created art and contributes to the bias when it comes to viewing, evaluating, and valuing AI-generated art. It is valid that such a bias will surely influence consumer and spectator tastes and beliefs, and thus affect the process of placing value on AI-generated works.

My response here is twofold:

The value I talk of in this paper is not a social, political, or economic value assigned to objects for their contribution to a resulting gain. But instead, I focus on aesthetic values that are not measured quantitatively or through a trade-off that comes with dependent variables. Aesthetic value is of a subjective nature, based on a nuanced understanding of the quality of the aesthetic experience of the artwork. It cannot be reduced to a utilitarian notion wherein common belief shapes value. It is wholly dependent on the aesthetic experience the object allows for. Aesthetic value is not dependent on feelings of pleasure or displeasure, but rather on feelings of affects (Beardsley, 1969) from the experience.[19] So, a feeling of displeasure from a pre-existing bias cannot contribute to the aesthetic judgement but rather forms a part of personal taste. Coming to the key difference here that I believe is pivotal. The assigning of aesthetic value is not through tastes or beliefs, it is instead based on an aesthetic judgement; the type of judgement that follows from a theoretical framework grounded in literature. Thus, this enquiry is not clouded by ideas of pre-existing bias against AI art, and I do so by providing valid arguments for why the aesthetic experience is incomplete in AI works.

Moving onto the other side of the bias coin: A pre-conceived societal notion is, in fact, an external aesthetic property that may linger in the mind of the spectator during the aesthetic experience, unbeknownst. This may affect the *affects of feelings* (Dickie, 1974) that rise within a spectator and thus form a different aesthetic attitude towards AI-generated art as opposed to human artworks. Regardless, whilst making an aesthetic judgement, one must make an informed evaluation considering the unity of experience. And in this case, the external property of preference to human art is evident and cannot be separated from the unity of the aesthetic experience. The reasons for placing a lower aesthetic value, in this paper, are grounded in an experiential theory of aesthetic value rooted in the necessity of certain properties (Argument 1).

3.3. Temperature and Chaos:

Parameters of temperature, or chaos, only complicate the value problem and allude to referring to creativity as something that can be controlled with adjusting the uncertainty of responses. Shanahan and Clarke (2023) present examples of altering the temperature in large language models to gauge the literary creativity it displays.

Understanding whether the aesthetic value of certain AI-generated works would vary if chaos/temperature was changed is vital. Increasing the chaos on a system like Midjourney would result in a much more creative output derived from uncertainty and surprise. Now, this could alter certain aspects of value but when it comes to aesthetic value: the previous argument still holds. Without a unified aesthetic experience, one can only assign a certain amount of aesthetic value. Let's call it φ. So φ is the aesthetic value of an artwork that does not have a unified aesthetic experience but still follows the aesthetic principles. Now, if we increase or decrease chaos, this could affect individual aesthetic properties like colour, structure, representation and so on. Reducing temperature would result in $0 \leq \varphi<1$ and increasing would result in $0 < \varphi \leq 1$. Whilst maintaining that the φ if an artwork providing a unified aesthetic experience would be $\varphi+n$ depending on the intensity of the experience. Therefore, φ remains the domain in which the aesthetic experience remains limited when speaking of generative models, due to the lack of the intentional relation.

3.4. AI Art in the Art World

The place of AI art in the art world remains open to interpretation. The enquiry I present of the aesthetic value of AI-generated works merely provides a starting point for artists to truly engage in AI accompanied creation. Although the importance of human artists in the art world is evident through this evaluation, this paper does not contribute to the existing bias. It simply presents a case to understand the place if AI art in the art world through an aesthetic point of view. But with increasing collaborations as well as human creativity in prompt generation, the place of AI in the art world is one that remains everchanging.

[19] Note that I do **not** adopt aesthetic hedonism. This is a view where aesthetic value is value because things having it give pleasure when experienced. In the view I depend, the unified experience is not reliant on pleasure but on a complete experience, which may have affects, any affects. The affects needn't be solely pleasurable but could also be paradoxical in nature. Beardsley's earlier work defended aesthetic hedonism, but my arguments merely draw upon his views.

Conclusion

In placing aesthetic judgment on the overall aesthetic experience of AI-generated artworks through an enquiry into its aesthetic properties and relations, this paper concludes that AI-generated works do have aesthetic value although not as high as those of artworks that provide a unified aesthetic experience. This conclusion is reached from understanding the lack of "intentional relations" in AI-generated works that form gaps in the aesthetic experience of AI art resulting in an aesthetic value lower than a work that allows for a unified aesthetic experience. The paper also suggests future research scope in computational creativity with regards to how diffusion models marry other properties within a visual work, as well as scope in understanding how the aesthetic values influence other dimensions of value like social and economic when it comes to the place of AI art in the art world.

References

Baumeister, R. F., Schmeichel, B. J., & DeWall, N. C. (2014). Creativity and consciousness: Evidence from psychology experiments. In P. S. Elliott & S. B. Kaufman (Eds.), *The philosophy of creativity: New essays* (pp. 185–198). Oxford Academic.

Beardsley, M. C. (1969). Aesthetic experience regained. *The Journal of Aesthetics and Art Criticism, 28*(1), 3-11.

Beardsley, M. C. (1970). The aesthetic point of view. *Metaphilosophy, 1*(1), 39-58.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*(1-2), 347-356.

Boden, M. A. (2003). *The creative mind: Myths and mechanisms*. Routledge.

Carroll, N. (1984). Hume's standard of taste. *The Journal of Aesthetics and Art Criticism, 43*(2), 181-194.

Cushen, P. J., & Wiley, J. (2012). Cues to solution, restructuring patterns, and reports of insight in creative problem solving. *Consciousness and Cognition: An International Journal, 21(3)*, 1166-1175.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486-492.

Dickie, G. (1974). Beardsley's theory of aesthetic experience. *The Journal of Aesthetic Education, 8*(2), 13-23.

Dietrich, A. (2004). The cognitive neuroscience of creativity. *Theoretical And Review Articles, 11*, 1011-1026.

Franceschelli, G., & Musolesi, M. (2022). DeepCreativity: Measuring creativity with deep learning techniques. *Intelligenza Artificiale, 16*(2), 151-163.

Freud, S. (1908). *Creative Writers and Day-Dreaming*. In Standard Edition of the Complete Psychological Works of Sigmund Freud (Vol. 9, pp. 143-153).

Fromm, E. (1959). *The Creative Attitude*. In H. H. Anderson (Ed.), *Creativity and Its Cultivation* (pp. 44-54). Harper & Row.

Gaut, B. (2010). The Philosophy of Creativity. *Philosophy Compass, 5(12)*, 1034-1046.

Goldman, A. H. (2006). The experiential account of aesthetic value. *The Journal of Aesthetics and Art Criticism, 64*(3), 333-342.

Hong, J.-W. (2020). Bias in perception of art produced by artificial intelligence. *Human-Computer Interaction*.

Horton Jr, C. B., White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. *Scientific Reports, 13*.

Jung, C. G. (1966). *The Spirit in Man, Art, and Literature*. Princeton University Press.

Kant, I. (1914). *Critique of judgement*. (Original work published 1790).

Livingston, P. (1998). Intentionalism in aesthetics. *New Literary History, 29*(4), 831-846.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review 83*, 435-450.

Shanahan, M. & Clarke, C. (2023, November). Evaluating large language model creativity from a literary perspective. *arXiv*. https://arxiv.org/abs/2312.03746

Shelley, J. (2022, Spring). The concept of the aesthetic. In *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2022/entries/aesthetic-concept/

Shevlin, H., Halina, M., Vold, K., & Crosby, M. (2019). The limits of machine intelligence. *EMBO Reports, 20*(49177).

Shevlin, H. (2020). Rethinking creative intelligence: comparative psychology and the concept of creativity. *General Philosophy of Science 11, 16*.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge University Press.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In R. J. Sternberg, *Handbook of Creativity.* Cambridge: Cambridge University Press.

Wallas, G. (1926). *The Art of Thought*. Harcourt, Brace and Company.

Wimsatt, W. K., & Beardsley, M. C. (1946). The intentional fallacy. *The Sewanee Review, 54*(3), 468-488.

Quantifying Bodies, Categorising Difference: Border AI Through the Lens of Racial Capitalism

Cherry Benson
Lucy Cavendish College, University of Cambridge

This paper explores the connection between racial capitalism and the development and deployment of AI technologies, using border AI as an illustrative example. Section 1, examines how racial capitalism, rooted in historical hierarchies and discrimination, influences the development and deployment of AI technologies. It highlights how this legacy perpetuates inequalities, privileging certain groups while disadvantaging others. Section 2 frames border AI, highlighting both its benefits and challenges. This section sets the stage for understanding how border AI can perpetuate existing inequalities and raise significant human rights concerns. Section 3 presents an analysis stemming from the ideas presented in Sections 1 and 2. Tracing the historical roots of AI technologies in border control, it highlights how pseudo-scientific racist ideologies and biometric quantification practices have shaped their foundations. Section 4 explores algorithmic accountability at EU borders and examines the EU Artificial Intelligence Act, revealing significant gaps in migrant protection. Although automating decision-making processes offers potential benefits, these systems often reinforce existing biases and lack transparency, complicating oversight and judicial review. The paper concludes by drawing upon the insights gleaned from the exploration and advocates for a shift towards a person-centred framework at the border that acknowledges and incorporates marginalised knowledge systems. This approach underscores the necessity for border control practices to prioritise human rights and dignity over technical progress and efficiency, paving the way for a more equitable future in AI deployment.

Keywords: Border AI, Racial Capitalism, AI Ethics, Algorithmic Bias, EU AI Act

Introduction

Before continuing, I would like to clarify my use of racial terms within this paper. "Race," as I understand it, originated from the colonial belief that white Western men were superior to all other human beings (Wynter, 2003). Following the words of scholars such as Patricia Hill Collins (1990), Paul Gilroy (1993) and Ruha Benjamin (2019), I understand "race" to be a pseudo-scientific social construct that stratifies people and organizes society based on physical traits like skin colour, hair texture, and eye colour.

When discussing racialised groups, I align with the rationale of scholars such as Bonilla-Silva (2003) and Collins (2000), using the terms "racially minoritised" and "racialised individual" to describe those identified as non-white. Racism constructs "race" by categorising racially minoritised individuals as the 'Other,' highlighting the socially constructed nature of racial categories and leading to systemic marginalisation that infiltrates and distorts societal structures, realities, and institutions (Omi & Winant, 2015; Bonilla-Silva, 2003).

To briefly define my understanding of 'Racial capitalism', it is a theoretical framework describing the intrinsic link between capitalism and racial exploitation. Originating from Cedric J. Robinson's "Black Marxism: The Making of the Black Radical Tradition" (2000), it argues that capitalism relies on and perpetuates racial inequalities, having developed through racial differentiation and subjugation. In critical race theory, racial capitalism is seen as the extraction of economic value from racialised bodies via exclusion, exploitation, and marginalisation. This perspective challenges the idea that racism is a mere aberration within capitalism, asserting instead that it is fundamental to the system's functioning. Scholars like Ruth Wilson Gilmore emphasise how racial capitalism perpetuates and exacerbates inequalities by continually producing racialised subjects deemed exploitable and disposable (Robinson, 2000; Gilmore, 2007).

In this paper, I will argue that historical racism is deeply embedded in AI technologies, using the case of border AI to illustrate this point. The analysis will demonstrate how these biases are perpetuated in modern systems, reinforcing existing power dynamics and discrimination. I will provide an overview of current AI technologies employed at borders, highlighting their inherent issues and biases. Due to the constraints of this paper, a comprehensive analysis of all technologies is not feasible. Instead, the focus will be on selected examples that illustrate the broader trend. I will also address counterarguments, evaluating the purported neutrality and efficiency of AI in border control. The research will specifically focus on EU policies, with a detailed examination of the EU AI Act, to contextualise the discussion within a regulatory framework. Through this focused analysis, the paper aims to shed light on the enduring impact of historical prejudices on present-day AI applications and advocate for more equitable and transparent technological practices.

1. New Technologies, Old Hierarchies

Capitalism emerged from and relied on racial hierarchies and discrimination to enable exploitation and accumulation. It did not precede racism, but rather colonial slavery, genocide, and dispossession - made possible by categorising differences by race, laid the groundwork (Robinson, 2005). Racial capitalism amplifies inequalities that naturalise the unjust distribution of resources, power and privilege. It reinforces the disposability and deviancy of certain groups, denying them rights and resources. Within this system, privileges of movement, labour access, and social safety nets are structurally conferred to some bodies while denied to those marked as threats.

Racial capitalism, with its pervasive influence on society, extends its legacy of discrimination into various technological realms, including AI. Despite the portrayal of AI systems as progressive tools (Eubanks, 2018), focused on optimisation and progress, they inherit assumptions from racial capitalism regarding the prioritisation of certain capabilities and whose interests technology should serve (Benjamin, 2020). The claimed technical neutrality of AI systems becomes questionable as they encode prejudice through unexamined design choices, a phenomenon labelled by Ruha Benjamin (2020) as "the new Jim Code" – a covert manifestation of racial bias masked by rhetoric promoting diversity, inclusion, and fairness. Contemporary racial capitalism persists by categorising groups based on their exploitation potential, keeping racialised individuals vulnerable to exploitative cycles of capitalism due to the enduring logic of white supremacy (Melamed, 2015). Racial capitalism provides a framework to examine the historical, present and future manifestations of racial inequity, offering insight into past discriminatory patterns encoded in present AI systems.

