Bias in Algorithmic Decision making in Financial Services
Barclays Response

Barclays is a transatlantic consumer and wholesale bank with global reach, offering products and services across personal, corporate and investment banking, credit cards and wealth management, with a strong presence in our two home markets of the UK and the US. With over 325 years of history and expertise in banking, Barclays operates in over 40 countries and employs approximately 85,000 people. Barclays moves, lends, invests and protects money for customers and clients worldwide.

Barclays welcomes the opportunity to engage with this consultation from the Centre for Data Ethics and Innovation. We provide our high level perspectives in response to the questions posed. We hope you find our views useful and we would be happy to discuss them further if helpful.

1. The Use of Algorithmic Tools

- *What algorithmic tools are currently being developed or in use?*
- *Who is developing these tools?*
- *Who is selling these tools?*
- *How are these tools currently being used? How might they be used in the future?*
- *What does best practice look like and are there examples of especially innovative approaches?*
- *What are the key ethical concerns with the increased use of algorithms in making decisions about people?*

Within Barclays, we utilise multiple algorithmic methods to develop and enhance our customer relationships and services. For example:

1. **Marketing of Products & Services:** We apply algorithms to customer data to predict evolving customer needs and to identify the services or products we provide that fulfil those needs.

2. **Credit Risk Assessment:** We apply algorithms to predict a customer's risk of default using Credit Scoring methods, and to assess a customer's ability to afford repayments when they apply for credit or a borrowing product. This ensures that customers only borrow to the extent that they can repay, are able to access credit, and the price of that credit is competitive.

3. **Fraud Identification**: We apply algorithms to identify fraudulent transactions with pattern recognition to identify variances against customers' spending profiles, protecting them against transaction fraud.

These algorithms are developed externally (with providers like SAS, Experian, FICO or Continuum (Anaconda) or other Open Source Packages) and are applied to different areas within Barclays, using internally or externally sourced data.

The algorithms used in decision-making are advancing at great pace, commensurate with increases in computing power and availability of data. This allows capture of complex non-linear relationships among data elements to minimise prediction errors. However, the underlying approach on training & validation of these models is relatively consistent and focuses on evaluating stability, robustness and objectivity of decision outcomes on independent data samples.

Increasing use of algorithms across an increasingly digital economy poses various ethical considerations, for example on transparency, fairness, and accuracy and explainability of decision-making systems. However, it is important to point out that these considerations are not specific to automated decision making only, but to the wider trade-offs resulting from the application of technology to business problems.

Privacy & Transparency: The use of algorithms requires data. In many cases the greater the amount of data available, the greater the scope for accuracy in the algorithm. There is therefore an ethical consideration of trade-off between privacy and accuracy. Organisations will need to consider the extent to which sourcing more data to feed an algorithm adds enough value in terms of increased accuracy to make it worthwhile. Transparency in the communication of what data is captured, and for what purposes, is critical to ensuring consumers understand what benefit they are receiving in return, so they can make an informed decision on the value of that exchange to them.

Fairness & Bias: A second ethical concern that can arise from algorithms in automated decision-making is the extent to which unintended consequences of biases in the design and development of the algorithm or use of data can have a detrimental impact on individuals and society. In some cases, unintended consequences can become visible only once the algorithm has been in operation for a period of time. There is an ethical consideration in deciding the extent to which it is fair to release an algorithm without a full understanding of its potential impact. As mentioned above, this is not a trade-off unique to Machine Learning algorithms. However, as the use of automated decision making systems increases, it is important to apply diversity in thinking and approach, as well as close monitoring of the Machine Learning systems in order to identify any unintended consequences early and have the ability to stop it immediately.

Accuracy & Explainability: A further ethical consideration is the choice of algorithms available, and the varying levels of accuracy & explainability they provide. For example, better predictive models powered by Machine Learning algorithms are more accurate in predicting defaults, enabling banks to provide credit to customers they may not have been able to extend credit to using more primitive algorithms. At the same time, these algorithms also operate as a Blackbox, meaning their selection techniques are hidden from users. Hence, the use of algorithms for automated decision-making will create a trade-off between explainability and accuracy. Data scientists and organisations will need to strike the balance between the accuracy of the prediction and the ability to demonstrate to individuals affected how a decision was made.

The concept of contestability quickly becomes a relevant principle to consider. The concept does not require a full explanation of how the algorithm works in detail for specific individual decisions, rather it allows the organisation to give a reasonable explanation of how the outcome was derived if individuals were to contest the outcome.

