Data Governance Network
Anchored by IDFC Institute

Algorithmic Explainability Working Group: Green Paper 001/Discussion Note

Key Takeaways

- **Standard-setting for algorithms must be contextual and specific**. Fairness, accountability, trustworthy and explainability mean different things in different use case scenarios and for different stakeholders.
- **Explainability may be at odds with efficiency and performance**. The more robust and sophisticated an algorithm, the less explainable its decision-making. This raises the question of whether full explainability is necessary for achieving FATE design's objectives.
- **In certain contexts, outcome mapping could serve as the proxy for explainability**. Assessing outcomes could help understand which tool to deploy, and determine if bias exists in the system.
- **FATE standards thus must be narrowly drafted, keeping not just sectoral but use case variance in mind**. Given how contextual each element of FATE is, operationalising a broad standard is impossible.

Intent

Automated Decision Systems (ADS) have proliferated in recent years, in India as elsewhere, enabled by a rise in big data. They are increasingly crucial components of consumer-market and citizen-state interaction. Thinking through effective, holistic regulatory guidelines that can inform both self- and government regulatory frameworks is therefore critical. A rich body of literature and experience shows the potential downsides of poorly implemented ADS in the absence of such frameworks.

The **Working Group on FATE (Fair, Accountable, Trustworthy and Explainable) Standards for ADS in India**, anchored by IDFC Institute's Data Governance Network and CPC Analytics, aims to develop such frameworks for specific use cases. This discussion paper summarises the key takeaways from its first session with a diverse group of academic, industry and policy experts. These takeaways will inform the Working Group's scope of work.

Explaining Explainability

"Black box" ADS driven by algorithmic processes have caused justifiable regulatory and public concern. However, there is no clear definitional framework for algorithmic harms. Some have suggested a human rights-based approach[1], while others have argued for considering the ethical implications of different types of algorithms[2] —

[1] McGregor, L., Murray, D., & Ng, V. (2019). International human rights law as a framework for algorithmic accountability. *International and Comparative Law Quarterly*, *68*(2), 309–343. doi:https://doi.org/10.1017/S0020589319000046

[2] Kerr, I., & Earle, J. (2013). Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy. *Stanford Law Review Online*, *66*(65), 65–72. Retrieved from

consequential (anticipating the likely consequences of a person's actions and helping with risk management), preferential (outcomes based on user preferences) and pre-emptive (takes decisions on behalf of users, restricting their alternatives and knowledge about alternatives).

Further complicating matters, bias and lack of fairness in algorithmic processes can rarely be assessed in isolation. They are more often marginal and counterfactual, assessed against the bias in the human decision-making the algorithmic process has replaced[3].

Given these variables, there is, unsurprisingly, no consensus on what explainability means — or if it is even necessary for the fairness and accountability that supposedly lie downstream of it. The GDPR has taken a broad view of the issue. Recital 71 looks at input processes — integrity of data collection to ensure accuracy and lack of bias — rather than unpacking algorithms themselves. This is important: algorithms that use simpler processes like linear regression can be unpacked, but more complex models are orders of magnitude more difficult[4].

The next section summarises the Working Group's recommendations keeping the highly contextual nature of FATE implementation in mind.

Key Takeaways

1. Need for specificity while setting algorithmic standards

The context and specificity of where the algorithm is deployed will be crucial. FATE standards are impossible to define in the abstract and the parameters, especially of fairness, will have to be defined on a case-by-case basis. This is partly because the principles of fairness may differ contextually and hence the standards across domains cannot have a one-size-fits-all approach[5]. For instance, in online marketplaces like a job market or credit market, protecting users from discrimination will need to be emphasised, while for search engine results, preventing filter bubbles or abusive/malicious content will be important.

It is also because fairness, accountability and trustworthy will mean different things to different stakeholders: data owners, algorithm users, developers and regulators. Developers will need to practice different kinds of accountability for regulators and users, for instance — for example, providing ethics training to employees,

https://review.law.stanford.edu/wp-content/uploads/sites/3/2016/08/66_StanLRevOnline_65_KerrEarle.pdf

[3] Cowgill, C. & Tucker, C. (2017). Algorithmic Bias: A Counterfactual Perspective. *NSF Trustworthy Algorithms Working Paper*. Retrieved from https://pdfs.semanticscholar.org/55ce/d34a39ed52ddcc7435b7637d5fda55210eed.pdf

[4] Hume, K. (2018). When Is It Important for an Algorithm to Explain Itself? *Harvard Business Review*. https://hbr.org/2018/07/when-is-it-important-for-an-algorithm-to-explain-itself

[5] Seng Ah Lee, M. (2019). Context-conscious fairness in using machine learning to make decisions. *AI Matters, 5*(2), 23–29. Retrieved from https://doi.org/10.1145/3340470.3340477

participating in third-party audits or setting up channels of consumer recourse. And as far as transparency and explainability go, the demand for both is likely to be much higher in the context of investment and credit decisions than it is in health diagnosis.

2. Explainability may be at odds with efficiency and performance
The more robust and sophisticated an algorithm, the less explainable its decision-making is even to experts in its field[6]. To illustrate, AI models with artificial neural networks may have significant capacity to interpret useful patterns for an insurance company, but the model explainability could be difficult even for people with adequate knowledge of machine learning[7]. This has profound implications in terms of not just understanding decision-making but also addressing AI-related harms [8].

This also leads to two important questions that will have to be addressed: Is full explainability necessary for achieving FATE design's broad objectives? And by looking at regulation in other jurisdictions and precedent from other sectors, is it possible to build a taxonomy for when explainability is important and when it is irrelevant to the purpose?

3. In certain contexts, outcome mapping could serve as the proxy for explainability
When complete explainability is not possible, outcome mapping of AI decisions can be a useful tool. Assessing outcomes can help understand which tool to deploy, and determine if bias exists in the system. Experts such as Geoff Hinton, for instance, have suggested that technology be regulated based on its performance. However, using outcome mapping as a proxy in a more nuanced manner is advisable given that in some cases, it might not shed light on the counterfactual. For example, it would be hard to decipher the counterfactual outcome of using an alternative algorithmic model in financial investing.

4. FATE standards thus must be narrowly drafted, keeping not just sectoral but use case variance in mind
Given how contextual each element of FATE is, operationalising a broad standard is impossible. While it is important to have agreement on broad data governance principles, it is equally important for legal and administrative norms to be developed by specific sectors that deal with these issues. This is echoed in the Ministry of Electronics and Information Technology's Draft Report on Cyber Security, Safety, Legal and Ethical Issues[9] which recommends that a future AI framework "should

[6] London, A. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15-21. doi: 10.1002/hast.973
[7] Bornstein, A, Is Artificial Intelligence Permanently Inscrutable?, Nautilus, 2016
[8] Ronan, H., Henrik, J., & Ignacio, S. (2020). Robustness and Explainability of Artificial Intelligence. Retrieved from https://publications.jrc.ec.europa.eu/repository/bitstream/JRC119336/dpad_report.pdf
[9] Committee — D, Ministry of Electronics & Information Technology. GoI. (2019). Report of Committee — D on Cyber Security, Safety, Legal and Ethical Issues. Retrieved from https://meity.gov.in/writereaddata/files/Committes_D-Cyber-n-Legal-and-Ethical.pdf.

define the broad principles and guidelines/requirements and allow organizations to design their own programs in compliance with these principles, with flexibility to adapt as the technology continues to evolve at a rapid pace … without introducing excessive bureaucracy”.