Examining the entire spectrum of the development and deployment of AI technology through this lens reveals that racially minoritised individuals bear the adverse impacts of these systems. Mathematical concepts crucial to AI development, such as statistics, were notably influenced by the work of Eugenicists like Galton, Pearson, and Fisher (UCL, 2021), which will further be explored in section 3. The representations of politics and white power are evident in the collection and storage of archival data, leading to the datafication of populations (Yale, 2015). Furthermore, the recent extraction of natural materials and resources essential for AI components disproportionately affects global South populations through exploitative labour and the depletion of local resources and environmental landscapes (Crawford, 2021). Building on these disparities is the unregulated and invisible work of data labourers responsible for maintaining the artificial intelligence data pipeline (Muldoon *et al.,* 2024). The application of AI technologies in societal domains, such as healthcare (Benjamin, 2019) and policing (Can You Make AI Fairer than a Judge? 2019), perpetuates discrimination against racialised individuals due to encoded biases in training data. Marginalised populations do not reap the benefits of systems that have a bias against them so deeply embedded.

To comprehend how systemic biases emerge at the intersection of technology and migration control, we now examine the specific use of AI in border management.

2. Framing Border AI

The concept of borders has existed for millennia, but with the advent of AI, surveilling and policing vulnerable communities at the border has become more efficient. Over the past two decades, borders have become critical zones for managing risk in Western societies (Hall *et al.,* 2021). Advanced security technologies, such as surveillance, data collection, predictive analytics, and physical barriers, have reinforced these borders and increased security (Alam *et al.,* 2023). These measures aim to manage risks and demonstrate to populations that governments are "doing something" regarding migration (Vallet and David, 2012; Hall & Clapton, 2021).

Border AI systems utilising biometrics are increasingly deployed in border control to algorithmically identify migrants deemed high-risk by analysing both bodily features and behavioural traits (Fors & Meissner, 2022). Biometrics encompasses fingerprints, iris and retinal scans, facial recognition, vein and blood vessel patterns and gait (European Parliament Directorate General for Parliamentary Research Services., 2021). Some experimental applications include AI-powered lie detectors determining truthfulness at the border through dubious emotion recognition (Lomas, 2022) and micro expression analysis (Foundation, 2021). DNA-based biometrics, measuring the living body, are also gaining prominence (Browne, 2015).

The use of new technologies, particularly automated decision-making systems, can streamline processes for public administrations and some applicants. Despite the benefit of increased efficiency for some applicants, most technologies primarily serve state authorities rather than migrants, asylum seekers, or refugees, whose interests are often overlooked in design and implementation (Ozkul, 2023). To highlight one noteworthy case where the needs of migrants have been included and centred in its design – Latvia introduced speech recognition tools to assist individuals in preparing for their citizenship applications. This self-test tool allows potential applicants to test their speech and knowledge in preparation for citizen tests. According to a 2019 survey conducted by the OCMA, a significant reason non-Latvians were not applying for citizenship was their fear of failing the tests required by the Citizenship Law. Thus this initiative addresses the needs of migrants directly (OCMA, 2021). The initiatives that include migrants in their design are primarily driven by grassroots efforts, often involving collaboration with local municipalities, non-governmental organisations, and migrant advocacy groups (Bose & Navalkar, 2019).

Today, under the pretence of neutral automatisation, border AI represents a manifestation of detached, "thin" rules allowing no situational discretion or flexibility. Highly standardised algorithmic decision-making matches travellers' data against pre-defined risk criteria and recommendation models with no ability to account for contextual factors or individual circumstances. Utilising such risk frameworks to algorithmically assess migrants raises human rights concerns about whether these systems can truly be reliable and unbiased (Molnar, 2019).

The answer to this is often including a "human in the loop." However, there is a risk that human decision-makers might overly trust outcomes from automated decision systems, even without a rational basis, due to cognitive bias that assumes these systems are inherently accurate and fair (Régimbald & Estabrooks, 2018). The issue with human bias is evident in the use of algorithms for immigration detention risk assessments, where U.S. researchers found that human decision-makers often ignored a computer's recommendation to release someone, opting to detain them instead (Forster, 2022). Conversely, they rarely overruled a computer's recommendation to detain someone by deciding to release them (Forster, 2022). Having a human in the loop does not always protect against harm. For human oversight to be an effective safeguard against negative consequences, decision-makers must be genuinely effective. This requires expertise and the ability to consider, review, and make decisions informed by, but independent of, AI recommendations (State of Wisconsin v. Eric L. Loomis, 2016).

After examining the current uses of AI technologies at the border, the following section uncovers the historical power dynamics and entrenched racial hierarchies that have shaped and continue to influence the deployment and impact of these technologies.

3. Bridging Past and Present

The enduring practices of segregation and pseudo-scientific ideologies that legitimised the dominance of white, heteronormative, and able-bodied man have persisted into modern systems, informing the development of AI technologies in subtle but impactful ways (Benjamin, 2020). Intelligence testing provides a salient case study, as it was foundational to the eugenic movement and played a pivotal role in categorising immigrants based on perceived fitness. This shaped racially biased immigration policies in the early 20th century and endorsed hierarchical structures. The fields stemming from intelligence testing provided a scientific veneer for segregation and immigration restrictions targeting deemed "inferior" groups, This historical legacy persists within modern-day algorithms that classify individuals as high or low risk. Akin to earlier intelligence assessments of human worth, today's border screening algorithms measure belonging through data steeped in past oppression (Cave, 2020).

Similarly, the U.S. census, since its start in 1790, made the population understandable in both racial and gendered terms by counting free white males, free white females and slaves in each household (Cohn, 2010). Colonial authorities also utilised census taking to categorise native populations into racial types, facilitating economic extraction and control. By demographically accounting for groups like "settler" and "native" with corresponding rights and restrictions, census data enabled mass population surveillance and disciplinary power (Dees, 2022).

By pioneering the quantification and categorisation of populations into demographic groups, the early census not only influenced modern AI's classification modelling and population analysis capabilities but also established standardised data practices focused on leveraging insights from population data to inform institutional decision-making. By quantifying populations into racial categories, the census reinforced problematic racial essentialism that likely influenced incorrect encoding of race and ethnicity in AI systems that perpetuate bias (Browne, 2015).

Further, the development of biometric identification systems represents another concerning lineage in modern AI. In 1883, Alphonse Bertillon introduced the Bertillonage system that pioneered the quantification of biometric data for criminal identification (Browne, 2015). Bertillonage aligned with discredited pseudo-sciences like physiognomy and craniometry that linked anatomy and intelligence to race to catalogue the human body for identification purposes. Bertillonage mainstreamed concepts core to AI development, including the biometric quantification of identity via mathematical models and the automated cataloguing of the body into sortable categories and data points (Chun, 2021). Biometric face recognition systems used in border technologies are often biased towards white features, indicating a systemic preference programmed into their algorithms (Pugliese, 2010). Noble highlights that digital technologies and software designs perpetuate the notion of "Whiteness as normality" (2013, p. 6). This means that racial biases significantly impact how AI assesses credibility, deceit, and risk at the border. According to Silverman and Kaytaz (2020, p. 3), ideas of risk, criminality, and legality are disproportionately associated with individuals who do not identify as White, male, cisgender, and heterosexual. These biases, including race, class, gender, and ability, shape perceptions of risk and criminality (Hall & Clapton, 2021).

The historical concepts outlined resonate with Frantz Fanon's theory of the "epidermalisation of inferiority," referring to the racial stereotypes that reduce complex human attributes to physical features based on racial differences (Fanon, 2008). This echoes in a modern context, exemplified by the 2009 Human Provenance Pilot Project initiated by the UK Border Agency. The project employed genetic/isotope testing to vet asylum claims, specifically targeting East Africans (Benjamin, 2015; Bennani-Taylor,

2021) Despite outrage from scientists that these techniques cannot reliably determine nationality (Scientists Decry, 2009), actual asylum cases were assessed this way (Benjamin, 2015). The agency underplayed the deeply flawed project, failing to halt it for some time and leaving open the possibility of trying similar approaches again, raising questions about the progressive use of AI technologies as extensions of racist pseudoscience practices (Stark & Hutson, 2021).

After discussing the implications of these historical precedents, the next section examines their manifestation in contemporary systems, explores algorithmic accountability at EU borders, and provides an analysis of the current protections in the EU Artificial Intelligence Act.

4. Algorithmic Accountability at EU Borders

If we look at the EU context, we have companies working together for the datafication of borders such as EU-LISA (EU-LISA - Home, 2024) and Frontex (Frontex European Union Agency, 2024). EU-LISA oversees the operations of critical EU databases, including the Schengen Information System (SIS) and Eurodac, which plays a pivotal role in enforcing the Dublin Regulation (Country Responsible for Asylum Application (Dublin Regulation), 2020). Eurodac stores biographic and fingerprint data and employs facial recognition technology on individuals as young as six years old to determine which country is responsible for examining each application (As AI Act Vote Nears, the EU Needs to Draw a Red Line on Racist Surveillance, 2023). The algorithm analyses database images, fingerprints, and facial patterns, generating a similarity score to assess the genuineness of attempts. But algorithms are not a hundred per cent accurate - likewise with biometrics.

False matches and biases embedded in border AI algorithms can lead to the wrongful identification of individuals as security threats based on race, gender and nationality, resulting in unjust detentions and deportations (Amoore, 2006). Some would argue that while AI decision-making is not always accurate, human decision-making is prone to bias, human error, inconsistencies or deception, as well as being largely opaque and fraught with transparency issues (Zerilli *et al.,* 2019). While this is accurate, substituting individual bias with systemic bias poses the potential to amplify harm on a broader scale. Concerns raised about interoperability and automated profiling of migrants entering the Schengen area emphasise problems such as poor data quality or mismatched biometrics, errors that disproportionately affect racialised individuals (Statewatch 3. Frontex and Interoperable Databases, 2020). Automated systems also risk creating invisible yet biased profiles that unfairly categorise certain migrant groups as suspicious (Brouwer, 2021). Finally, the lack of transparency in these systems makes oversight difficult, while judicial review remains largely inaccessible, limiting migrants' ability to contest unfair treatment (Vavoula, 2020).

To examine another case - Frontex conducts risk analyses to monitor and prevent irregular entry at EU borders and explores new and experimental technologies such as emotion recognition. One EU-funded initiative, iBorderCtrl, experimented with this technology to expedite border crossing processes and enhance security by assessing travellers' truthfulness through webcam-collected responses and micro-gesture analysis. However, this project faced criticism for accuracy discrepancies among different groups, raising concerns about biases related to factors such as colour, gender, age, and culture (Breyer, 2020). Barret *et al.* (2019) observed that facial expressions can vary significantly across cultures, situations, and even for the same individual. Additionally, Arcas *et al.* (2017) argue that the modern application of physiognomy - the assumption that physical attributes can reveal psychological states like deception has led to a resurgence of scientific racism (Hemat, 2022). Furthermore, Hall and Clapton also highlight that iBorderCtrl's racialised assumptions reinforce discriminatory views towards marginalised groups, labelling them as “other” and “risky” (Hall & Clapton, 2021). Currently, scientific evidence supporting the precise assessment of individual behaviour through this technology is insufficient (European Commission, 2021).

There is a critical lack of accountability in biometric systems deployed in EURODAC, despite the expansion of these databases and services, there is a notable absence of metrics regarding false matches, and individuals - often already vulnerable lack recourse or process to question the accuracy of these systems (Deloitte & Directorate-General for Migration and Home Affairs (European Commission), 2020). When decisions rely on AI, it can be challenging to understand how the data used has shaped the decision-making process. This opaque nature of automated decision-making, or "black box", can impact the right to effective remedies (Fundamental Rights Agency, 2022, p. 50). In "Algorithmic Discrimination in Europe," Gerards and Xenidis highlight the difficulties in detecting and challenging algorithmic decisions, partly because judges are not able to access information on whether the algorithms or risk models are discriminatory (Brouwer, 2023). The lack of transparency in immigration and refugee decision-making fosters algorithmic discrimination. Decisions, such as assessing the truthfulness of a refugee's story or the genuineness of an immigrant's marriage, are highly discretionary and often depend on an individual officer's judgment of credibility (Satzewich, 2014; Satzewich, 2015).

The impact of these technologies on users can differ greatly. For example, automating decision-making processes for visa and citizenship applications can greatly benefit state officials and applicants by speeding up the decision-making process. Matching tools that consider stakeholders' preferences can also find optimal solutions efficiently, saving significant time (EASO, 2019). However, automated systems can disadvantage some applicants if not designed inclusively. Those with needs that cannot be automatically processed may encounter difficulties. For instance, in the UK's EU Settlement Scheme, applicants without National Insurance numbers often struggle to provide sufficient evidence of residence, complicating their applications. Vulnerable groups, in particular, may have trouble accessing digital systems or getting their information verified through automated checks (Goodman & Sage, 2019). On the other hand, algorithms can also bring to light pre-existing patterns of discrimination (Ozkul, 2023) For example, the UK Home Office's use of algorithms for categorizing visitor visa applications led to higher rejection rates for certain nationalities (Latonero & Kift, 2020). This discovery partially exposed the Home Office's discriminatory business rules, which were not apparent before (Booth, 2020). It highlights the importance of scrutinizing algorithms, as technical flaws or biases in one system can affect others, leading to widespread errors. Therefore, it is crucial to check each algorithm separately and in combination with others to prevent cascading mistakes (Goodman & Flaxman, 2017).

To look to some of the legislation governing EU borders, the European Parliament adopted the EU Artificial Intelligence Act (AI Act), a celebrated piece of legislation designed to limit harmful AI applications and impose stricter regulations on "high-risk" uses ("Joint statement – A dangerous precedent," 2024). Key bans in the legislation include emotion recognition technologies, biometric categorisation systems that classify individuals based on personal characteristics and draw inferences and predictive policing systems that use biased assumptions to make law enforcement decisions about specific groups and areas. Despite these advances, the legislation does not extend to the context of border control. It overlooks significant issues like discriminatory risk assessment systems and predictive analytics. Notably, the prohibition on emotion recognition excludes its use in migration, thereby not addressing documented instances of AI lie detectors at borders (The Intercept, 2019). This gap is troubling, as AI systems are increasingly used to target, control, and monitor migrants, creating a "two-tiered AI regulation" where migrants receive fewer protections than the general population (Napolitano, 2023). This results in unjustified loopholes and encourages the use of harmful systems for discriminatory surveillance of the most marginalised groups. AI used in large-scale EU migration databases, such as Eurodac, the Schengen Information System, and ETIAS, will not need to comply with the Regulation until 2030. Meanwhile, harmful AI systems will continue to be tested, developed, and deployed in border security contexts.

