AIFS0050

Written evidence submitted by The Centre for Protecting Women Online

Dr. Ángel Pavón Pérez, Prof. Miriam Fernandez and Prof. Olga Jurasz on behalf of the Centre for Protecting Women Online.

The Centre for Protecting Women Online[1] is a vehicle for understanding and addressing challenges posed to women's online safety through a novel, interdisciplinary and ambitious research agenda. It develops cross-sectoral, collaborative outputs and interventions which inform law, policy, technology development and practice to reduce online harms suffered by women and girls; minimise anti-social behaviours online whilst promoting pro-social behaviours and help build tech/software that helps ensure accountability, credibility and helps facilitate justice.

The Centre is comprised of five interwoven work streams:

- Law & Policy,
- Human Behaviour,
- The Future of Responsible Tech,
- Ethical and Responsible Tech/AI and
- Policing.

Additionally, the Centre has partnered with numerous organisations to help achieve its core objectives.[2] We are pleased to have the opportunity to respond to Public Accounts Committee inquiry in relation to AI in financial services. Our submission draws on our expertise in online safety and responsible AI to address how AI systems in financial services can impact consumer outcomes, with particular attention (but not limited) to fairness towards women. Should you require further information we would be happy to assist you by providing oral evidence and engaging in further policy development.

Summary

As Artificial Intelligence (AI) becomes more embedded in financial services—from credit scoring to customer support—its potential to improve consumer outcomes is growing rapidly. But with this promise comes risk, especially for women and other marginalised groups.

In this response to the UK Parliament's Public Accounts Committee inquiry on AI in financial services, we at the Centre for Protecting Women Online highlight both the opportunities and challenges AI presents. Done well, AI can enhance financial inclusion by analysing non-traditional data to support underserved consumers. It can also personalise services, increasing accessibility for those in remote areas or with disabilities.

However, if left unchecked, AI risks reinforcing historical biases. AI models trained on historical financial data can inadvertently replicate and even amplify embedded discriminatory practices. With data reflecting decades of prejudice, particularly against women and

[1] The Centre for Protecting Women Online, (2024) available at: https://university.open.ac.uk/centres/protecting-women-online/#:~:text=About%20the%20Centre,interdisciplinary%20and%20ambitious%20research%20agenda

[2] 'Partnerships' (The Centre for Protecting Women Online) available at: https://university.open.ac.uk/centres/protecting-women-online/partnerships

minoritised groups, these systems risk reinforcing disparities. For example, studies have revealed that gender biases can emerge in AI-driven credit scoring through proxy variables such as part-time work patterns—a reflection of remaining societal inequities—and ultimately result in biased decisions against women. Worse still, even if AI is initially unbiased, AI can drift over time or be manipulated, requiring constant monitoring and adjustment.

To address these risks, the response emphasises the importance of both designing AI with built-in fairness and ensuring exhaustive and continuous oversight. There is a strong case for sharing and using sensitive data—not to discriminate, but to audit and mitigate bias. Controlled access to information like gender or ethnicity, accompanied by robust security measures, could enable auditors to validate and correct unfair outcomes. Regular algorithmic audits, transparent decision-making processes, and clear consumer recourse mechanisms are vital safeguards.

Another challenge lies in defining fairness itself. Competing fairness definitions often conflict, and often it’s impossible to satisfy them all at once. Without clear regulatory guidance, financial institutions are left to make difficult ethical trade-off decisions alone. Therefore, we emphasise that clear guidance is needed on which definitions of fairness to apply and what level of bias, if any, can be tolerated. Decisions about these standards should involve collaboration between responsible AI experts, domain specialists (such as financial professionals), and regulators.

With deliberate safeguards and collaboration between regulators, developers, and researchers, AI can become a tool not just for innovation—but for equity. As the UK shapes its AI future, fairness must be built in from the start.

What benefits to consumers might arise from using AI in financial services? For example, could AI be used to identify and provide greater assistance to vulnerable consumers?

