AIFS0073

Written evidence submitted by the Working Group on fAIr Credit, Credit Research Centre, University of Edinburgh

The Credit Research Centre (CRC), established in 1997, is an impartial research entity dedicated to the study of credit. The Centre specialises in credit risk modelling—particularly within consumer credit—while also addressing broader issues related to lending, borrowing, and financial capability and well-being. The CRC has been fostering collaboration between academic researchers and industry practitioners through a number of initiatives, e.g. Credit Scoring & Credit Control conference, the latest being the Working Group on fAIr Credit.

Credit risk models, particularly credit scoring, represent early and well-established examples of algorithmic decision-making, employing mathematical, statistical, and machine learning (ML) techniques to assess borrowers' creditworthiness. Applications of ML in credit scoring date back to at least 1996 (Desai et al., 1996). The history of credit scoring provides valuable lessons for comparing AI and human decisions, as automated credit scoring models began replacing subjective assessments by credit officers as early as the 1970s.

Besides, the credit industry is shaped by a longstanding and evolving regulatory framework concerned with fairness, transparency, and equality in credit granting. Beginning with the Equal Credit Opportunity Act in the United States (ECOA, 1974), this framework has developed into a complex set of regulations across different jurisdictions. A more comprehensive overview can be found in Andreeva et al. (2004) and Kim et al. (2024a).

The initial assumption of early regulation was that, to ensure fairness and equality, it would be sufficient to exclude prohibited characteristics from training data and decision-making processes, making algorithms 'blind' to prohibited variables such as gender or race. However, this approach did not automatically eliminate historical biases, as will be discussed further. Yet a significant unintended consequence has been that lenders stopped collecting prohibited characteristics altogether due to the fear of being accused of discrimination.

Given this context, the credit industry's experience is highly pertinent to the current inquiry — particularly with respect to the third and fourth sets of questions:

- *What are the benefits and risks to consumers arising from AI, particularly for vulnerable consumers?*
- *How can Government and financial regulators strike the right balance between seizing the opportunities of AI but at the same time protecting consumers and mitigating against any threats to financial stability?*

Benefits and Risks of AI to Consumers, with a Focus on Vulnerable Groups

Machine learning (ML), as a subset of AI, offers significant potential benefits for consumers, including those who are financially vulnerable. Key advantages include:

- **Improved predictive accuracy**, enabled by the capacity to model complex, often non-linear relationships and interactions between variables.
- **Enhanced personalisation**, such as more granular Know Your Customer (KYC) processes and tailored product offerings based on individual needs and circumstances.
- **Early detection of financial distress**, whereby AI/ML can anticipate potential difficulties and enable timely interventions or bespoke solutions for vulnerable consumers.

However, these benefits are accompanied by substantial risks, including—but not limited to—reduced transparency, model instability (e.g. due to overfitting), and **amplified bias**. This submission focuses on the latter, as it represents both a technically and ethically complex issue with far-reaching implications.

The potential for AI/ML to perpetuate and even exacerbate existing biases is well-documented in credit and beyond (Fuster et al., 2022). Historical data reflect real-world inequalities. Consequently, models trained on such data are likely to replicate these disparities. In credit scoring, for instance, if certain demographic groups—such as those defined by ethnicity or gender—historically experienced higher default rates due to systemic disadvantage, models will assign them lower credit scores. Even when protected characteristics, such as race or gender, are excluded (as required by law), associated variables[1] (e.g. income, employment status, residential stability) can act as 'proxies', leading to unequal outcomes. See Andreeva & Matuszyk (2019), Kim et al. (2024b) for examples.

The greater predictive power of AI/ML models reinforces these disparities as compared to traditional statistical methods. Thus, if historically disadvantaged groups—many of whom overlap with vulnerable consumers— showed high default rates, they are likely to face lower acceptance rates and higher borrowing costs under more accurate, but less transparent, AI/ML-driven models.