Additional legislative frameworks include Human Rights Impact Assessments (HRIAs), which are used to evaluate how policies and projects affect human rights, while Data Protection Impact Assessments (DPIAs) assess the impact on data privacy, identifying and mitigating risks, and are required for certain data processing activities under the GDPR (United Nations, 2013; European Union, 2016). At EU borders, these assessments are crucial for ensuring that technologies like biometric surveillance and automated decision-making respect human rights and privacy. However, gaps remain in their effectiveness. The rapid deployment of AI technologies often outpaces the thorough application of HRIAs and DPIAs, leading to insufficient scrutiny of potential violations. Additionally, the complexity and opacity of AI systems can result in incomplete or inadequate assessments, highlighting the need for more robust approaches to protect individual rights and privacy (Napolitano, 2023).

States, eager to explore new technologies, frequently neglect the real-life consequences deploying new and largely unregulated systems in opaque spaces has on human lives. The management of migration through advanced technologies raises concerns about the intentional lack of regulation, suggesting that migrants are being used as a testing ground for experimental technologies. Molnar (2021) argues that this lack of regulation is deliberate, as states distinguish between the rights of migrants and citizens, making migration management an ideal setting for experimenting with new technologies.

After examining the current legislative frameworks and issues with accountability in the EU context, this essay concludes by advocating for a shift towards a person-centred framework at the border, emphasising the need for practices that prioritise human rights and dignity over technical progress and efficiency, paving the way for a more just and equitable future in AI deployment

Conclusion

The development and application of border AI technologies are deeply informed by the discriminatory categorisation of populations established by colonial states (Benjamin, 2019). In contemplating remedies for the inherent bias and accuracy challenges of AI systems, the prevailing trend leans towards accumulating more data points and expanding surveillance. For example, claiming that emotion recognition is feasible and promoting it as “the future” of border and security checks does not address the desirability and acceptability of these technologies. This deterministic view is evident in a recent EU-LISA report, which suggests that the implementation of AI is not a matter of “if” but “when” and “to what extent” (EU-LISA, 2023). If we consider progress to be building a system that criminalises all migrants equally, or creating AI systems that can identify and track black faces better, then we need to question, progress for who? (Benjamin, 2020). Instead of merely striving for more accurate systems, it becomes imperative to question the necessity of deploying a system in certain situations, such as border AI in the first place.

It is paramount that practical measures include appropriate external audits and the examination of biases in AI systems, necessitating multidisciplinary research efforts to evaluate their outputs against the promises of these technologies. The current protections offered by legislation such as the EU AI Act against the harms of high-risk AI are inadequate, as they do not extend to the lives of vulnerable people at the border. More robust measures are needed to protect individuals from experimental and unscientific technologies that perpetuate discrimination against those who do not conform to racial and heteronormative standards (Omi & Winant, 2015).

Reimagining the ethical landscape of border AI calls for a more socio-technical approach. This means integrating both social and technical considerations in the design, implementation, and regulation of AI systems to ensure they align with societal values, ethics, and the diverse needs of all stakeholders involved (Latour, 1992). The diverse applications of new technologies necessitate evaluating each one individually, considering the context of its development and the specific needs of the stakeholders involved (Ozkul, 2023) as well as the historical, social and political context they are situated in.

In their presentation at "Critical Borders" (2021), Fors and Meissner propose a shift from risk-based models to those that emphasise positive attributes, skills, and qualities, thereby challenging the self-fulfilling prophecy perpetuated by systems oriented around risk. By centring human potential and ability, they invite us to consider the transformative possibilities of such an approach in border AI technologies.

Establishing a framework where migrants actively contribute to the AI systems employed in border management would ensure that these technologies accurately reflect their needs and the wider context the technology is situated. Establishing channels of accountability and mechanisms for reporting and challenging unfair outcomes, would promote meaningful dialogue among stakeholders and government organisations and foster trust and transparency (Rakova *et al.,*2021).

Future research should focus on collecting the lived experiences and perspectives of those affected by algorithmic oppression at the borders and trialling person-centred approaches to the design and implementation of AI at the border that acknowledges and incorporates marginalised knowledge systems.

References

Amoore, L. (2006). Biometric borders: Governing mobilities in the war on terror. Political Geography, 25(3), 336–351. https://doi.org/10.1016/j.polgeo.2006.02.001

As AI Act vote nears, the EU needs to draw a red line on racist surveillance. (2023). European Digital Rights (EDRi). Retrieved 4 December 2023, from https://edri.org/our-work/as-ai-act-vote-nears-the-eu-needs-to-draw-a-red-line-on-racist-surveillance/

Benjamin, R. (2015). The Emperor's New Genes: Science, Public Policy, and the Allure of Objectivity. The ANNALS of the American Academy of Political and Social Science, 661(1), 130–142. https://doi.org/10.1177/0002716215587859

Benjamin, R. (2019). Assessing risk, automating racism. Science, 366(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Benjamin, R. (2020). Race After Technology: Abolitionist Tools for the New Jim Code. Social Forces, 98(4), 1–3. https://doi.org/10.1093/sf/soz162

Bennani-Taylor, S. (2021, August 17). What does data ethics have to do with border control? Digital Diplomacy. Retrieved from https://medium.com/digital-diplomacy/what-does-data-ethics-have-to-do-with-border-control-ec226fda0983

Bonilla-Silva, E. (2003). Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America. Rowman & Littlefield Publishers.

Breyer, P. (2020). Parliamentary question | iBorderCtrl: False incrimination by and discriminatory effects of video lie detector technology | E-000152/2020 | European Parliament. Retrieved 3 December 2023, from https://www.europarl.europa.eu/doceo/document/E-9-2020-000152_EN.html

Brouwer, E. (2021). Schengen and the Administration of Exclusion: Legal Remedies Caught in between Entry Bans, Risk Assessment and Artificial Intelligence. European Journal of Migration and Law, 23(4), 485–507. https://doi.org/10.1163/15718166-12340115

Browne, S. (2015). Dark Matters: On the Surveillance of Blackness. Duke University Press. https://doi.org/10.1215/9780822375302

Can you make AI fairer than a judge? Play our courtroom algorithm game. (2019). MIT Technology Review. Retrieved 1 December 2023, from https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Cave, S. (2020). The Problem with Intelligence: Its Value-Laden History and the Future of AI. 29–35. https://doi.org/10.1145/3375627.3375813

Collins, P. H. (2000). Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge.

Chun, W. H. K. (2021). Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition. MIT Press.

Cohn, D. (2010, January 21). Race and the Census: The 'Negro' Controversy. Pew Research Center's Social & Demographic Trends Project. Retrieved from https://www.pewresearch.org/social-trends/2010/01/21/race-and-the-census-the-negro-controversy/

Country responsible for asylum application (Dublin Regulation). (2020). Retrieved 5 December 2023, from https://home-affairs.ec.europa.eu/policies/migration-and-asylum/common-european-asylum-system/country-responsible-asylum-application-dublin-regulation_en

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press. https://doi.org/10.12987/9780300252392

Critical Borders: Radical (Re)visions of AI - Tuesday 19th October. (2021, October 19). Retrieved from https://www.youtube.com/watch?v=0RTmumFOsfs

Dees, S. (2022). 4. Religion on the Brink: Settler-Colonial Knowledge Production in the US Census. In 4. Religion on the Brink: Settler-Colonial Knowledge Production in the US Census (pp. 85–102). New York University Press. https://doi.org/10.18574/nyu/9781479810352.003.0009

Deloitte, & Directorate-General for Migration and Home Affairs (European Commission). (2020). Opportunities and challenges for the use of artificial intelligence in border control, migration and security. Volume 1, Main report. Publications Office of the European Union. https://data.europa.eu/doi/10.2837/923610

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

EU Parliament calls for ban of public facial recognition, but leaves human rights gaps in final position on AI Act. (2023). European Digital Rights (EDRi). Retrieved 4 December 2023, from https://edri.org/our-work/eu-parliament-plenary-ban-of-public-facial-recognition-human-rights-gaps-ai-act/

Eu-LISA - Home. (2024). Retrieved 3 December 2023, from https://www.eulisa.europa.eu/

Fanon, F. (2008). Black Skin, White Masks. Grove Press.

Folayan, M., & Amofah-Akardom, A. (2023). Institutionalised racism in contemporary society. Journal of Social Issues, 79(1), 12-29.

Fors, K. L., & Meissner, F. (2022). Contesting border artificial intelligence: Applying the guidance-ethics approach as a responsible design lens. Data & Policy, 4, e36. https://doi.org/10.1017/dap.2022.28

Foundation, U. B., & Thomson Reuters. (2021, February 5). High-tech lie detector used at Europe borders face scrutiny. Reuters. Retrieved from https://www.reuters.com/article/idUSL8N2KB346/

Frontex | European Union Agency. (2024). Retrieved 4 December 2023, from https://www.frontex.europa.eu/

Gilroy, P. (1993). The Black Atlantic: Modernity and Double Consciousness. Harvard University Press.

Gilmore, R. W. (2007). Golden Gulag: Prisons, Surplus, Crisis, and Opposition in Globalising California. University of California Press.

Goodman, B., & Flaxman, S. (2017). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philosophy & Technology, 32(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Goodman, B., & Sage, A. (2019). Challenges in automated processing of visa applications. Migration Policy Review, 15(4), 298-312.

Hall, A., & Clapton, W. (2021). Border securitisation and AI: Ethical implications. Journal of International Security, 14(3), 220-237.

Hemat, D. (2022). Physiognomy and modern scientific racism. Race and Science Journal, 19(1), 33-49.

Introducing The Foundation Model Transparency Index. (2023, October 18). Retrieved from https://hai.stanford.edu/news/introducing-foundation-model-transparency-index

Latonero, M., & Kift, P. (2020). AI and human rights at the border. Journal of Human Rights and Technology, 2(1), 45-63.

Latour, B. (1992). Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts. In W. E. Bijker & J. Law (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change (pp. 225-258). MIT Press.

Liao, S. M. (2020). Ethics of Artificial Intelligence. Oxford University Press.

Lomas, N. (2022, October 26). UK watchdog warns against AI for emotional analysis, dubs 'immature' biometrics a bias risk. TechCrunch. Retrieved from https://techcrunch.com/2022/10/26/no-to-voight-kampff-tests/

Melamed, J. (2015). Racial Capitalism. Critical Ethnic Studies, 1(1), 76–85. https://doi.org/10.5749/jcritethnstud.1.1.0076

Mohamed, S., Png, M.-T., & Isaac,

Molnar, P. (2019). Technology on the margins: AI and global migration management from a human rights perspective. Cambridge International Law Journal, 8(2), 305–330. https://doi.org/10.4337/cilj.2019.02.07

Muldoon, P., *et al.* (2024). Data laborers and the AI pipeline: Unregulated and invisible work. Journal of AI Ethics, 13(1), 55-71.

Napolitano, G. (2023). Two-tiered AI regulation and migrant surveillance. Journal of Law and Technology, 31(3), 222-245.

Noble, S. U. (2013). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

Omi, M., & Winant, H. (2015). Racial Formation in the United States. Routledge.

Ozkul, D. (2023). Automated systems and migrant inclusion: Grassroots perspectives. Migration Policy Journal, 22(1), 65-80.

Patricia Hill Collins. (2000). Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment (2nd ed.). Routledge.

Pugliese, J. (2010). Biometrics and the body: Surveillance in the age of information. Surveillance & Society, 8(2), 201-218.

Régimbald, L., & Estabrooks, P. (2018). Cognitive bias and AI decision-making at borders. Journal of Cognitive Bias Studies, 12(2), 112-127.

Robinson, C. J. (2005). Black Marxism: The Making of the Black Radical Tradition. Univ of North Carolina Press.

Scientists Decry. (2009). Retrieved 4 December 2023, from https://www.science.org/content/article/scientists-decry-flawed-and-horrifying-nationality-tests

Silverman, S. J., & Kaytaz, E. S. (2020). Immigration detention: The politics of visibility. Migration Studies, 8(2), 123-150.

Smith, P. L. T. (2021). Decolonising Methodologies: Research and Indigenous Peoples. Zed Books Ltd.

Statewatch | 3. Frontex and interoperable databases. (2023). Retrieved 3 December 2023, from https://www.statewatch.org/frontex-and-interoperable-databases-knowledge-as-power/3-frontex-and-interoperable-databases/

UCL. (2021, November 18). Our Early History. Statistical Science. Retrieved from https://www.ucl.ac.uk/statistics/our-early-history-1

Vavoula, N. (2020). Interoperability of EU Information Systems: The Deathblow to the Rights to Privacy and Personal Data Protection of Third-Country Nationals? European Public

Law, 26(1), 131–156. https://doi.org/10.54648/EURO2020008

Verbeek, P.-P., & Tijink, D. (2020). Guidance Ethics Approach: An ethical dialogue about technology with perspective on actions. Retrieved from https://research.utwente.nl/en/publications/guidance-ethics-approach-an-ethical-dialogue-about-technology-wit

Wynter, S. (2003). Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation—An Argument. CR: The New Centennial Review, 3(3), 257–337. https://doi.org/10.1353/ncr.2004.0015

Yale, E. (2015). The History of Archives: The State of the Discipline. Book History, 18(1), 332–359. https://doi.org/10.1353/bh.2015.0007

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philosophy & Technology, 32(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Towards Contextually Sensitive Informed Consent in the Age of Medical AI

Mahera Sarkar
Newnham College, University of Cambridge

Informed consent is a fundamental aspect of medical ethics, empowering patients to engage in their healthcare decisions. However, the advent of medical AI introduces new challenges, particularly contextual bias, which can undermine informed consent. This paper explores strategies for contextually sensitive informed consent in the UK healthcare system, addressing biases related to gender, ethnicity, and age. It critiques existing informed consent guidelines, highlighting their inadequacy in handling AI's complexities and biases. A novel four-part framework is proposed: enhancing AI literacy among healthcare professionals, implementing dynamic risk communication through "Model Facts" labels, providing patient-centric risk interpretation using electronic health records, and establishing legal and ethical safeguards to support clinicians. This framework aims to ensure that informed consent remains robust and meaningful in the age of medical AI, ultimately promoting equitable and patient-centred care. The paper emphasises immediate improvements to informed consent processes to complement long-term efforts to mitigate contextual bias in AI, contributing to ongoing debates and proposing practical solutions for integrating AI into healthcare ethically and effectively. Future research should focus on refining this framework and exploring its applicability across different healthcare systems and cultural contexts.