- **Expanded Financial Inclusion:** AI can support broader financial inclusion by removing traditional barriers to access. For individuals with mobility challenges or those living in remote or underserved regions, AI-powered tools—such as mobile banking apps, voice recognition systems, and automated customer service—can offer accessible, user-friendly financial services without the need to travel to physical branches. This not only empowers users to manage their finances more independently but also opens opportunities for economic participation that were previously out of reach.
- **Personalised Services:** AI chatbots and virtual assistants can provide 24/7 support, helping customers (including those with mobility impairments or in remote areas) access information in their native languages. However, if not properly implemented, these services may pose risks. For example, while chatbots may be suitable for handling procedural queries, they might be inappropriate for offering advice on complex matters like mortgages or investments. Incorporating a human-in-the-loop approach could help mitigate these concerns.

What is the risk of AI increasing embedded bias? Is AI likely to be more biased than humans?

- **Historical Bias in Data Feeding AI:** AI models learn from historical financial data, which have shown to reflect decades of biased practices against women and other marginalised groups.

  Bias in AI refers to systematic and unfair discrimination that can arise when AI models learn from biased or unrepresentative data. Since these models are trained on historical data, any biases present in the data—whether due to historical social inequalities, inadequate collection methods, or inaccurate labelling- can be replicated and even amplified.[3] This can lead to disproportionate harm for minoritised groups (e.g. women or individuals from minority ethnic backgrounds), who are often underrepresented or misrepresented in the data, further exacerbating existing disparities and marginalising vulnerable populations in critical decision-making processes.[3]

  Examples of AI Bias in Finance: In 2021, Women's World Banking, in collaboration with the University of Zurich, audited credit processes for gender bias in India, Mexico, and Colombia. They found "reject inference bias" in all three markets, meaning that a substantial subset of women who were rejected for loans should, in fact, have been granted them.[4] Similarly, Amazon discovered bias in its AI-driven recruitment system in 2018, which systematically downgraded resumes containing keywords associated with women, reflecting historical hiring biases embedded within the training data —and in turn limiting women's access to financial opportunity through high-quality employment.[5] This resulted in the inadvertent exclusion of qualified female applicants, underscoring how biased historical data can perpetuate existing inequities through automated decision-making systems.

- **Use of Sensitive Data and Its Proxies to Train AI:** If the data is biased, using information such as gender or its proxies (e.g. maternity leave) can result in biased AI outcomes. Even seemingly neutral data that do not contain sensitive information can be correlated with protected characteristics (e.g. gender). If AI is trained in this biased data, it can lead to discrimination even when the AI model doesn't explicitly see attributes like gender or ethnicity. For example, an algorithm can pick up on attributes like hours worked as a proxy for gender (since women are far more likely to work part-time).[6] This could result in biased outcomes, such as a loan approval system automatically downgrading applicants who work fewer hours, assuming lower financial reliability, and thereby disproportionately affecting women. These proxy effects mean an AI system can **recreate bias** against women or minoritised groups even when it

[3] Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9). https://dl.acm.org/doi/fullHtml/10.1145/3465416.3483305

[4] HARNESSING THE POWER OF DATA: Inclusive Growth and Recovery Challenge Impact Report. Data.org. 2020. https://data.org/reports/challenge-impact/

[5] Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. October 2018. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

[6] Women and the UK economy. House of Commons Library. 2025. https://commonslibrary.parliament.uk/research-briefings/sn06838/

isn't "intending" to, simply because certain inputs serve as indirect representations for protected characteristics.[7]

Example of AI Bias Amplification Due to Proxies in Finance: Women's World Banking's research found that a lending AI in emerging markets used mobile phone GPS data on how long a business owner stays at their shop – a metric biased against women entrepreneurs who spent more time working from home due to caregiving duties. As a result, many creditworthy women were flagged as less creditworthy compared to men, a clear case of an AI unintentionally penalising women via a proxy variable.[8]

