immerse
EDUCATION

Breaking the Feedback Loop: Ethical Interruptions for Just Algorithms

(Discuss the concept of machine learning fairness and the measures needed to prevent AI bias)

Majd Al Bitar

Algorithms increasingly mediate decisions once entrusted to human judgment, yet their authority often conceals a disquieting truth: they can quietly encode and magnify structural injustice. In 2020, Robert Williams, a Black man from Michigan, was wrongfully detained after a flawed facial recognition system misidentified him as a suspect. Held for 30 hours, Williams became what Garvie (2022) characterises as a casualty of a self-reinforcing feedback loop, wherein flawed precedents were reified as facts and shielded from scrutiny. His ordeal exposes a deeper crisis in algorithmic decision-making, where discrimination and opacity are not aberrations but embedded design features. Although often presented as neutral, machine learning systems inevitably inherit the inequities embedded within their training data. Barocas and Selbst (2016) demonstrate that when outputs are recycled as inputs, a process known as feedback looping, early errors compound and become institutionalised. Conventional audits attempt to address such failures only after harm has materialised. What is needed instead are anticipatory safeguards that intervene at the point of emergence, embedding ethical interruptions as deliberate mechanisms to surface, suspend, and correct systemic bias before it crystallises into policy.

Grasping why such interventions are indispensable requires recognising how bias not only persists but escalates within algorithmic systems. Lum and Isaac (2016) reveal that predictive policing models trained on historical arrest records disproportionately target racialised neighbourhoods, prompting heavier surveillance, more arrests, and reinforcing their initial assumptions. Likewise, Obermeyer et al. (2019) found that a healthcare algorithm underestimated Black patients' risk because it used past expenditure as a proxy for need, misreading lower costs as lower risk. Eubanks (2018, cited in Gordon 2019) argues that such systems routinely profile and penalise the poor under the guise of efficiency, while Benjamin (2019) warns that they embed racial hierarchies directly into computational logic. These are not incidental glitches; they are engineered feedback patterns that perpetuate structural harm.

Opacity intensifies this harm, enabling bias to operate unchecked behind the veneer of technical objectivity. As Pasquale (2015) observes, many deep learning models function as black boxes, protected by trade secrets and inaccessible even when shaping credit or employment outcomes. Noble (2018) further contends that search engines not only echo societal biases but actively entrench them. Efforts to counteract this through transparency reports and audits remain piecemeal, retrospective, and easily manipulated, offering little insight into real-time behaviour. Ethical interruptions would pierce this opacity from within, compelling hidden mechanisms into view before their consequences unfold.

Drawing on the principles of chaos engineering, Basiri et al. (2016) advocate embedding controlled glitches that pause algorithms at ethically sensitive junctures. A sentencing tool, for example, could trigger an interruption if racial correlations exceed a defined threshold, surfacing its decision variables for human review. This becomes especially

critical in neuromorphic computing, where brain-inspired chips continuously reconfigure themselves through real-time feedback. Bias could propagate rapidly within such adaptive architectures, and interruptions would function as circuit breakers, forcing systems to disclose their internal logic before distortions ossify.

Although critics fear potential delays, most algorithmic judgments, such as welfare eligibility or predictive policing, are not inherently time-critical. Benjamin (2019) proposes establishing Interruption Panels composed of engineers, ethicists, legal scholars, social scientists, and affected communities to evaluate flagged cases. Ethical interruptions would reframe algorithmic outputs as provisional hypotheses rather than immutable truths. By embedding friction, reflection, and accountability within systems designed for frictionless optimisation, they counter the logic of surveillance capitalism that Zuboff (2019) warns against. If algorithms are to shape the future, they must first be taught to pause, and in that pause, to reckon with the humanity they aspire to model.

References:

- Barocas, S. and Selbst, A.D. (2016) ‘Big data’s disparate impact’, California Law Review, 104(3), pp. 671–732.
- Basiri, A., Behnam, N., de Rooij, R., Hochstein, L., Kosewski, L., Reynolds, J. and Rosenthal, C. (2016) ‘Chaos engineering’, IEEE Software, 33(3), pp. 35–41.
- Benjamin, R. (2019) Race after technology: Abolitionist tools for the New Jim Code. Cambridge: Polity.
- Garvie, C. (2022) A forensic without the science: Face recognition in U.S. criminal investigations. Washington, DC: Center on Privacy & Technology at Georgetown Law.
- Gordon, F. (2019) ‘Virginia Eubanks (2018) Automating inequality: How high-tech tools profile, police, and punish the poor. New York: Picador, St Martin’s Press’, Law, Technology and Humans, 1(0), pp. 162–164.
- Lum, K. and Isaac, W. (2016) ‘To predict and serve?’, Significance, 13(5), pp. 14–19.
- Noble, S.U. (2018) Algorithms of oppression: How search engines reinforce racism. New York: NYU Press.
- Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) ‘Dissecting racial bias in an algorithm used to manage the health of populations’, Science, 366(6464), pp. 447–453.
- Pasquale, F. (2015) The black box society: The secret algorithms that control money and information. Cambridge, MA: Harvard University Press.
- Zuboff, S. (2019) The age of surveillance capitalism: The fight for a human future at the new frontier of power. New York: PublicAffairs (Hachette Book Group).