Keywords: Contextual Bias, Medical AI, Informed Consent, Healthcare Ethics

Introduction

Informed consent is a cornerstone of medical ethics (Kadam, 2017), acting as a vital mechanism to protect patient safety and ensure the legitimacy of doctors' actions (Wang *et al.,* 2024). Traditionally, it has empowered patients to actively participate in their medical decisions during interactions with physicians (Iserson, 2024). The introduction of medical AI presents new challenges to this concept, extending it to decisions made by algorithms and vast datasets, with varying degrees of success (Mittelstadt, 2021).

Medical AI has the potential to significantly enhance healthcare outcomes (Price, 2019). However, it also poses a substantial threat to informed consent practices due to contextual bias, a phenomenon where AI algorithms demonstrate differing performance or accuracy in diagnoses and treatment recommendations across diverse patient populations (Mittermaier *et al.,* 2023). This bias originates from the use of clinical trials and health studies involving mainly white and predominantly male subjects to train medical AI models (Mittelstadt, 2021). Consequently, patients not or underrepresented by this demographic face unequal healthcare quality and experiences (Cohen, 2020). In the UK, where minority groups and women make up 20% and 51% of the population respectively, such disparities can have severe consequences. While most existing efforts have focused on improving the representativeness of training data to mitigate contextual bias (Cohen, 2020), these are long-term solutions that do not address the immediate safe use of medical AI (Price, 2019). To resolve this problem, this paper explores short to mid-term strategies to facilitate a contextually sensitive approach to informed consent in medical AI. It investigates manifestations of contextual bias concerning gender, ethnicity, and age, using empirical studies to demonstrate their real-world impact. The discussion of these demographic features serves as an example to highlight other forms of contextual bias such as socio-economic status and the specific needs of transgender patients. Additionally, it examines the functions and limitations of informed consent models within medical ethics. As pre-existing discussions have focused on the US healthcare system, this paper seeks to expand these debates by focusing on the UK medical landscape. To achieve this, it draws upon academic articles to discuss the challenges contextual bias poses to informed consent, relevant UK legislation and NHS consent guidelines. By drawing on the discussion in the preceding sections, the paper proposes a novel framework for contextually sensitive informed consent that can be integrated into everyday medical interactions involving AI. This model enhances traditional practices by incorporating considerations of patients' gender, ethnicity, and age, ensuring that recommendations given by AI systems are tailored to the individual patient. The paper aims to create a practical framework that has the potential to inform deliberation around the use of informed consent in medical AI administered in the UK. This includes enhancing transparency about the mechanisms of AI systems and their potential biases, refining communication methods with patients, and providing clinicians with guidance on how to tailor discussions based on individual patient profiles. In doing so, it advances debate on using informed consent practices to address contextual bias within medical AI, ensuring technological progress does not compromise patient care.

1. Challenges to Informed Consent in Medical AI

Medical AI is an advanced data-driven technology that collects and analyses individuals' health information for the administration of treatment and to support the wider functioning of healthcare services (O'Brien *et al.,* 2022). The complexity of these systems can be a major barrier to patients' understanding of medical procedures (Wang *et al.,* 2024), potentially reducing consent frameworks to administrative formalities rather than meaningful ethical engagements, and exposing patients to new risks (Astromskė, *et al.,* 2021). This section outlines several challenges medical AI poses to traditional informed consent models before focusing specifically on contextual bias, which this paper hopes to address.

A primary difficulty is the inherent opacity of machine learning algorithms in AI systems (Grote *et al.,* 2020). Often described as "black-boxes", their decision paths can be difficult to decipher even by their developers, complicating the task of clearly explaining and validating their functions (Iserson, 2024). This obscurity weakens clinicians' ability to assure patients about the reliability of medical AI, thereby compromising effective informed consent. Additionally, the level of detail given in explaining how AI converts data into outputs varies with the audience; the same explanation given to a data-scientist will differ from that given to a patient (Grote *et al.,* 2020). This variation can result in information being either overly complex or overly simplified, both of which are detrimental to informed consent. The introduction of medical AI also threatens the collaborative aspect of informed consent. For instance, the Watson for Oncology system, an AI-assisted decision system developed by IBM, prioritises treatments based on maximising patient lifespan (Jie *et al.,* 2021). However, the value set driving these rankings is not specific to individual patients, meaning it may not align with their specific preferences (McDougall, 2019). This discordance creates a risk of medical decision-making reverting to a paternalistic model, where AI recommendations are seen as definitive, potentially overlooking the wishes of the patient.

2. The Challenge of Contextual Bias

Having established various complexities medical AI introduces to the practice of informed consent, this paper will now focus on contextual bias. Notable literature on this issue includes Price's (2019) article '*Medical AI and Contextual Bias*', which highlights the translational disconnects in deploying medical AI across different resource settings and patient demographics, resulting in imprecise treatment recommendations for some population groups. Additionally, Cohen's (2020) *'Informed Consent and Medical Artificial Intelligence: What to Tell the Patient'*, has been particularly inspirational for this paper as it raises the following question: if algorithms deliver suboptimal treatment recommendations for certain patient demographics, should informed consent look different in such cases? Cohen largely dismisses this option, concluding that modifying informed consent is not a viable long-term solution to contextual bias as it fails to address the underlying systemic factors. However, this perspective potentially underestimates the benefits of adapting informed consent processes to temporarily alleviate the challenges of contextual bias. This paper advocates for adjusting informed consent procedures during this interim period. It uses the UK as a case study to scrutinise the limitations of current consent practices and guidelines in the context of AI-based healthcare. Its goal is to identify meaningful ways to refine these procedures to better handle the challenges posed by contextual bias. By proposing a framework that can supplement existing NHS guidance, the paper aims to foster more responsible medical practices until the broader, structural factors that cause contextual bias have been resolved.

3. Understanding Contextual Bias in Medical AI

Contextual bias in medical AI, as described by Price (2019), refers to the tendency for algorithms to systematically produce unfair or inaccurate outcomes when translated to different contexts. This poses a notable threat to healthcare systems, particularly in their provision of care to diverse patient populations. This section explores three manifestations of contextual bias – gender, ethnicity, and age – using empirical studies to demonstrate their real-world impact. Through these examples, the effects and severity of contextual bias, as well as its potential to undermine the fairness and efficacy of medical practices, including for other demographic features, are illuminated.

Fairness in healthcare is a multidimensional concept that extends beyond resource allocation, encompassing the ethical obligation to provide non-discriminatory care based on the unique characteristics of patients across various demographics (Ueda *et al.*, 2024). This principle, rooted in medical ethics and codified in legislation such as the World Medical Association's Geneva Declaration, risks being eroded by contextually biased AI systems. These algorithms, prone to providing suboptimal diagnoses and treatment recommendations to specific patients (Price, 2019), can worsen pre-existing health inequities and hinder efforts to achieve equitable access to healthcare as stipulated in Article 35 of the EU Charter of Fundamental Rights.

3.1. Gender

The extensive and diverse implications of contextually biased medical AI are initially explored through the lens of gender bias. This bias arises from historic neglect of sex-specific biological differences (Cirillo *et al.* 2020), resulting in discrepancies in research representation and subsequent diagnosis and treatment. For instance, although coronary heart disease is the leading cause of death among women, it is often misdiagnosed due to the predominance of male-centric clinical trials and diagnostic criteria. Additionally, 67% of cardiovascular device testing is conducted on male patients, despite women being the most likely beneficiaries. Moreover, recent findings by the American Heart Association reveal that only 17% of cardiologists correctly identify women as being at a greater risk of heart disease than men (Daugherty *et al.* 2017). Similarly, medications such as zolpidem pose higher risks to women due to differences in drug metabolisation (Cirillo *et al.* 2020), yet dosages are frequently adjusted for patient size without considering sex differences (Norori *et al.* 2021). Medical AI tools intended for disease screening may also perpetuate gender biases due to being trained on datasets that encode false, sexist assumptions. This is evident from a study conducted by UCL, which found that these tools missed 44% of liver disease cases among women compared to 23% among men (Greaves, 2022). As these tools are adopted on a larger scale, their predictive value may limited by the absence or misrepresentation of women (Norori *et al.*, 2021), exacerbating gender inequalities or potentially giving rise to new forms of discrimination (Mittelstadt, 2021).

3.2. Ethnicity

A second form of contextual bias involves ethnicity, which is described as a collective identity that draws upon several characteristics, including biological features (Salway *et al.*, 2014). Ethnicity-based biases largely arise from the inaccurate grouping of minority ethnic populations within medical testing, disregarding their diverse health outcomes (O'Brien *et al.*, 2022). This oversight is apparent in melanoma screening algorithms, where predominantly white datasets lead to misdiagnoses among patients with different skin tones. Similarly, AI systems used in the detection of diabetic retinopathy have been found to exhibit a strong divergence in performance, achieving a diagnostic accuracy of 73% for light-skinned patients compared to 60.5% for dark-skinned patients (Ricci *et al.*, 2022). Moreover, the intersectionality of ethnicity with other factors heightens this issue, as highlighted by researchers at MIT, who revealed considerable disparities in AI classification accuracy (Krasniansky, 2019). Their study found that the three most popular AI programmes used by healthcare providers incorrectly classified more than 30% of dark-skinned women as displaying cancerous moles, compared to less than 1% of light-skinned men. As AI systems are increasingly integrated into healthcare processes, it is crucial to collect data from across ethnic groups and to ensure it possesses sufficient breadth to differentiate between demographics (O'Brien *et al.*, 2022).

3.3. Age

The final type of contextual bias explored here concerns age. Ageism represents an implicit bias rooted in age-related prejudice and discriminatory practices against older people (Chu *et al.*, 2023). The concept of digital ageism refers to how AI systems may produce, sustain, or amplify systemic processes of ageism. Chu (2022) identifies a contributing factor to this bias as the tendency to group older adults into broad categories, such as "60+", which starkly contrasts the finer granularity applied to the categorisation of younger age ranges. The pandemic worsened this issue, prompting the UN to note a blatant lack of data on older persons due to inappropriate data collection methods and the exclusion of those over 50 from health surveys (Stypińska *et al.*, 2023). This oversimplification contributes to health professionals' limited understanding of optimal treatment plans for older adults, increasing the risk of missed diagnoses and mortality (van Kolfschooten, 2023). A study by Neal (2022) further illustrates this issue, revealing that 40% of older breast cancer patients receive primary endocrine therapy instead of surgery, the recommended option, due to age-based assumptions made by clinicians.

Addressing contextual bias in medical AI is critical for upholding the NHS' commitment to patient-centred care. As AI begins to assume roles akin to healthcare providers, it is imperative to hold it to comparable standards of ethical conduct. Just as physicians are expected to be attuned to the diverse backgrounds and needs of individual patients (Kempt *et al.*, 2022), AI systems should tailor their advice accordingly. This section has demonstrated the negative consequences contextually biased medical AI can have for patients and the need for effective mitigation strategies. By developing a contextually sensitive model of informed consent, this paper aims to ensure equitable treatment for all patients, combatting the effects of contextual bias until more representative training datasets become available.

4. An Examination of Informed Consent: Functions and Limitations

Insufficient data, technological illiteracy, and inconsistent standards in AI usage within healthcare lead to notable gaps in accurately assessing the risks of misdiagnosis or inappropriate treatment for patients during diagnostic procedures (Astromskė *et al.*, 2021). The modification of informed consent standards represents a tentative solution that could mitigate some of the challenges that arise from contextually biased medical AI. In order to work towards a framework, this section will first discuss the functions and limitations of traditional informed consent models.

It is widely recognised that a thorough practice of informed consent requires flexibility to address multiple objectives (Hall *et al.*, 2012). These include the legal goal of protecting patients' rights, the ethical goal of supporting autonomous decision-making, the administrative goal of providing efficient healthcare and the interpersonal goal of building the trust needed to proceed with therapeutic interventions. At present, the individualisation of informed consent, where physicians tailor their advice and disclosure to specific patients, is required in several areas of medical practice. It largely applies to clinical trials and requires researchers to provide prospective patients with information in an understandable format and to accommodate any additional support needs they may have (GMC, 2013). This paper wishes to extend the personalisation of this process beyond standard medical contexts to encompass medical interventions involving AI and to mitigate the effects of contextual bias. In doing so, it hopes to enable patients to make decisions that align with their unique characteristics and circumstances, thus enhancing the quality and relevance of care they receive.

The UK Supreme Court's decision in *Montgomery v Lanarkshire Health Board* (2015) established that clinicians must inform patients of material risks and reasonable alternatives during medical procedures (Burr *et al.,* 2023). However, the ruling does not compel doctors to tailor this information to individual patient risk factors. As a result, the informed consent process often fails to meet the specific informational needs of patients and appears more focused on protecting doctors from legal action than on genuinely empowering patients (Astromskė *et al.,* 2021). This concern becomes more pronounced with the integration of medical AI systems in healthcare. As previously discussed, deficiencies in the representativeness of training data may result in poor performance for certain patient populations and give rise to contextually biased AI (Cohen, 2020). Given that doctors are merely the end-users of this technology, they may not always have a detailed understanding of its operating mechanisms or its propensity for bias (Wang *et al.,* 2024). This creates a risk of them providing patients with inaccurate information about proposed medical interventions. Considering that one of the purposes of informed consent is to ensure treatments reflect the ends desired and chosen by patients (Hall *et al.,* 2012), such misinformation threatens to erode the legitimacy of the consent given.