Another example of AI Bias Amplification Due to Proxies**:** For example, in healthcare, AI models trained on biased data can exacerbate existing racial disparities. A widely used U.S. healthcare algorithm underestimated the health needs of Black patients by using healthcare spending as a proxy for health risk.[9] This approach introduced significant racial bias because Black patients, who often face systemic barriers to accessing healthcare, tend to have lower medical expenditures despite having similar or worse health conditions compared to White patients. As a result, the algorithm inaccurately concluded that Black patients were healthier, leading to under-treatment and delayed diagnoses, with the number of Black patients eligible for additional care reduced by half. The underlying issue was that healthcare costs, influenced by racial inequities in access and care, are a poor stand-in for actual health needs. When the researchers replaced this cost-based proxy with direct measures of health status, such as the number of chronic conditions, the bias was reduced by 84%, highlighting the profound impact that biased data and proxies can have on AI-driven decisions.

- **Difficulties in Defining Bias:** One significant risk lies in how fairness is defined. There are multiple ways to define what constitutes a fair and unbiased AI system, and these definitions often involve trade-offs. For example, consider an AI system that decides whether to grant loans. One definition of fairness might require that the same percentage of loans be approved for men and women. Alternatively, fairness could be defined in terms of equal error rates—such as ensuring that the system denies loans to eligible applicants (false negatives) at the same rate for both genders.

  Furthermore, it is mathematically impossible to satisfy all fairness definitions simultaneously[10]—except under highly idealised conditions, which are rarely achievable in practice, or when some degree of bias is considered acceptable.[11]

[7] Pavón Pérez, Á., Fernandez, M., Al-Madfai, H., Burel, G., & Alani, H. (2023, April). Tracking Machine Learning Bias Creep in Traditional and Online Lending Systems with Covariance Analysis. In *Proceedings of the 15th ACM Web Science Conference 2023* (pp. 184-195).

[8] Innovative AI for Women's Financial Inclusion. Data.org. https://data.org/stories/womens-world-banking/

[9] 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), 447-453.

[10] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[11] Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023, June). The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 400-422).

Therefore, clear guidance is needed on which definitions of fairness to apply and what level of bias, if any, can be tolerated. Decisions about these standards should involve collaboration between responsible AI experts, domain specialists (such as financial professionals), and regulators.

- **Other risks emerge after AI is deployed**: Even if initially unbiased, AI can become biased over time due to changes in data distribution[12] or adversarial attacks.[13] Monitoring frameworks are necessary to ensure AI remains fair and unbiased.

- **AI vs Human Bias – Who's Worse?:** AI is not inherently more biased than humans, but it **can apply bias more consistently and opaquely**. A human loan officer might have individual prejudices, but an AI trained on biased data will systematically apply those prejudices to every decision, potentially impacting far more people. For instance, some fintech lenders have achieved gender-neutral outcomes by consciously engineering and testing their algorithms for fairness. In one audited digital credit model in India, women were equally likely as men to be approved and received similar loan terms; the few slight disparities were traced to fewer women applying, not the AI's treatment of gender.[14]

  Thus, AI systems have the potential to exhibit reduced bias compared to some human decision-making when responsible AI principles and guidelines—such as ensuring fairness, enforcing transparency, incorporating rigorous auditing practices, and maintaining human oversight—are rigorously implemented. However, these positive outcomes require deliberate and continuous efforts. By default, an AI system will reflect the patterns—whether fair or biased—that exist in its training data. In general, the risk is that without these responsible AI safeguards, biased AI can scale discrimination faster and more widely than individual human decision-makers, and its complexity can make bias harder to detect. Thus, implementing robust responsible AI frameworks is essential to ensure that AI not only avoids amplifying historical biases but actively contributes to equitable outcomes.

  Furthermore, even the inclusion of human oversight—often referred to as a "human-in-the-loop" or mixed AI-human decision-making approach—comes with its own challenges. One key risk is automation bias,[15] where human reviewers may over-rely on AI outputs, accepting them without sufficient scrutiny, especially when the system appears sophisticated or consistently confident. On the other hand, algorithm aversion[16] can occur when humans under-rely on AI, disregarding its input even when it performs better than human judgement overall. These dynamics can undermine the benefits of combining AI and human decision-making, leading to either the unchecked propagation of algorithmic bias or the reintroduction of human subjectivity and inconsistency. Therefore, human oversight must be carefully designed, with clear roles, proper training, and decision-making accountability to ensure it enhances rather than compromises the fairness and accuracy of AI-supported systems.