Conversely, if a group has historically demonstrated stronger credit performance—e.g. women relative to men (Andreeva & Matuszyk, 2019)—ML models may benefit that group by assigning higher credit scores.

Although such situations may be justified from legal and business points of view, they are not necessarily perceived as *fair* by a general public. In the context of AI applications, the following aspects present particular problems. The first one is **lack of agreement regarding what exactly is fair**, and this is not a new problem, which AI has simply brought to the forefront. The second one is need for **justification for decisions** that may be problematic with ML/AI due to its opacity.

Are new regulations needed or do existing regulations need to be modified because of AI?

The societal imperative to address historical inequalities and ensure fairness, particularly for legally protected groups, is clear. Yet achieving fairness is not straightforward. Current regulatory approaches—both in the UK and globally—typically rely on *formal* or *procedural fairness*, i.e. excluding protected characteristics from decision-making models. This reflects the **legal principle of equal treatment**, often referred to as direct discrimination. However, so-called "blind" fairness does not necessarily eliminate structural disadvantage or ensure improved outcomes for marginalised groups.

Public and media concern centres instead on *substantive fairness*, or equality of outcome. High-profile cases (e.g. Apple Card as described by Vigdor, 2019) highlight how public **perceptions of unfair outcomes** persist even in the absence of direct discrimination/ unequal treatment. Apple Card was accused of the sexist treatment because of giving women lower credit limits as compared to men. However, the investigation by the New York State Department of Financial Services did not find any evidence of discrimination (NYSDFS, 2021).

[1] We deliberately used the term 'associated', since association is wider than 'correlation', which is the standard term for an unwanted relationship, however, is problematic, since it is often understood as 'linear dependency'. If a model had a U-shaped dependence on age, it could be quite discriminatory and yet uncorrelated. ML/AI can introduce discriminatory dependencies that are invisible to a linear correlation measure. The definition of fairness in regulation should clarify the term "correlation".

Equality law does address unequal outcomes via the concept of indirect discrimination, where a neutral rule or practice disproportionately disadvantages a particular group. Such practices may be lawful if they pursue a legitimate aim and are proportionate.

Nonetheless, there is an increasing pressure to modify unequal outcomes to reflect societal expectations of fair, unbiased treatment. A number of technical approaches to bias mitigation have been developed in computer science and AI research, such as re-balancing training datasets or optimising models to achieve more equitable outcomes (Kim et al., 2024a). These approaches often require access to protected characteristics—for example, ensuring gender balance among approved applicants requires data on applicants' gender. Another notable solution by Breeden & Leonova (2021) posits that if prohibited characteristics were known, a pre-model could be built that would subtract off all discriminatory bias so that otherwise correlated variables could be used without adjustment or resampling.

This presents a legal and practical dilemma. The use of protected characteristics in technical solutions, such as training set balancing, may be interpreted as violating equal treatment and amount to direct discrimination. Additionally, as mentioned earlier, many lenders do not collect this data, due to legal and ethical concerns.

First, regulatory clarification is needed to confirm that **the use of prohibited characteristics is permissible for fairness monitoring** and algorithmic bias mitigation. Such clarification would facilitate the development and application of methods designed to achieve more balanced outcomes. An alternative remedy could be the creation of a ‘golden sample’ by regulators that includes protected characteristics, enabling lenders to test their models for bias.

Second, it is currently unclear **what constitutes a fair or unbiased outcome**. Multiple definitions of bias exist in the technical literature (Kozodoi et al, 2022), yet none are codified in regulation. Clarifying which fairness metrics and benchmarks are appropriate in specific contexts would assist decision-makers in navigating these complexities. A related issue is the lack of quantification of proportionality in the legal definition of indirect discrimination. Many technical definitions of fairness are mutually incompatible, and it is not possible to satisfy all simultaneously. Despite this, fairness is often referred to in abstract terms, without identifying which specific definitions or principles are being invoked. There is limited understanding of how technical measures of fairness align with public perceptions. Further research, followed by regulatory guidance, is required to bridge this gap. Clear guidelines that integrate legal standards, technical practices, and societal expectations would assist organisations in making fair decisions.

Third, clarification is required on **the use of ‘proxy’ variables**, i.e. variables associated with prohibited ones. As noted above, due to hidden associations models can unintentionally discriminate against vulnerable social groups, even when protected characteristics such as gender or race are excluded by design (Andreeva & Matuszyk (2019), Kim et al. (2024b)). Such indirect or unintentional bias will be amplified by AI models, depending on the data used.

Below are several examples where AI models indirectly infer sensitive attributes through proxies in the data:

• First or last names serving as predictors of gender or ethnicity;

• Postal codes, geo-tags, revealing socioeconomically disadvantaged areas or ethnic enclaves;

• Language features from LLM (e.g., accent, vocabulary complexity) suggesting gender, education level, race, or even health conditions;

• eKYC data, including skin tone and visual appearance, being used—implicitly or explicitly—to estimate gender, income level, or ethnic background;

• Digital footprints being leveraged to deduce disability, gender, age or socioeconomic status.

Conclusion

The credit industry's long experience with algorithmic decision-making provides valuable insights for understanding the role of AI in financial services. While AI/ML can enhance predictive power and enable personalised solutions, it can also amplify historical biases. Current regulatory approaches focus on procedural fairness but fall short of addressing structural inequalities or aligning with public expectations. Greater clarity is needed on the use of protected characteristics for fairness testing, the selection of appropriate fairness metrics, and the integration of legal, technical, and societal perspectives. Regulatory support for representative datasets and guidance on best practices would help institutions ensure that AI contributes to more equitable financial outcomes.

On behalf of the Working Group on fAIr Credit, Credit Research Centre

Prof Galina Andreeva,
Personal Chair in Societal Aspects of Credit,
Director, Credit Research Centre
University of Edinburgh Business School

Dr Joseph Breeden,
Chief Executive Officer at Deep Future Analytics LLC

April 2025

References

Andreeva, G., Crook, J., & Ansell, J. (2004). Impact of anti-discrimination laws on credit scoring. Journal of Financial Services Marketing, 9(1), 22–33. https://doi.org/10.1057/palgrave.fsm.4770138

Andreeva, G., & Matuszyk, A. (2019). The law of equal opportunities or unintended consequences: The impact of unisex risk assessment in consumer credit. Journal of Royal Statistical Society, Series A, 182(4), 1287–1311. https://doi.org/10.1111/rssa.12494

Breeden, J., & Leonova, E. (2021). Creating Unbiased Machine Learning Models by Design. Journal of Risk and Financial Management.

Desai, V.S., Crook, J.N., & Overstreet G.A.J. (1996). A comparison of neural networks and linear scoring models in the credit union environment, European Journal of Operational Research, 95(1), 24 – 37.

ECOA (1974). Equal Credit Opportunity Act. https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/JOFI.13090

Kim, S., Andreeva, G., & Rovatsos, M. (2023a). The 40-year journey for fairness in credit: A systematic review and future research directions [SSRN Working Paper 4669379]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4669379

Kim, S., Andreeva, G., & Rovatsos, M. (2023b). The double-edged sword of Big Data and Information Technology for the disadvantaged: A cautionary tale from Open Banking [CRC Working Paper 01/2023]. https://www.crc.business-school.ed.ac.uk/research/working-papers

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083–1094. https://doi.org/10.1016/J.EJOR.2021.06.023

NYSDFS (2021). Report on Apple Card investigation. https://s3.documentcloud.org/documents/20521283/rpt_202103_apple_card_investigation.pdf

Vigdor, N. (2019, November). Apple Card Investigated After Gender Discrimination Complaints . The New York Times. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html