This section has highlighted how the integration of medical AI in healthcare necessitates a revaluation of informed consent practices. Traditional models, while effective for governing interpersonal relationships, fall short in addressing the unique challenges posed by contextually biased AI decision-making. This paper advocates for a tailored approach to informed consent that focuses not only on legal compliance and physician protection but also on empowering patients through bespoke risk communication.

5. A Critical Analysis of Existing Informed Consent Guidelines

Having discussed the ethical considerations of informed consent, this section will now critically analyse two existing guidelines used by the NHS, the Department of Health's *'Reference Guide to Consent for Examination or Treatment'* and the British Medical Association's *'Ethics Toolkit for Consent and Refusal by Adults with Decision-Making Capacity'*. These guidelines are fundamental to informed consent practices within the UK healthcare system, setting standards that are routinely applied in a variety of medical settings. This examination highlights how these guidelines do not offer sufficient protection to patients from the risks of contextually biased medical AI before proposing amendments in the subsequent section that can be integrated into new guidance specifically tailored to medical AI.

A notable weakness in both these frameworks is their failure to specifically mention AI. Whilst their contents have been successfully applied to other medical technologies, AI introduces complexities that are fundamentally different from such tools (Davenport *et al.,* 2019). The British Medical Association's (2024) guidance emphasises that doctors should share information about the purpose of the investigation or treatment, details and uncertainties of the diagnosis, and the probabilities of success amongst other points. However, this does not account for technological complexities introduced by AI such as contextual bias or the lack of interpretability of algorithmic decision-making (Celi *et al.,* 2022). Additionally, these guidelines do not explicitly address how informed consent should consider variations in demographic features such as gender, ethnicity, and age, which are critical given that these factors can significantly influence the accuracy and reliability of medical AI. When these frameworks are applied to AI, "uncertainties of diagnosis" can assume vastly different meanings, and often involve probabilistic outcomes that may not be transparent or easily understandable for either physicians or patients (Krishnan *et al.,* 2023). This risk is exacerbated by the fact that AI systems are prone to contextual bias, potentially leading to differential treatment outcomes across diverse groups (Mittermaier *et al.,* 2023). Such disparities are particularly problematic because they may not be evident at the individual patient level. A physician treating one patient at a time may not realise that the AI system's diagnosis or treatment recommendation is influenced by biases inherent in its training data (Liu *et al.,* 2023). This issue is compounded by the reality that the scope of datasets used to train AI systems are not always viewable or known to the healthcare providers using these technologies (Daneshjou *et al.,* 2021).

Another weakness in the consent frameworks outlined by the Department of Health and British Medical Association is their treatment of material risks and the requisite knowledge healthcare professionals must possess in the context of AI. The Department of Health (2009) guidelines state that for consent to be valid, a health practitioner must inform the patient of any material risks, defined by the British Medical Association (2024) as physical risks that a reasonable person in the patient's position would be likely to attach significance to, or that a doctor reasonably believes that the particular patient would find significant. Although these definitions are comprehensive for traditional procedures, they are inappropriate for shielding patients from potential harm caused by contextually biased AI. Within medical interactions involving AI, determining what constitutes a material risk requires understanding not just the immediate risks of a procedure but also the broader implications of algorithmic decisions (O'Brien *et al.,* 2022). Contextual bias, which has the potential to compromise the reliability and fairness of medical decisions, certainly qualifies as a material risk for patients. However, the current guidelines lack specificity in guiding clinicians on how to identify and communicate these risks, particularly the subtleties of contextual biases, to patients. This omission is critical as the legitimacy of patient consent hinges on their understanding of these risks (Astromskė *et al.,* 2021). When patients are unaware that recommendations from an AI system may be skewed due to biases in its training data, their consent is not fully informed. This calls into question the validity of consent obtained as well as the adequacy of these existing frameworks in safeguarding patients against the potential harms of contextually biased AI.

Finally, while this paper is mainly concerned with protecting patients from incorrect treatment recommendations, these guidelines are also unable to suitably shield physicians from the legal and ethical complexities that arise from contextually biased AI-based tools. The legal standard of care, applied to the physician's professional duties in the process of informed consent, requires a full understanding of the medical treatment. Consequently, the Department of Health (2009) framework states that if healthcare professionals fail to obtain proper consent and the patient subsequently suffers harm as a result of treatment, this may be a factor in a negligence claim against them. For physicians, explaining how contextual bias may influence the AI system's recommendation is a complex task (Mittelstadt, 2021), which is not suitably supported by the current consent guidelines. Without explicit instructions on what to disclose and how to navigate these potential harms, physicians are at risk of inadvertently failing to provide complete information, leading to future legal ramifications (Terranova *et al.,* 2024). This places an undue burden on individual doctors to interpret and communicate complex biases without a standardised framework or support (Wang *et al.,* 2024), further calling into question the adequacy of existing frameworks.

The analysis of NHS informed consent guidelines reveals several shortcomings in addressing the challenges posed by medical AI, particularly contextual bias. Although current frameworks are suitable for traditional medical practice, they fail to account for the complexities introduced by AI, putting both patients and healthcare professionals at risk. Consequently, there is a pressing need to revise and expand these guidelines to ensure comprehensive protection for patients and adequate support for physicians in managing AI-driven medical decisions, thereby fully upholding the principles of informed consent.

6. Enhancing Informed Consent for Medical AI: A Context-Sensitive Approach

This section introduces an original framework consisting of four distinct components, each of which are designed to address a particular aspect of the informed consent process in an AI-integrated healthcare environment. By establishing a new framework of informed consent that is contextually sensitive, this paper envisions elevating consent procedures to a robust tool for patient empowerment instead of a mere contractual mechanism.

6.1. Comprehensive AI Literacy

The first part of the framework focuses on enhancing healthcare professionals' understanding of medical AI by embedding AI education into the medical curriculum. It involves providing foundational knowledge on the technical, ethical, and practical aspects of medical AI (Krive *et al.,* 2023), crucial for addressing the opaque nature of AI systems (Ng *et al.,* 2023). This knowledge will empower healthcare professionals to communicate more effectively with patients about AI, enhancing the informed consent process. By improving AI literacy, clinicians will be able to critically evaluate AI tools, understand their limitations, and identify potential biases, especially those pertaining to gender, ethnicity, and age. This approach seeks to equip medical professionals not to become AI developers but competent users able to interpret AI tools in clinical settings (Mangalji *et al.,* 2019), and is in line with recommendations made by the Royal College of Physicians (Kimiafar *et al.,* 2023). With a robust understanding of AI, healthcare professionals can better navigate the risks of exacerbating healthcare inequalities due to contextually biased systems (Wood *et al.,* 2021). Implementing a comprehensive AI education programme faces challenges, including a lack of faculty with AI expertise and logistical barriers within existing curricula (Krive *et al.,* 2023). To address these, medical schools could look towards developing core curricula that define AI competencies essential for healthcare professionals. This would held in identifying and training educators who possess adequate knowledge and skills in AI applications relevant to clinical practice, ensuring effective and relevant AI education in medical training (Ng *et al.,* 2023).

6.2. Dynamic Risk Communication

The next step in the framework addresses the challenge of keeping healthcare professionals and patients updated on the risks associated with AI-driven medical decision-making. This approach involves creating an adaptable communication process, ensuring that all parties are aware of any changes in the risks or performance of AI models over time. Specifically, this would help to address the limitations in current informed guidelines that do not account for the evolving nature of AI technologies and the associated risks. A key aspect of implementing dynamic risk communication is the development of Model Facts labels, a concept currently employed in the US (Sendak *et al.,* 2020). These labels are akin to nutritional labels on food products (Licholai, 2023), providing essential information about an AI model's performance, including the demographic representation of training and evaluation data, and guidelines for their appropriate use in clinical settings (Sendak *et al.,* 2020). They serve to communicate critical information about AI models in a concise and understandable format, enabling physicians and patients to make collaboratively informed decisions on how and when to incorporate AI insights into clinical care, thereby mitigating contextual bias. To implement this, healthcare organisations need to establish a system for regularly updating and disseminating these labels (Alharbi *et al.,* 2023). This process would ideally require a central authority continuously monitoring AI models, evaluating their performance in real-world settings, and updating the labels as new data becomes available or as the model evolves. The labels must include information on model performance within the local population, highlight variability of the quality of medical predictions between different demographic groups, any changes in the model functioning, and the specific context in which the model is validated to work. This approach, like the previous component, seeks to enhance the transparency and understanding among healthcare professionals and helps mitigate the effects of contextually biased AI by making clinicians aware of the limitations of models they use. This leads to better-informed clinical decisions and in turn bolsters informed consent processes. However, barriers to implementation include potential information overload for healthcare professionals, the need for ongoing training to understand and interpret the Model Facts labels, and the logistics of regularly updating and disseminating these labels. Overcoming these requires collaboration between healthcare providers, AI developers, and regulatory bodies to ensure that the information provided is relevant, accurate and actionable. The creation of a designated NHS Model Facts Assessment Unit would further alleviate this. By integrating dynamic risk communication into actual informed consent practices, it provides a mechanism for healthcare professionals to stay informed about the AI tools they use, thus empowering them to communicate risks more effectively to patients and make better-informed medical decisions.

6.3. Patient-Centric Risk Interpretation

This component is arguably the most direct response to the threat contextually biased medical AI poses to patients. It builds on established principles of personalised risk communication, advocating for providing patients with individualised information about the specific risks and benefits of AI-assisted recommendations (Han *et al.,* 2013). This addresses variations in AI performance that correlate with a patient's ethnicity, gender, age, and other factors often neglected in standard risk communication (Noul, 2024). The implementation of this step involves using Electronic Health Records, real-time patient-centred records that function as digital versions of patients' paper charts, to inform patients about how an AI system's output may be influenced by their unique health and demographic profile, predicting and explaining potential biases (Sokhack, 2023). This method goes beyond general explanations about AI functionalities and focuses on how its decision-making might exhibit biases when applied to their specific case. The effect of this is to create a more transparent informed consent process that is tailored to each patient's circumstances. While the previous component, dynamic risk communication, focuses on keeping healthcare professionals and patients informed about general updates in AI model risks and performance, this step concentrates on individualised communication. It requires healthcare professionals to convey personalised risk information in a manner that is understandable to the patient, potentially using proven tools such as customised printed materials, visual aids, or interactive media (Green, 2011). This approach upholds the true nature of informed consent as an instrument that enables patients to make their own health-related decisions (Astromskė *et al.,* 2021). Challenges in implementing this include the abstract nature of risk information and the time constraints of clinical practice (Han *et al.,* 2013). Despite these barriers, this remains a worthwhile initiative that represents a significant step towards countering the one-size-fits-all approach often seen in healthcare, particularly in the deployment of medical technologies (Noul, 2024).

6.4. Legal and Ethical Safeguards

The final part of the framework aims to protect physicians by establishing clear standards and guidelines for obtaining valid informed consent for use of medical AI. The traditional legal standard of care necessitates that physicians have a full understanding of all medical treatment and care options to effectively inform patients (Astromskė *et al.,* 2021). However, the complexity of medical AI introduces a higher level of difficulty in understanding and explaining these systems and their potential biases, often placing an unreasonable burden on healthcare professionals (Wang *et al.,* 2024). Instead, this paper recommends defining new standards that detail the necessary level of AI understanding for different roles within healthcare. These include *consumers*, clinicians who use AI tools in patient care, *translators*, who act as intermediaries between AI developers and clinical practitioners, and *developers*, who are responsible for the technical development of AI tools (Ng *et al.,* 2023). Consumers, forming the majority of the clinical workforce, must understand how to select and apply tools effectively and be equipped to discuss AI usage with patients within the informed consent process. Translators must ensure that AI tools are properly validated and integrated into clinical settings, making certain they are practical and safe for patient care. Developers, often with a background in both medicine and computer science, must ensure the efficiency of medical AI and work to reduce biases within them. By differentiating between these tiers, the exact duties of clinicians become clear and should be codified by regulatory bodies such as the Department of Health to give rise to corresponding legal and ethical responsibilities. The maintenance of material or physical risk comparisons is another critical aspect of these safeguards (BMA, 2024). These should be assessed in relation to AI Model Labels, electronic health records, and the patient's values, ensuring that treatment recommendations uphold principles of autonomy and that patient preferences drive decision-making (McDougall, 2019).

This section has outlined a four-part framework to enhance informed consent in AI-integrated healthcare, addressing the specific challenge of contextual bias. The first component, AI Literacy, equips healthcare professionals with core knowledge to understand and communicate the intricacies of AI to patients. Second, Dynamic Risk Communication, which seeks to introduce Model Fact labels for medical AI, ensures healthcare interactions allow for AI's evolving nature, maintaining informed consent as a continuous process. Third, Patient-Centric Risk Interpretation, directly addresses contextual bias by customising risk information to the individual patient's background, ensuring informed consent is not only comprehensive but also personalised. Finally, the framework incorporates Legal and Ethical Safeguards, which offer a structured approach to protect both patients and physicians. Collectively, these components move towards a more robust medical environment that remains patient-focused in the face of technological advancement.

Conclusion

The central research aim of this paper has been to explore how modifications to informed consent can address the challenges posed by contextual bias in medical AI, specifically focusing on the UK healthcare system but with implications for global practices. Unlike the perspectives offered by Cohen (2020), who advocates for long-term solutions such as reducing dataset biases, and Price (2019), who discusses the systemic nature of bias in AI deployment, this paper emphasises practical enhancements to informed consent procedures to mitigate contextual bias in the short to mid-term. In doing so, it extends and refines the debates initiated by these scholars, suggesting that immediate changes to informed consent practices can substantially complement long-term strategies. Moreover, by creating a four-part framework, this paper contributes a structured approach that actively engages with the complexities posed by medical AI. While this framework cannot solve the structural problems that give rise to contextual bias, it serves as both a response to the identified deficiencies in current practices, and an example for future adaptations in diverse healthcare settings worldwide. Future research should refine and explore implementation strategies for this framework as well as its applicability and adaptability in different national contexts and healthcare systems, which each have their own guidelines and cultural norms. Such initiatives mark a crucial step towards a future where medical AI not only advances healthcare outcomes but does so in a manner that is just, empathetic, and patient-centred.

References

Alharbi, A. *et al.* (2023) 'Factors Influencing the Implementation of Medicine Risk Communications by Healthcare Professionals in Clinical Practice', *Research in Social and Administrative Pharmacy*, 19(1), p. 50.

Article 35 - Healthcare (2015) *European Union Agency for Fundamental Rights*.

Astromskė, K. *et al.* (2021) 'Ethical and Legal Challenges of Informed Consent Applying Artificial Intelligence in Medical Diagnostic Consultations', *AI & Society*, 36(2), pp. 511, 512, 517.

British Medical Association (2024) *Guidance for Doctors on Patient Consent*, p. 8.

Burr, N.E. *et al.* (2023) 'Individualised Consent for Endoscopy: Update on the 2016 BSG Guidelines', *Frontline Gastroenterology*, 14(4), pp. 273-274.

Celi, L.A. *et al.* (2022) 'Sources of Bias in Artificial Intelligence that Perpetuate Healthcare Disparities — A Global Review', *PLOS Digital Health*, 1(3), p. 12.

Chu, C. *et al.* (2022) 'Digital Ageism: Challenges and Opportunities in Artificial Intelligence for Older Adults', *The Gerontologist*. Edited by S. Meeks, 62(7), p. 950.

Chu, C. *et al.* (2023) 'Age-Related Bias and Artificial Intelligence: A Scoping Review', *Humanities and Social Sciences Communications*, 10(1), p. 2.

Cirillo, D. *et al.* (2020) 'Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare', *Digital Medicine*, 3(1), pp. 2, 4, 7.

Cohen, I. (2020) 'Informed Consent and Medical Artificial Intelligence: What to Tell the Patient?', *The Georgetown Law Journal*, 108(6), pp. 1464, 1468.

Davenport, T. *et al.* (2019) 'The Potential for Artificial Intelligence in Healthcare', *Future Healthcare Journal*, 6(2), p. 95.

Daneshjou, R. *et al.* (2021) 'Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms', *JAMA Dermatology*, 157(11), p. 1363.

Daugherty, S. *et al.* (2017) 'Implicit Gender Bias and the Use of Cardiovascular Tests Among Cardiologists', *Journal of the American Heart Association*, 6(12).

Department of Health (2009) *Reference Guide to Consent for Examination or Treatment*, pp. 5, 13.

General Medical Council (2013) *Good Practice in Research and Consent to Research*.

Green, N. (2011) 'Artificial Intelligence and Risk Communication', *AI and Health Communication*, (Papers from the AAAI 2011 Spring Symposium).

Greaves, M. (2022) Gender Bias Revealed in AI Tools Screening for Liver Disease, *UCL News*.

Grote, T. *et al.* (2020) 'On the Ethics of Algorithmic Decision-Making in Healthcare', *Journal of Medical Ethics*, 46(3), p. 208.

Hall, D. *et al.* (2012) 'Informed Consent for Clinical Treatment', *Canadian Medical Association Journal*, 184(5), pp. 534, 536.

Han, P. *et al*. (2013) 'The Value of Personalised Risk Information: A Qualitative Study of the Perceptions of Patients with Prostate Cancer', *BMJ Open*, 3(9), pp. 1-2.

Iserson, K. (2024) 'Informed Consent for Artificial Intelligence in Emergency Medicine: A Practical Guide', *The American Journal of Emergency Medicine*, 76, pp. 225-228.

Jie, Z. *et al.* (2021) 'A Meta-Analysis of Watson for Oncology in Clinical Application', *Scientific Reports*, 11(5792), p. 1.

Kadam, R.A. (2017) 'Informed Consent Process: A Step Further Towards Making it Meaningful!', *Perspectives in Clinical Research*, 8(3), p. 107.

Kempt, H. *et al.* (2022) 'Relative Explainability and Double Standards in Medical Decision-Making', *Ethics and Information Technology*, 24(2), p. 20.

Kimiafar, K. *et al.* (2023) 'Artificial Intelligence Literacy Among Healthcare Professionals and Students: A Systematic Review', *Frontiers in Health Informatics*, 12(168), p. 2.

Krasniansky, A. (2019) *Understanding Racial Bias in Medical AI Training Data - Bill of Health*.

Krishnan, G. *et al.* (2023) 'Artificial Intelligence in Clinical Medicine: Catalysing a Sustainable Global Healthcare Paradigm', *Frontiers in Artificial Intelligence*, 6, p. 5.

Krive, J. *et al.* (2023) 'Grounded in Reality: Artificial Intelligence in Medical Education', *JAMIA Open*, 6(2), pp. 6, 7.

Licholai, G. (2023) *It's Time For 'Nutrition Labels' In Artificial Intelligence*, *Forbes*.

Liu, M. *et al.* (2023) 'A Translational Perspective Towards Clinical AI Fairness', *Digital Medicine*, 6, p. 3.

Mangalji, A. *et al.* (2019) 'Preparing for the Future of Medicine: Considering the Need for Data-Literate Physicians', *British Columbia Medical Journal*, 61(8), p. 326.

McDougall, R. (2019) 'Computer Knows Best? The Need for Value-Flexibility in Medical AI', *Journal of Medical Ethics*, 45(3), pp. 157-158.

Mittelstadt, B. (2021) *The Impact of Artificial Intelligence on the Doctor-Patient Relationship*. Council of Europe, pp. 46, 48, 49, 51.

Mittermaier, Mirja *et al.* (2023) 'Bias in AI-Based Models for Medical Applications: Challenges and Mitigation Strategies', *Digital Medicine*, 6(1), p. 1.

Neal, D. *et al.* (2022) 'Is There Evidence of Age Bias in Breast Cancer Healthcare Professionals' Treatment of Older Patients?', *European Journal of Surgical Oncology*, 48(12), p. 2401.

Ng, F.Y.C. *et al.* (2023) 'Artificial Intelligence Education: An Evidence-Based Medicine Approach for Consumers, Translators, and Developers', *Cell Reports Medicine*, 4(10), pp. 2, 8.

Norori, N. *et al.* (2021) 'Addressing Bias in Big Data and AI for Healthcare: A Call for Open Science', *Patterns*, 2(10), pp. 2, 3.

Noul, D. (2024) *Personalised Healthcare: Fusing Biomedical Engineering and AI for Customised Patient Care*.

O'Brien, N. *et al.* (2022) A*ddressing Racial and Ethnic Inequities in Data-driven Health Technologies*. Institute of Global Health Innovation, Imperial College London, pp. 7, 10, 11, 18.

Population of England and Wales (2022).

Price, N. (2019) 'Medical AI and Contextual Bias', *Harvard Journal of Law & Technology*, 33(1), pp. 66, 67, 68, 100.

Ricci Lara, M.A. *et al.* (2022) 'Addressing Fairness in Artificial Intelligence for Medical Imaging', *Nature Communications*, 13(1), p. 1.

Salway, S. *et al.* (2014) 'Race Equality and Health Inequalities: Towards More Integrated Policy and Practice', *Race Equality Foundation*.

Sendak, M. *et al.* (2020) 'Presenting machine learning model information to clinical end users with model facts labels', *Digital Medicine*, 3(1), p. 2.

Sokhach, D. (2023) *The Role of AI in Personalised Healthcare, Institute of Entrepreneurship Development*.

Stypińska, J. *et al.* (2023) 'AI Revolution in Healthcare and Medicine and the (Re)emergence of Inequalities and Disadvantages for Ageing Population', *Frontiers in Sociology*, 7, p. 4.

Terranova, C. *et al.* (2024) 'AI and Professional Liability Assessment in Healthcare. A Revolution in Legal Medicine?', *Frontiers in Medicine*, 10, p. 3.

The NHS Constitution for England (2023).

Ueda, D. *et al.* (2024) 'Fairness of Artificial Intelligence in Healthcare: Review and Recommendations', *Japanese Journal of Radiology*, 42(1), p. 4.

Van Kolfschooten, H. (2023) 'The AI Cycle of Health Inequity and Digital Ageism: Mitigating Biases Through the EU Regulatory Framework on Medical Devices', *Journal of Law and the Biosciences*, 10(2), p. 5.

Wang, Y. & Ma, Z. (2024) 'Ethical and Legal Challenges of Medical AI on Informed Consent: China as an Example', *Developing World Bioethics*, pp. 3, 4, 5.

What is an Electronic Health Record (EHR)? | *HealthIT.Gov* (2019).

Wood, E. *et al.* (2021) 'Are We Ready to Integrate Artificial Intelligence Literacy into Medical School Curriculum: Students and Faculty Survey', *Journal of Medical Education and Curricular Development*, 8, p. 1

Virtue Under Pressure: The Case for an Exemplarist Virtue Ethics Framework to Build Artificial Moral Agents for High-Pressure Tasks

Harry Collins
St. Edmunds College, University of Cambridge

This paper examines artificial moral agents (AMA) and seeks to justify their use to perform tasks involving high situational pressures that significantly impact human moral decision-making even when there is consensus on the correct decision. Moreover, these tasks can lead to moral injury for human decision-makers if their moral code has been violated, either by themselves or by external situational pressures. I argue that AMAs can potentially negate these concerns, particularly AMAs utilising exemplarist virtue ethics, a flexible approach to normative ethics, allowing agents to learn from experience to emulate the virtue of selected exemplars. To this end, I propose the outlines of an exemplarist framework for building AMAs using reinforcement learning from human feedback, where feedback is provided by moral exemplars in given tasks. Processes for selecting candidate tasks, identifying exemplars, and developing AMAs to emulate those exemplars are provided. Finally, potential objections are considered, both against the idea of exemplarist AMAs and their feasibility. I conclude that exemplarist AMAs for high-pressure tasks are promising candidates to perform high-pressure moral tasks, reducing moral injury for humans, although issues such as minimising cross-cultural disagreement on moral decisions and how well agents capture morally relevant features in their environment to emulate exemplars need further exploration through practical experimentation.

Introduction

As artificial intelligence systems become more advanced and require less human oversight, there are questions about whether such systems should be used for moral decision-making and over what kind of artificial moral agents (AMAs) should be developed (Wallach & Allen, 2008). Therefore, this paper aims to determine what kind of AMAs, if any, humanity should be striving to build. I will argue that AMAs are justified in tasks with high situational pressure that negatively impact human decision-making and can lead to moral injury (MI). Moreover, I will argue for an exemplarist virtue ethics approach to training AMAs to reproduce domain-specific moral exemplars' virtue.

Firstly, I will introduce AMAs and the kind I will consider. Next, I will overview virtue ethics, particularly the exemplarist approach. Then, I will address whether any AMAs should be used for moral decision-making, justifying them in high-pressure domains by showing that high situational pressure diminishes human virtuous behaviour, potentially also leading to MI. Next, I will suggest what kind of AMAs should be developed by presenting theoretical outlines for a virtue-based framework for training AMAs to emulate moral exemplars in high-pressure domains, showing a theoretical end-to-end implementation and addressing potential objections.

1. Defining AMAs

First, I will determine which AMA definition will be used. The term "artificial moral agent" was introduced by Wallach and Allen (2008, p. 4), stating that AMAs are robots that act independently from real-time human supervision and make moral decisions, meaning determining the right action based on ethical values.

However, determining to what degree a bot (software agent or physical robot) is a moral agent requires more specific definitions. For instance, a bot that alerts when you go over a speed limit and a vehicle-driving bot that must decide whose lives to prioritise in unavoidable crashes both have ethical impacts, yet the latter requires complex moral decision-making capabilities. Moor (2006, pp. 19-20) suggests four levels of moral agent:

1. Ethical impact agent: any machine that can have ethical consequences.
2. Implicit ethical agent: machines that reflect moral values without explicit representation.
3. Explicit ethical agent: machines that recognise and take morally relevant

information into account and can make explicitly moral decisions.
4. Full moral agents: machines with human-level moral agency.

Levels 1 and 2 lack the capacity for moral considerations. Level 3, however, is more advanced. Misselhorn (2022, p. 34) compares these AMAs to chess bots that recognise all chess-relevant information, evaluate it, and determine the best move. Explicit ethical agents recognise "general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations" (Moor, 2009, p. 12), meaning they can adapt to situations not explicitly accounted for, like how chess bots successfully evaluate novel positions. These AMAs can theoretically emulate moral decision-making without human moral decision-making features, like "consciousness, intentionality, and free will" (Misselhorn, 2022, p. 35) required for full moral agents. As there is much debate over defining consciousness and whether machines can achieve it (Searle, 1980; Himma, 2009), I will argue for explicit ethical agents which appear more practically achievable. Moreover, this avoids existential objections to AMAs as these agents do not require general superintelligence (Chalmers, 2010). They also lack emotions, so moral considerations towards them need not be considered. Next, I will overview virtue ethics as I will later argue for virtue-based AMAs.

2. Virtue Ethics

2.1. Overview

This section will explain virtue ethics and exemplarist approaches. Virtue ethics is a normative ethical theory focused on cultivating strong moral dispositions (virtues) like honesty and helpfulness to determine the right kind of person to be (Hursthouse & Pettigrove, 2023). Virtues must be learnt through repeated practice, so one honest act does not make someone honest, nor does a single lie make someone dishonest. Sometimes lying may be appropriate, like in the "Murderer at the door" thought experiment (Varden, 2010). Someone runs into your house to hide. A murderer then appears, asking if the would-be victim is inside. Although honesty is generally virtuous, lying would save the victim's life. To determine this, virtuous agents require phronesis, "the "practical wisdom" [...] learned by acting in social situations and gives those agents that possess this quality the ability to make new or novel judgments" (Sullins 2021, p. 136). Learning from experience is critical to habituating virtuous behaviour. I will now overview exemplarist virtue ethics as a learning approach, as I will later argue for its applicability to AMAs.