[12] Shao, M., Li, D., Zhao, C., Wu, X., Lin, Y., & Tian, Q. Supervised Algorithmic Fairness in Distribution Shifts: A Survey.

[13] Mehrabi, N., Naveed, M., Morstatter, F., & Galstyan, A. (2021, May). Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*(Vol. 35, No. 10, pp. 8930-8938).

[14] In a world of gender bias, Lendingkart's AI-based credit model stands apart. Women's World Banking. 2022. https://www.womensworldbanking.org/insights/in-a-world-of-gender-bias-lendingkarts-ai-based-credit-model-stands-apart

[15] Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In Decision making in aviation (pp. 289-294). Routledge.

[16] Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

What data sharing would be needed to make AI more effective in financial services, and will there be a need for legislative change to achieve that?

While the effectiveness of AI systems in financial services often depends on access to rich and diverse datasets, sharing such data must be done carefully and ensure safeguards for individual privacy. Even seemingly innocuous variables can carry significant implications when they act as proxies for protected characteristics. Although much of this data can be shared under the right conditions, our focus here is solely on sensitive characteristics like gender, and proxies such as maternity leave, due to their central role in identifying and mitigating algorithmic bias.

- **Use of Sensitive Information and Its Proxies to Train AI:** In many jurisdictions, including the United States, laws such as the Equal Credit Opportunity Act and the Fair Housing Act have prohibited the use of sensitive personal data—such as race, gender, or age—in lending or insurance decisions, and sometimes restrict even the storage of such data to avoid direct discrimination.[17] The European Union's AI Act (2024) limits the use of sensitive attributes in high-risk AI systems, typically allowing them only for fairness auditing and bias mitigation.[18] UK law does not outright forbid including sensitive information in training or analysis if the aim is to detect or prevent bias rather than to discriminate. In fact, regulators encourage using such data in controlled ways to audit algorithms. The Information Commissioner's Office (ICO) advises that organizations may need datasets containing protected characteristics to test how their AI systems perform for different groups and to retrain models to avoid discriminatory effects.[19]

  Some state-of-the-art fairness-aware machine learning approaches, which can require sensitive data during AI training to measure and reduce bias (although other bias mitigation approaches do not require sensitive information during the training). Importantly, excluding sensitive features (a technique often called "fairness under unawareness") does not imply eliminating bias, as there might be other sources of biases. For example, as mentioned before, other variables in the data may act as proxies—for example, names, postcodes, or education histories might correlate strongly with ethnicity or gender. These proxy variables can inadvertently reproduce discriminatory patterns, even if the sensitive attributes themselves are removed. Therefore, sensitive information should be shared and used exclusively, and with the appropriate safeguards, to audit and mitigate AI bias. Organisations should also consider other sources of biases, and treat potential proxies for sensitive information with caution during AI development and consider assessing their impact during model training and evaluation.

- **Use of Sensitive Information to Audit AI:** Furthermore, not having access to sensitive data can actually make it harder to detect and correct bias.[20] Most state-of-

[17] Equal Opportunity Act. Sandra F. Braunstein, Director, Division of Consumer and Community Affairs. 2008. https://www.federalreserve.gov/newsevents/testimony/braunstein20080717a.htm

[18] EU AI Act, Article 10. https://artificialintelligenceact.eu/article/10/

[19] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination

[20] How Anti-Discriminatory Measures Can Worsen AI Bias Anton S. Ovchinnikov, INSEAD and Queen's University, Canada. 2023. https://knowledge.insead.edu/operations/how-anti-discriminatory-measures-can-worsen-ai-bias

the-art approaches for identifying and mitigating bias in AI require access to such sensitive information.[21] Thus, sensitive information and its proxies should only be used to audit bias or in bias mitigation strategies. Furthermore, strong security measures should be implemented to protect this data, and to ensure that is only used for these safeguarding purposes. For example, third parties could be responsible for storing it, allowing access only for auditing and bias mitigation purposes.