2. Bias Identification and Mitigation

- *To what extent (either currently or in the future) do we know whether algorithmic decision making is subject to bias?*
- *At what point is the process at highest risk of introducing bias? For example, in the data used to train the algorithm, the design of the algorithm, or the way a human responds to the algorithm's output.*
- *Assuming this bias is occurring or at risk of occurring in the future, what is being done to mitigate it? And who should be leading efforts to do this?*
- *What tools do organisations need to help them identify and mitigate bias in their algorithms? Do organisations have access to these tools now?*
- *What examples are there of best practice in identifying and mitigating bias in algorithmic decision making?*
- *What examples are there of algorithms being used to challenge biases within existing systems?*

Decision-making in general can be biased, though any bias can be partially mitigated through values and standards. In the case of algorithms, any impacts from bias may further be amplified due to the scalable nature and potential for unintended consequences. We identify potential bias risks below:

Design & Interpretation Bias: Primary risk of bias comes in the design of algorithms and in the interpretation and use of their output by humans.

This could be mitigated by establishing ethics principles and aligning process design to organisational values and holding them to the same ethical standards as human-driven decisions. This can also include system designs with the ability to monitor and control bias to ensure no individual or group of individuals are systematically disadvantaged, unless these decisions can be justified.

A lack of diversity in the teams involved in preparing and assessing the data set, designing the algorithm, and testing any outcomes, can also result in unintended bias. Diversity in thinking, cultural backgrounds, as well as gender, ethnicity and professional specialism can help minimise the risk of unintended bias. It is increasingly important to have multidisciplinary teams involved in the end-to-end design, development and implementation of algorithms.

Information Bias: The use of inherently biased data in training algorithms is a further potential risk of unintended bias. If data inputs have bias, the model will learn those biases and further amplify them ending up in self-fulfilling prophecies. This risk grows significantly with use of Machine Learning involving non-linear functions applied on large scale of data.

For example, to develop a new credit scoring model which predicts credit defaults from applications for a credit product, banks will have to use default data from customers who were previously accepted for credit. However, if the data on previously accepted customers were biased for any reason and not representative of all applicants, the model will develop predictors from the previously accepted population (biased sample) and would assume them as predictors for all applicants, resulting in new applicants potentially being declined credit based on predictors not suitable for them. To correct this bias risk, banks would have to collect unbiased data by accepting all applications, which may result in significant business cost and customer detriment making this a difficult trade-off.

In Barclays case,

1. We have a strong culture of running many tests at small scale to collect unbiased data in a systematic and incremental way. We believe this allows us to minimise the risk of collecting data and to incrementally remove any potential bias in the training of the algorithms.

2. We also continuously look to challenge and augment data sources to include a broad set of characteristics to get a full picture of consumer behaviour and reduce reliance on a small number of characteristics susceptible to bias.

3. Also, while introducing new algorithms, we are cautious with validations on independent samples and testing at small scale in parallel runs to validate our algorithms' effectiveness in achieving the intended outcome.

4. Decisions taken on the basis of algorithms are also continually monitored for their performance including customer response or reactions providing a feedback loop to identify any potential bias in data or design of algorithms.

With regards to methods for identification of bias of protected attributes (Gender, Ethnic Groups, Race, Age etc) in decision making – an example approach would be to be 'Bias Aware', as below:

1. **Training vs Validation Usage:** Protected Attributes should not be used in the training of algorithms for use cases where using them may cause discrimination in breach of regulation. However, these attributes can be used to identify biased outputs through validation. To do this, organisations should be able to collect information on protected attributes.

2. **Identification of Proxies:** These attributes should be analysed for potential proxies in the data (e.g. Ethnic group to post code correlations) to identify key risk variables for bias mitigation.

3. **Validation of Prediction Accuracy:** Prediction outputs of algorithms should be validated for accuracy within populations defined by protected attributes to establish that predictions hold true across population segments. For example, a probability of default of X% should result in X% default rates when assessed for any particular ethnic groups with statistically significant populations.

4. **Algorithmic Independence Assessment**: If a protected attribute (e.g. ethnic groups) on its own differentiates a performance outcome (e.g. credit default) significantly, its inclusion in training of a particular algorithm should add significant prediction accuracy. We should be able to contrast the performance of the algorithm when a protected attributes is excluded vs included. Where an algorithm is found to **not** achieve an increase in prediction accuracy by including protected attributes with significant predictive power, it is essentially triangulating for the protected attribute with proxies from remaining available data and hence, is potentially biased. This would need mitigation by reducing influence of key risk variables

(identified in step 2) at the cost of predictive power to achieve equality of opportunity i.e. fairness.

3. Public Engagement

- *What are the best ways to engage with the public and gain their buy in before deploying the use of algorithms in decision making? For example, should a loan applicant be told that an algorithm is being used to assess their loan application?*
- *What are the challenges with engaging with the public on these issues?*
- *What are good examples of meaningful public engagement on these issues?*

It is important to be transparent (where relevant) with customers regarding decision making with algorithms, and the GDPR provides a comprehensive framework for meeting this challenge, accompanied by ICO guidance. For instance, under GDPR Article 22, where a firm makes a 'fully automated decision' that has a 'significant effect' on an individual, it must put in place safeguards to protect the individual, notably including the right to obtain human intervention, to express his/her point of view and to contest the decision. The ICO's guidance also notes that the individual has a right to an explanation of the decision.

The level of understanding about algorithms will range amongst the public. However, the average customer will need to be educated in this space. It is therefore important to engage all stakeholders to raise awareness on the nature and benefits of algorithms in helping individuals and organisations make better decisions. For customers, and broader society, any awareness raising efforts should highlight what they can do to influence algorithmic decisions to achieve a positive outcome (for example – what are the ways to build good credit history to be able to access credit at the best price in future). For regulators, any engagement should seek to demonstrate the robustness and stability of the algorithmic decisions to build trust.

However, any public engagement needs to take into account various challenges:

- Customers may have limited understanding of decision making algorithms. Therefore, it is necessary to explain automated decision-making in a manner that is both understandable by a non-specialist but also is sufficiently detailed.
- Customers may want to control the use of their data for decision-making but may also suffer interest fatigue due to the frequent notifications of data processing (e.g. cookie data).
- Machine Learning algorithms have trade-offs of interpretability vs accuracy. Any public engagement would need to balance this trade-off and the associated benefits and costs.
- Customers may have different levels of scepticism of algorithms and reactions to decisions driven by algorithms (especially in case of adverse decisions).
- Algorithms, by definition, are statistical & impersonal. Hence, any public engagement needs to include education around human involvement in the development and use of algorithms, especially with the heightened customer sensitivity around automation.

With respects to good examples of meaningful public engagement the GDPR sets out a 'layered approach' which is recommended by both the ICO and EDPB. This involves key information being provided up front, with the option for customers to access further detail separately. For example, with respect to credit checks, the industry, in consultation with the ICO, developed an approach to facilitate GDPR compliance by lenders and credit reference agencies in a manner that is digestible to customers and focused on their key information requirements. As such the lenders provide a short notice that provides an explanation of how credit scoring works and what data is shared. The customer can then follow this through to a more detailed explanation of how credit reference agencies operate and generate credit scores, and can also explore any additional information about the lending decisions of the particular lender.

4. Regulation and Governance

- *What are the gaps in regulation of the use of algorithms?*
- *Are there particular improvements needed to existing regulatory arrangements to address the risk of unfair discrimination as a result of decisions being made by algorithms?*

The GDPR provides information on the use of algorithmic decision-making in financial services. In order to help firms meet the obligations under the GDPR both the ICO and the European Data Protection Board have provided guidance. Key considerations for firms are the identification and mitigation of (unfair) bias, and transparency and governance.

The FCA Handbook sets out extensive regulatory obligations for financial services firms and provides overarching high-level principles requiring firms to treat customers fairly, maintain appropriate systems, and maintain appropriate governance.

The question of whether there are gaps is a difficult one. On one hand the GDPR's requirements for 'fairness', 'transparency' and a legitimate basis for processing essentially cover the key risks to individuals, and in the case of lending there is also the FCA regime. However, regulations tend to be technology neutral (which is understandable to avoid regulations no longer being applicable when new technologies are utilised) however at the same time it is essential that regulations are written with technologies and more specifically 'use cases' in mind. As more complex use cases evolve, it will be important that the regulators collaborate and cooperate to provide further guidance in lieu of further regulation being provided.

With regards to improvements needed to existing regulatory arrangements, one area of focus could be to set a level playing field by standardising access to data across industries, and explicitly confirming any interpretability requirements of algorithms into key decisions. This would ensure that industry works with the same constraints on their choice of data and algorithms to arrive at decisions.

Also, as noted earlier, to better identify and mitigate any 'unfair' bias in automated decision-making, it should be possible to explore the possibility of collecting and processing data on customers' 'protected characteristics'. However, it is important to consider the following:

- Article 9 of the GDPR places strict controls on the processing of ‘special category data’ (SCPD). These are: “data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation…”

Whilst Article 9 does not include age or sex, there is overlap with the ‘protected characteristics’ under the Equality Act. Firms may also have incomplete datasets as under GDPR SCPD cannot be processed unless a limited set of exemptions apply. Increased collection of data entails various risks (including security incidents and data breaches) and as such, we would be cautious with regards to a regulatory expectation that firms should collect such data in order to mitigate bias in algorithms.