2.2.Exemplarist Virtue Ethics

Zagzebski's (2013) exemplarist approach suggests that, to become virtuous, moral exemplars (agents with admirable moral qualities, also called phronimos) must be identified and learnt from by observing their actions. Right and wrong moral actions are determined as follows: "a wrong act = an act that the phronimos characteristically would not do, and he would feel guilty if he did = an act such that it is not the case that he might do it = an act that expresses a vice = an act that is against a requirement of virtue (the virtuous self)". Rather than defining a set of virtues, exemplars embody virtue. Exemplars can be identified through admiration, defined as "attraction that carries the impetus to imitate" (Zagzebski, 2013, p. 201). After identifying potential exemplars, people should critically reflect on whether they are suitable to learn from. Further details will be given when applying this to AMA development. However, assuming moral exemplars can be identified, why not teach other humans to emulate them rather than AMAs? I will now seek to justify AMAs based on situationist critiques of virtue and MI.

3. If Any: How situationism and Moral Injury Justify AMAs in High-Pressure Domains

3.1. Situationism

Several debates surround whether AMAs should exist, with Formosa and Ryan (2021, p. 9) suggesting that debates should be more specific towards morally appropriate or inappropriate use cases. Hence, I will focus on a specific use case, showing how situational pressures and MI impact human moral decision-making, thus justifying AMAs in high-pressure domains/tasks. Firstly, I will discuss situationism, the argument that "variance in human behaviour is typically a function of the situation [...] rather than any traits of character" (Upton, 2009, p. 104), claims decisions are based more on situational pressures than universal virtues. Situational pressures are contextual factors that can influence behaviour, like the Milgram experiment (Milgram, 1963), where participants were instructed to electrically shock someone in another room if they got a word-pair recall question wrong, with shock intensity increasing each time. Participants were frequently urged to continue by an authority figure. The other person screamed in pain but was not actually shocked, unbeknownst to the participants. Over two-thirds of participants continued shocking after the other person feigned unconsciousness. Situationists argue that the authority figure's pressure explains why most participants inflicted deadly shock levels, as the large random sample negates the possibility that all participants were cruel. This study has been criticised for being unrealistic (Orne & Holland, 1968). However, participants believed their situation was real, giving genuine responses. For virtue ethicists, one cruel behaviour does not make someone cruel, although sufficient situational pressure clearly diminishes most participants' virtuous behaviour. Virtue ethicists may also argue that virtues cannot be simplified into behaviours (Kupperman, 2001). Whilst virtue is more about character than individual actions, virtuous agents should use phronesis to make good decisions, particularly when there is a clear consensus on the morally correct decision, as in situationist experiments.

Another example showing how time pressures can impact helping behaviours is the Good Samaritan experiment (Darley & Batson, 1973). Here, theology students travelled to a building to discuss the Good Samaritan parable under differing amounts of time pressure. One-third had to rush (high-pressure), another third were due to be just on time (medium-pressure), and another third had excess time (low-pressure). While travelling, participants encountered someone who needed help. Over 63% of low-pressure students helped versus sub-10% of high-pressure students, showing that as time pressure increased, helping decreased. Many situationist experiments similarly demonstrate virtuous behaviour decreasing as situational pressure increases (Alzola, 2008). Hence, some situationists argue that virtues have minimal behavioural impact versus situational pressure (Doris, 1998). Although some virtue ethicists dismiss this as oversimplifying virtue as behaviour in a single situation (Kupperman, 2001), high pressure clearly negatively impacts behaviour approximated as virtuous. However, this does not refute exemplarist virtue ethics, as a minority of participants exhibit exemplary moral behaviour despite situational pressures, like the 10% of high-pressure students who helped. Therefore, AMAs that learn from these exemplars would morally outperform most humans, justifying them. Furthermore, de Bruin *et al.* (2023) provide empirical evidence suggesting virtuous behaviours are more stable than situationists claim, demonstrating that the "ability to withstand the pressure and act virtuously is particularly present in mid-range situations" (470), so virtuous behaviour only diminishes under very high pressure for most humans. Hence, AMAs are justified where high situational pressure is expected because AMAs emulating exemplars displaying virtue despite high pressures would prioritise helping distressed individuals over timeliness or refuse to electrically shock someone when pressured to. Next, I will show how MI links to situationism, reinforcing this justification.

3.2. Moral Injury (MI)

Moral injury is "the strong cognitive and emotional response that can occur following events that violate a person's moral or ethical code" (Williamson *et al.*, 2021, p. 453), with these events either perpetrated or observed by that person. For instance, healthcare staff may feel unable to offer appropriate care if given inadequate supplies or when managing overly high workloads. MI can involve long-term feelings of shame, altered beliefs and self-destructive coping mechanisms. Coimbra *et al.* (2024) also associate it with an increased risk of suicidal ideation, burnout, depression, anxiety and Post-Traumatic Stress Disorder. Therefore, if someone suffers from MI, they may struggle to perform to previous standards, resulting in negative outcomes for them and anyone impacted by such lowered standards. For instance, 8-out-of-10 UK National Health Service (NHS) doctors suffered from MI during the COVID-19 pandemic due to increased workloads and pressure (Rimmer 2021, p. 1). The NHS also suffered record departures post-pandemic, citing stress and work-life balance issues (Savage, 2022; Best, 2021, p. 2). The remaining healthcare workers then face even more pressure due to understaffing. As shown by situationists, higher pressures diminish most people's moral decision-making, meaning higher chances of MI if a poor decision violates someone's morals. This creates a vicious cycle of increasing pressure, leading to a greater risk of MI and vice-versa. Therefore, AMAs are further justified in tasks where high situational pressure can be expected, as they can mitigate both the drop-off in humans' virtuous behaviour as situational pressure increases and MI risk. Other factors could also reduce situational pressures and MI risk, such as hiring more staff to reduce workloads in healthcare. However, high situational pressure can still arise, as shown by the pandemic. As high situational pressure is key to MI, both moral decision-makers and those affected by moral decisions could benefit from AMAs to further reduce situational pressures, meaning humans will be less likely to violate their moral codes.

4. What Kind: Establishing an Exemplarist Virtue Framework for AMAs

Having demonstrated that AMAs can be beneficial in domains with high situational pressure, I will argue for what kind of humanity should strive to build. As established, the focus will be on explicit ethical agents. Whilst lacking the human capacities necessary to be responsible for their actions, they can still perform morally desirable actions without human oversight (Anderson & Anderson, 2007, p. 19), either by entirely taking over tasks or by advising humans. Firstly, I will outline existing approaches to building AMAs and their issues before outlining an alternative exemplarist framework that addresses them. Then, I will defend this framework from potential criticisms like the frame problem, responsibility gaps, and relativism.

4.1.Current Approaches

AMAs are generally built in two ways: top-down and bottom-up (Allen & Wallach, 2009, p. 106). Top-down approaches impose an ethical theory onto a bot and are popular as they can be applied to existing systems. Deontological approaches add moral rules to follow, such as not to lie. Although seemingly intuitive, it disregards outcomes. For example, a self-driving car may be unable to avoid a low-speed crash, either with an elderly person ahead or with a young, athletic person to the side. The rule may be not to steer towards pedestrians, meaning the car would hit the elderly person. However, were it to steer into the young person, they would likely not be seriously injured, whereas the collision could kill the more fragile elderly person. Deontological approaches ignore this, meaning the AMA cannot learn from the outcome. Real-world examples of poor deontological outcomes include Google's Gemini image generator, where images of white people or images with historically accurate diversity levels could not be generated even when explicitly requested (Raghavan, 2024). The rules were implemented to enhance diversity yet had detrimental, offensive outcomes like portraying Nazis as black men, highlighting issues with situational inflexibility.

Alternatively, consequentialist approaches evaluate which action brings the best outcome. Winfield (2014) shows an implementation where a robot can predict all possible outcomes of actions taken in its environment. Its goal is to move to a point whilst avoiding a hole, but if a bot representing a human (H-bot) is likely to fall into the hole, it must prevent this by colliding with them. When a single H-bot headed towards the hole, the bot could predict outcomes and prevent them from falling in. The bot was successful 33-out-of-33 times. However, when a second H-bot was added, the optimal outcome became more difficult to compute. The experiment was again repeated 33 times. In 3 cases, both H-bots were saved; in 16, one was saved, but in 14 cases, neither were saved. Winfield explains that if the bot detects the second H-bot slightly after the first then the consequences being calculated completely change, with the bot "dithering" whilst considering new information. This highlights an issue: outcomes are not always certain, and new factors can drastically alter them, so outcomes alone cannot be relied upon. Related to this is the frame problem, where "potentially every new piece of information may have an impact on the whole cognitive system of an agent" (Misselhorn, 2022, p. 40), meaning it is difficult for top-down AMAs to discern what will change after an action. Therefore, top-down AMAs explicitly following moral rules or targeting specific consequences face fundamental challenges.

Bottom-up approaches do not explicitly employ ethical theories; instead, ethics are developed by learning from experience (Allen & Wallach, 2009, p. 107). This can be done via machine learning (ML) algorithms, giving machines "the ability to learn without explicitly being programmed" (Samuel, 1959). Modern ML models are very good at finding patterns in data and decisions to complete tasks at human levels or better, such as playing video games and language generation (Jordan & Mitchell, 2015). The frame problem is less significant as bottom-up agents learn from all data given to them, using prior learning experience to determine what data is morally relevant. However, without explicit ethical theories, explaining and controlling bottom-up agents' decisions is difficult due to the complexity of ML algorithms, making them impractical.

4.2. Towards an Exemplarist Virtue Framework for AMAs

Having shown issues with top-down and bottom-up approaches, I will now argue for a hybrid, virtue-based approach to AMAs that alleviates these issues. Allen and Wallach (2009, p. 107) suggest that virtue ethics represents a promising hybrid between both approaches, as virtues can be approximately represented top-down by behaviours, whilst moral character is cultivated by learning from experience akin to a bottom-up ML approach. Decisions are made using artificial phronesis where the AMA applies its learnt knowledge to new situations. Therefore, frame problem concerns are reduced as the AMA self-determines what information is morally relevant for emulation. Also, concerns over the inadequacy of general rules or potential outcomes alone are reduced, as virtue ethics enables agents to use their own judgment based on experience in any given scenario. However, few implementations have been attempted, largely due to questions over how to represent virtue (Vishwanath *et al.*, 2022, p. 666). Brewer (2009) argues that virtue is un-codifiable and can only address overall moral character. However, to cultivate virtue in an AMA, virtue must be learnt from virtuous acts, necessitating the approximation of virtue into measurable behaviours. Whilst some nuance is lost, if the AMA can learn to emulate virtuous behaviours, this is sufficient. The question of which virtues should be represented is also challenging, as virtues may be interpreted differently in different settings, and there is no universally agreed set of virtues.

To avoid these issues, I propose that to emulate exemplary human performance in high-pressure situations, an exemplarist framework for training AMAs is appropriate due to its practicality, and I will now show a theoretical outline for such a framework. Rather than explicitly modelling individual virtues, an exemplarist AMA would learn to emulate virtuous behaviour shown by moral exemplars in the target domain/role. The top-down element is that the AMA's ethics are based on moral exemplars exhibiting virtuous behaviours, whilst the bottom-up element is the learning process cultivating these behaviours. This would require human feedback during training, which modern techniques like Reinforcement Learning through Human Feedback (RLHF) make possible. RLHF enables artificial agents to act and learn directly from human feedback as to how desirable an action is (Christiano *et al.*, 2017). With this, an AMA could learn both to perform a general task and to emulate exemplary decision-making in high-pressure situations. Furthermore, the learning process avoids top-down approaches' situational inflexibility and means the AMA can learn from decisions resulting in poor outcomes. Moreover, as ML innovation continues, both phronesis and the data given to the AMA can become more complex and nuanced. For instance, large language models have shown emergent reasoning capabilities (Wang *et al.*, 2024), and Chella *et al.* (2020) show how giving AMAs continuous inner dialogue to explain their reasoning can improve artificial phronesis, also adding further explainability to AMAs, making it easier to understand and challenge their decisions. To avoid unwanted biases, bias auditing tools like Aequitas (Saleiro *et al.*, 2018) could be implemented. The AMAs can also be benchmarked and evaluated against the exemplars' to ensure similarity. Therefore, this framework is promising for implementing AMAs to make moral decisions in high-pressure domains as capabilities are measurable and can potentially increase in the future. Such AMAs are theoretically capable of performing tasks and making moral decisions when required without having to explicitly determine whether the situation is morally challenging or not, as they always seek to emulate their exemplars. Depending on the use case, they could either advise humans or take over tasks. For instance, driving AMAs would drive a car to get to a location, yet would also emulate exemplars' virtue in moral scenarios like unavoidable collisions by determining the exemplar's most likely action. Having outlined how exemplarist AMAs could be implemented, a process is required to identify suitable high-pressure domains where moral decision-making is required, and AMA effectiveness is heavily dependent on selecting high-quality moral exemplars, so I will now elaborate on these elements.

4.3. Selecting Suitable Domains and Exemplars

To identify domains/tasks where high situational pressures diminish moral decision-making, situationist-style analysis can help analyse whether this occurs by setting tasks where situational pressures increase until the task's highest pressure levels are tested. As shown previously, NHS workers reported high MI levels, largely due to situational pressures impacting decision-making, thus highlighting this domain's potential suitability. Driving is another domain where high situational pressure like traffic levels, weather and time-to-react impact decision-making (Soares *et al.*, 2021) and where people make moral decisions that differ depending on situational pressure (Johnson *et al.*, 2023). Therefore, researchers should first seek domains where MI or high situational pressure have been reported. Then, interviews could be conducted with individuals operating in that domain to identify whether there are tasks that involve high-pressure moral decision-making, cause MI, and have measurable virtuous behaviours. If so, situationist-style experiments can be established to measure whether moral decision-making ability diminishes at high pressure. Having shown how to identify potentially suitable domains, I will now illustrate how exemplars could be identified.