- **Balancing Privacy and Open Data:** Another key challenge in AI and finance research is the lack of open financial data. Measures should be put in place to enable researchers to study financial AI systems while ensuring that data can be audited more easily by regulators. This would foster and advance research not only about the study of risks of AI in finance, but also about the development of mitigation strategies for those risks.

What sort of safeguards needs to be in place to protect customer data and prevent bias?

- **Bias identification:** As previously mentioned, there are various fairness metrics used to measure bias in AI, many of which involve trade-offs. Before developing AI systems and planning mitigation strategies, it is crucial to clearly define the fairness objectives, aligning with applicable legislation where relevant.

- **Bias mitigation strategies:** A wide range of bias identification and mitigation approaches have been proposed in the AI literature.[22] These approaches are typically classified based on the stage at which they are applied:
  - *Pre-processing approaches*, which are applied before training (e.g. modifying or balancing the data);
  - *In-processing approaches,* which are integrated during training (e.g. optimizing for fairness alongside performance); and
  - *Post-processing approaches*, which are applied after the model has been trained (e.g. adjusting decision thresholds).

  AI developers should carefully assess which of these approaches is/are most appropriate for their specific use case, based on context, data availability, and regulatory requirements.

- **Incorporating Protected Attributes for Fairness Checks:** Paradoxically, one safeguard against bias is the inclusion of protected attributes (e.g. gender, ethnicity) in the modelling process—not for use in decision-making, but to monitor and mitigate potential bias. Most current bias detection and mitigation methods require access to such sensitive information to evaluate and correct disparities.

[21] Ashurst, C., & Weller, A. (2023, October). Fairness without demographic data: A survey of approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-12).

[22] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

- **Regular Audits:** Financial institutions should undergo algorithmic audits to detect bias and ensure compliance with fair lending and consumer protection laws. This requires periodic reviews of AI decisions and their impact on different vulnerable customers.

- **Accountability Mechanisms**: Clear lines of accountability must be established for when AI systems produce harmful or unfair outcomes. For instance, if an AI system unfairly denies credit to a customer in urgent need—such as for medical treatment—there must be clarity on who is responsible: the developers, the institution deploying the model, or the data providers. Regulators, AI developers, financial institutions, and researchers should work together to define robust accountability frameworks that align with the high-risk nature of AI in financial services.

- **Transparency, Explainability, and Consumer Recourse:** AI-driven decisions must not operate as "black boxes." Banks and fintech companies should provide clear, understandable explanations for automated outcomes—especially adverse ones such as loan denials or unfavourable pricing—to enable consumers to understand, contest, and seek recourse when needed.

- **Legislation:** A promising step towards fairer AI is the UK's signing of the Council of Europe's Framework Convention on AI,[23] which establishes equality and non-discrimination as one of its core principles throughout the AI system lifecycle. However, further clear regulatory guidance is needed to define which fairness definitions should be followed in different fintech applications and what levels of bias, if any, are tolerable. Legislation should facilitate effective bias mitigation—such as allowing access to protected characteristics strictly for the purposes of monitoring and mitigation—while ensuring this information is not misused. For example, the EU AI Act permits the use of sensitive data only for bias detection and correction in high-risk AI systems. Additionally, regulations should mandate regular audits, both internal and external, and clarify lines of accountability for biased outcomes in AI systems. Developing these standards should be a collaborative effort involving responsible AI experts (from both academia and industry), domain specialists (such as financial professionals), and regulators.

- **Education and Awareness**: Developers and financial stakeholders involved in designing AI systems must be well-informed about the risks of bias and the importance of responsible AI practices. Companies deploying or purchasing AI-driven tools should ensure these systems have undergone fairness assessments and follow responsible AI principles. Equally important is public education: customers interacting with AI in financial services must understand their rights, including the right to explanation and recourse when decisions—such as loan rejections or pricing—are made automatically.

April 2025

[23] The framework convention on artificial intelligence. Council of Europe. 2024. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence