IBM response to CMA call for information on Algorithms

IBM is a global provider of technology products and services and has played a leading role in the development and delivery of Artificial Intelligence. In an IDC report published in July 2020, IBM was ranked as the largest global provider of AI software platforms for the fifth successive year for 2019.

As outlined in our Principles for Trust and Transparency, IBM has long argued that AI systems need to be transparent and explainable. We have contributed to and support the OECD AI Principles, and in particular the need to "commit to transparency and responsible disclosure" in the use of AI systems. We have also contributed to the European Commission High Level Expert Group on AI and to its Assessment List for Trustworthy AI (ALTAI) which we believe will be a valuable tool to help guide organisations during the process of designing and building AI in a responsible way.

Our comments are aimed primarily at question 5 relating to Section 3 of the paper.

Whilst principles are admirable and can help communicate a company's commitments to citizens and consumers, we believe that requiring appropriate disclosure, based on use-case and end-user, should be the default expectation for many companies creating, distributing, or commercialising AI systems. We support targeted policies that would increase the responsibilities for companies to develop and operate trustworthy AI. Given the ubiquity of AI, there will be no one-size-fits-all rules that can properly accommodate the many unique characteristics of every industry making use of this technology and its impact on individuals but we can define an appropriate risk-based AI governance policy framework based on three pillars:

1. ***Accountability*** *proportionate to the risk profile of the application and the role of the entity providing, developing, or operating an AI system to control and mitigate unintended or harmful outcomes for consumers.*
2. ***Transparency*** *in where the technology is deployed, how it is used, and why it provides certain determinations.*
3. ***Fairness and security*** *validated by testing for bias before AI is deployed and re-tested as appropriate throughout its use, especially in automated determinations and high-risk applications.*

In order to help organisations develop and deploy AI systems IBM has continued to develop a range of open tools which are available to all. We offer these as examples of tools which may be useful to the CMA as work develops in this area.

1. **Accountability**

**IBM Watson Openscale**

Watson OpenScale provides an open platform which allows businesses to operate and automate AI at scale - irrespective of how the AI was built and where it runs. Available via the IBM Cloud and IBM Cloud Private, it infuses AI with trust and transparency, explains outcomes, and automatically eliminates bias.

Tools include an explainability feature that gives detailed insights into the structures of AI models and the results they provide, giving users the power to back up their decisions. If there is an issue with the quality of a model, such as a potential bias, IBM Watson OpenScale is able to provide alerts so teams can quickly respond. Users can also trace the creation of their models to the data that the information was sourced from. Visualisations allow stakeholders to understand the source of decision making that the models follow, as well as quality issues, and drill into the specific model transactions.

2. **Transparency**

**AI Fact Sheets**

IBM has made available a first-of-its-kind-methodology for how to assemble documentation or "fact sheets" about an AI model's important features, such as its purpose, performance, datasets, characteristics, and more. These can be accessed via an AI FactSheets 360 website. With each step of the methodology, we describe the issues to consider and the questions to explore with the relevant people involved in creating and consuming the facts that go into an AI FactSheet. The website also shares an approach to AI Lifecycle Governance and a collection of example FactSheets, research papers, and other resources for anyone to use.

The concept of an AI FactSheet is very flexible. Different users will need different types of information. Likewise, different AI applications or use cases will implicate different information needs. Also, an AI FactSheet is not meant to explain every technical detail or disclose proprietary information about an algorithm. Rather, the goal is to promote human decision-making in the use, development, and deployment of AI systems, while also accelerating developers' education on AI ethics and their broader adoption of the concepts of transparency and documentation.

**Explainability 360**

AI Explainability 360 is a comprehensive open source toolkit of state-of-the-art algorithms that support the interpretability and explainability of machine learning models. We believe this helps develop the theory and practice of responsible and trustworthy AI. A doctor diagnosing a patient may benefit from seeing cases that are very similar or very different. An applicant whose loan was denied will want to understand the main reasons for the rejection and what she can do to reverse the decision. A developer may want to understand where the model is more or less confident as a means of improving its performance.

As a result, when it comes to explaining decisions made by algorithms, there is no single approach that works best. There are many ways to explain. The appropriate choice depends on the persona of the consumer and the requirements of the machine learning pipeline. AI Explainability 360 deals with this diversity of explanation with algorithms for case-based reasoning, directly interpretable rules, post hoc local explanations, post hoc global explanations, and more.

3. **Fairness and Security**

**Fairness 360**

AI Fairness 360 is a comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias.

Machine learning models are increasingly used to inform high-stakes decisions about people. Although machine learning, by its very nature, is always a form of statistical discrimination, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage. Bias in training data, due to either prejudice in labels or under-/over-sampling, yields models with unwanted bias.

The AIF360 Python package contains ten different bias mitigation algorithms, developed by the broader algorithmic fairness research community, to mitigate that unwanted bias. They can all be called in a standard way, very similar to scikit-learn's fit/predict paradigm. In this way, we hope that the package is not only a way to bring researchers together, but also a way to translate our collective research results to data scientists, data engineers, and developers deploying solutions in a variety of industries. AIF360 is a bit different from currently available open-source efforts due its focus on bias mitigation (as opposed to simply on metrics), its focus on industrial usability, and its software engineering. AIF360 contains two tutorials on credit scoring and on predicting medical expenditures.

**Adversarial Robustness 360**

The Adversarial Robustness 360 Toolbox is an open source software library which supports both researchers and developers in defending deep neural networks against adversarial attacks, making AI systems more secure. Its purpose is to allow rapid crafting and analysis of attack and defence methods for machine learning models.

The Adversarial Robustness 360 Toolbox provides an implementation for many state-of-the-art methods for attacking and defending classifiers. It is designed to support researchers and AI developers in creating novel defence techniques and in deploying practical defences of real-world AI systems. For AI developers, the library provides interfaces that support the composition of comprehensive defence systems using individual methods as building blocks.