Zagzebski (2013) argues that moral exemplars can be identified through admiration, and verifying that they are worthy of imitation. For full, generally intelligent moral agents, locating universal exemplars is difficult. However, for specific domains/tasks, situationist-style experiments can identify exemplary behaviour under pressure. For healthcare workers, this may involve asking staff which of their peers they admire and then verifying whether they are exemplary by, for example, analysing patient satisfaction surveys during periods of very high demand. For driving, Johnson *et al.* (2023, p. 6) suggest that prosocial, cooperative drivers generally align with virtuous traits like benevolence and end up in accidents far less frequently than other drivers. They also found that most drivers' decision to self-sacrifice or self-preserve in unavoidable accident scenarios changed depending on time-to-react, testing this by asking what they would do in a survey and then in real-time simulations. 76.8% of participants self-sacrificed in the survey, but only 22.8% of participants self-sacrificed in both survey and simulation. This highlights that participants admired self-sacrifice, so exemplars can be identified by displaying admirable behaviour under pressure like those 22.8% of participants. Therefore, to identify exemplars, interviews should be conducted with those performing the selected task, asking which peers they admire morally. This could be used in conjunction with situationist-style experiments. Once identified, exemplars' suitability must be verified via audit to ensure they lack unwanted biases, e.g. underestimating black hospital patients' needs (Obermeyer *et al.*, 2019). Having outlined a theoretical framework for virtuous AMA, I will demonstrate a theoretical end-to-end AMA implementation.

4.4. Theoretical Implementation

Having presented the framework, I will show a simple theoretical implementation based on Winfield's (2014) consequentialist bot experiment. The initial set-up is the same, where the moral agent's task is to travel to a point whilst avoiding a hole, and if they notice that a human might fall into the hole, they should display helping behaviour by colliding with them to prevent the fall. Firstly, interviews would be conducted to establish the task, whether high situational pressures can occur, whether there is a risk of MI, and whether there are clearly measurable virtuous behaviours. Here, the task of moving to a point is simple, with the potential for high situational pressure with humans heading towards a hole in the ground. Failing to save someone due to the pressure of the situation could lead to MI, and there is a clearly virtuous behaviour of not hesitating and saving as many humans as possible. Next, situationist tests would be devised to determine whether there is sufficient situational pressure to diminish most humans' moral decision-making skills. For instance, participants are faced with saving one human, then two for heightened situational pressure. Assuming these results are the same as Winfield's results for the consequentialist bot, all participants save the human in the low-pressure scenario. In the high-pressure scenario, 14-out-of-33 participants save no humans, 16-out-of-33 rescue one, and 3-out-of-33 rescue both. Here, virtuous behaviour diminishes with increased pressure, so the task is suitable for AMAs. Exemplars can also be identified as the 3 participants who saved both humans. They may be verified by analysing whether they consistently display these behaviours when performing similar tasks. Next, potential exemplars are audited for unwanted biases. Then, an environment is created with as much information as possible for an AMA to be trained via RLHF to perform the task and similar tasks, such as different routes with different numbers of humans. Chosen exemplars give feedback as to whether the decisions made by the AMA align with what they would do. Once trained, tested and measured against exemplars, the AMA should be able to independently perform the task whilst efficiently processing moral dilemmas that may occur, like which humans to prioritise if not all can be saved or whether multiple humans can be saved, without dithering. It would demonstrate its exemplars' virtuous behaviours without individual virtues being explicitly programmed, and the lack of hard coding means it may more easily adapt to new situations than Winfield's consequentialist bot. Having illustrated a theoretical implementation, I will defend this framework against potential objections that have yet to be considered.

4.5. Objections to the Framework

A key objection regards cultural disagreement over who exemplars are, as different cultures may admire different behaviours (Kotsonis, 2020, p. 228). This is highlighted by Awad *et al.*'s (2018) global survey of responses to moral dilemmas for driving. Whilst some moral preferences were global, many differed culturally, such as the propensity to spare those obeying traffic laws versus jaywalkers. However, whilst cultural preferences may vary, the core approach of developing strong moral character remains, and this framework's goal is not to solve all moral dilemmas universally but to emulate virtuous exemplars whose moral decision-making ability in specific domains withstands high situational pressures. Indeed, Macintyre (1981) argues that virtues must be interpreted by the community using them, and Zagzebski (2013) states that "identification of exemplars is revisable" (p. 200), so exemplars can differ by culture. However, this can raise objections regarding moral relativism, meaning if morals are relative to cultural attitudes, there is no objective morality. Basing moral judgements on exemplars within cultures and domains can seemingly support relativism (Kotsonis, 2020, p. 229). However, this framework does not claim that exemplar's actions are always correct, but that they exhibit virtuous behaviours in specific tasks/domains. For instance, whilst cultural preferences for whose treatment healthcare workers should prioritise may differ, exemplars should still be generally virtuous, e.g. kind and helpful, without their decision-making ability diminishing under pressure, like not neglecting patients despite high stress. Such a virtuous nature is universal, although cultures may interpret specific virtues differently. Macintyre (1981) suggests that reflecting on virtue enables the changing of morals for societies, so universal moral truths can be gradually built towards. Although this does not fully refute relativist objections, this framework's purpose is only to match human morality, not to exceed it, and whilst this significantly challenges the feasibility of cross-cultural AMAs, localised solutions or AMAs designed for specific tasks where there is cross-cultural consensus are still possible.

There may also be objections regarding moral deskilling. Vallor (2015) suggests that offloading tasks to AMAs can result in losing the moral skills required for the task. This would be a major issue if AMAs were to take over too many responsibilities from humans. However, explicit ethical AMAs cannot be responsible for their actions, so humans must critically evaluate them constantly to ensure that they are performing similarly to exemplars, and they require consistent human feedback. Therefore, although some tasks may be passed to AMAs, exemplars will still need to teach them, and these AMAs should only be used where most humans' moral decision-making is already poor.

A related concern is responsibility for AMAs' mistakes. As this framework does not involve full moral agency, responsibility should fall jointly between all parties developing the AMA. However, Sparrow (2007) demonstrates the possibility of responsibility gaps occurring when an AMA is not designed to break an ethical code but does so unforeseeably without human oversight. Therefore, nobody appears responsible for the AMA's action. This should be combated by ensuring that domains/tasks are narrow enough that most general moral scenarios can be addressed in the AMA's training. Then, whether the risks of an AMA failing are worth the potential benefits must be carefully evaluated.

Another objection may be that ML can perpetuate unwanted biases held by exemplars, such as racism or sexism (Fazelpour & Danks 2021). High-profile examples include the aforementioned Gemini case, so experts in domains besides the exemplars', like critical race scholars, feminist theorists and philosophers, should be involved in the selection and training process to ensure potential biases are found and eradicated before deployment. Ultimately, this is not a reason to avoid this approach, but it shows that great care should be taken to avoid perpetuating biases in these systems.

Finally, I will address possible objections to the proposed ML approach. One objection may be that ML algorithms cannot guarantee outputs (Kläs & Vollmer, 2018), so they will not always make decisions in line with their exemplars. Whilst true, certain decisions can be guaranteed by hardcoding deontological rules that override the ML output to comply with certain laws or regulations, such as never deactivating a life support system. Also, situationism shows that human decisions cannot be guaranteed under high pressure, so AMAs emulating exemplars would be more consistent in this regard. This is ultimately virtue ethics' goal, to promote strong moral character, not necessarily to always make the correct moral decision, and AMAs can learn from their mistakes via RLHF to constantly improve. A final practical objection may be that training environments cannot offer all the relevant moral information needed for moral decision-making. However, exemplars could suggest important moral features to capture for given tasks, and practical experimentation is required to determine how specific a task must be and how much information is required for an AMA to accurately emulate exemplars in that task.

5. Conclusion

I have argued that AMAs can be justified by highlighting a specific area where they can be beneficial whilst avoiding existential and feasibility concerns, demonstrating how humans' virtuous behaviours diminish under high situational pressures, potentially leading to MI, therefore justifying AMAs that can match exemplary human performance under high pressure. Additionally, I showed the suitability of an exemplarist, a virtue-based framework for building AMAs to perform moral tasks where high situational pressure impacts human performance and presented a theoretical implementation. Future work could build on and practically test this framework and experiment with training approaches, such as asking exemplars to imagine they are machines when giving training feedback because human and machine morals may not always align. For example, in Winfield's (2014) experiment, if a human were the moral agent preventing others from falling in the hole, self-preservation may also be a factor. However, for this level of AMA, there is no self to preserve, enabling different potential actions like jumping into the hole to reduce the falling distance. Therefore, practical experimentation is required to further develop AMAs, but overall, this paper presents a clear justification and an outline of a theoretical framework for practically applying exemplarist virtue ethics to AMAs.

References

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms or abdication of human responsibility. *Robot ethics: The ethical and social implications of robotics*, pp. 55-68.

Alzola, M. (2008). Character and environment: The status of virtues in organisations. *Journal of Business Ethics,* 78, pp. 343-357.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine,* 28(4), pp. 15-15.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature,* 563(7729), pp. 59-64.

Best, J. (2021). Undermined and undervalued: how the pandemic exacerbated moral injury and burnout in the NHS. *BMJ,* 374(1858).

Brewer, T. (2009). *The retrieval of ethics.* Oxford University Press, USA.

Chalmers, D. J. (2016). The singularity: A philosophical analysis. In *Science fiction and philosophy: From time travel to superintelligence,* pp. 171-224.

Chella, A., Pipitone, A., Morin, A., & Racy, F. (2020). Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7, 16.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems,* 30.

Coimbra, B. M., Zylberstajn, C., van Zuiden, M., Hoeboer, C. M., Mello, A. F., Mello, M. F., & Olff, M. (2024). Moral injury and mental health among health-care workers during the COVID-19 pandemic: meta-analysis. *European Journal of Psychotraumatology*, 15(1), 2299659.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27(1), pp. 100.

de Bruin, B., Zaal, R., & Jeurissen, R. (2023). Pitting virtue ethics against situationism: An empirical argument for virtue. *Ethical Theory and Moral Practice,* 26(3), pp. 463-479.

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous*, 32(4), pp. 504-530.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760.

Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *AI & society*, 36(3), pp. 839-851.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11, pp. 19-29.

Hursthouse, R., & Pettigrove, G. (2023). Virtue Ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Fall 2023)*. Metaphysics Research Lab, Stanford University. https://plato-stanford-edu.ezp.lib.cam.ac.uk/archives/fall2023/entries/ethics-virtue/

Johnson, K. A., Berman, S., Pavlic, T. P., Ulhas, S. S., Elkins, J. K., & Ravichander, A. (2023). Virtuous Vehicles: Identifying the Values Profiles of Human Drivers as a Basis for Programming Virtuous Decision-Making in Self-driving Cars.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp. 255-260.

Kläs, M., & Vollmer, A. M. (2018). Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In *Computer Safety, Reliability, and Security: SAFECOMP 2018* Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37 (pp. 431-438). Springer International Publishing.

Kotsonis, A. (2020). On the limitations of moral exemplarism: Socio-cultural values and gender. *Ethical Theory and Moral Practice*, 23(1), pp. 223-235.

Kupperman, J. J. (2001). The indispensability of character. *Philosophy*, 76(2), pp. 239-250.

MacIntyre, A. (2013). *After virtue.* A&C Black.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology,* 67(4), pp. 371.

Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible*

Artificial Intelligence: Interdisciplinary Perspectives. Cambridge: Cambridge University Press, pp. 31–49.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems,* 21(4), pp. 18-21.

Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now,* 72, pp. 12-14.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), pp. 447-453.

Orne, M. T., & Holland, C. H. (1968). On the ecological validity of laboratory deceptions. *International Journal of Psychiatry,* 6(4), pp. 282-293.

Raghavan, P. (2024, February 23). Gemini image generation got it wrong. we'll do better. *Google*. https://blog.google/products/gemini/gemini-image-generation-issue/

Rimmer, A. (2021). Covid-19: Eight in 10 doctors have experienced moral distress during pandemic, BMA survey finds.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint* arXiv:1811.05577.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), pp. 210-229.

Savage, M. (2022, February 26). Stressed NHS staff in England quit at record 400 a week, fuelling fears over care quality. *The Guardian*. https://www.theguardian.com/society/2022/feb/26/stressed-nhs-staff-quit-at-record-rate-of-400-a-week-fuelling-fears-over-care-quality

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences,* 3(3), 417-424.

Soares, S., Lobo, A., Ferreira, S., Cunha, L., & Couto, A. (2021). Takeover performance evaluation using driving simulation: a systematic review and meta-analysis. *European Transport Research Review,* 13, pp. 1-18.

Sparrow, Robert (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), pp. 62–77.

Upton, C. L. (2009). Virtue ethics and moral psychology: The situationism debate. *The Journal of Ethics*, 13(2), pp. 103-115.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28, pp. 107-124.

Vishwanath, A., Bøhn, E. D., Granmo, O. C., Maree, C., & Omlin, C. (2023). Towards artificial virtuous agents: games, dilemmas and machine learning. *AI and Ethics,* 3(3), pp. 663-672.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., ... & Yang, H. (2024). Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint* arXiv:2401.06805.

Williamson, V., Murphy, D., Phelps, A., Forbes, D., & Greenberg, N. (2021). Moral injury: the effect on mental health and implications for treatment. *The Lancet Psychiatry*, 8(6), pp. 453-455.

Winfield, A. F., Blum, C., & Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection. In *Advances in Autonomous Robotics Systems: 15th Annual Conference,* TAROS 2014, Birmingham, UK, September 1-3, 2014. Proceedings 15 (pp. 85-96). Springer International Publishing.

Zagzebski, L. (2013). Moral exemplars in theory and practice. *Theory and Research in Education,* 11(2), pp. 193-206.

CJAI

CAMBRIDGE
JOURNAL OF
ARTIFICIAL
INTELLIGENCE