When Algorithms Fail: The Case for Moral Repair (Extended Abstract)

Pak-Hang Wong[1], Gernot Rieder[2]

[1]Hong Kong Baptist University
[2] University of Bergen

pakhangwong@hkbu.edu.hk, gernot.rieder@uib.no

As concerns over the societal impacts of AI and algorithmic decision-making intensify, most scholarly and regulatory efforts have focused on identifying risks and implementing safeguards to prevent harm. These approaches are grounded in the belief that algorithmic harm can be avoided through well-defined rules and ethical design. However, this prevention-first mindset has left a critical gap: the lack of attention to post-harm scenarios, i.e. cases and situations where individuals have already been harmed by an algorithmic system. Following Charles Perrow's (1984) argument that failure is a normal aspect of complex systems, this paper expands on the concept of the *algorithmic imprint* – first introduced by Ehsan and colleagues (2022) – to explore the enduring psychological and social effects of algorithmic harm. Revisiting the Ofqual grading controversy in Bangladesh and highlighting its lasting effects on both students and teachers, we argue that neither decommissioning harmful systems nor reversing flawed decisions is sufficient to fully address the resulting harm. Instead, we advocate for a more robust response rooted in the principles of moral repair, outlining what such a process should entail and why it is essential for restoring trust and justice in the wake of algorithmic failure. More specifically, drawing Margaret Walker's (2006) pioneering work, we argue that in the context of algorithmic harm, moral repair

1. involves identifying and holding to account those responsible for harm, which may include developers, deployers, and those who have authorized the use of the algorithmic system;
2. requires wrongdoers and those who share responsibility to acknowledge and address the harm done to individuals, which may include not only decommissioning, reversal of decisions, and compensation, but also a *range of moral gestures* such as (public) admissions and apologies, a commitment to making amends and seeking reconciliation, and assurances of non-recurrence, all aimed at both validating and relieving the potential suffering, anger, and alienation of victims of wrong;
3. considers the state of moral norms and standards within a community, their absence or insufficiency when a system causes harm, and the need to reaffirm their relevance and value to those affected as well as to society at large;
4. focuses on restoring confidence among individuals that shared moral standards will be respected in the design and implementation of algorithmic systems and that efforts to encourage and enforce moral conduct will be supported;
5. calls for rebuilding reasonable hope that moral understandings and those responsible for upholding them are worthy of trust, providing assurance that principles of fairness and justice will be maintained;
6. aims to reconnect on both practical and moral grounds those who have done wrong and those who have been harmed, working toward a *climate of mutual respect* that in cases of serious injury may be difficult to establish.

Of course, the moral repair of algorithmic harm is far from straightforward, as the complex sociotechnical nature of algorithmic systems – with their inherent opacity and unpredictability (see, e.g., Santoni de Sio and Mecacci, 2021) – poses difficult theoretical and practical challenges to formulating satisfactory responses. Yet it is a critical issue that researchers and policymakers must confront. Only through deeper reflection and sustained engagement can we begin to craft more realistic governance frameworks that acknowledge the inevitability of residual risk, dare to look beyond preventative measures, and embrace responsibility for the full spectrum of algorithmic harm.

The full paper is forthcoming in *Science and Engineering Ethics*, titled "After Harm: A Plea for Moral Repair after Algorithms Have Failed". https://doi.org/10.1007/s11948-025-00555-y

Acknowledgements

GR was supported by the Research Council of Norway under the grant number 315580 "CoPol: COVID-19 contact tracing as Digital Politics" for this work.

References

Ehsan U, Singh R, Metcalf J and Riedl M (2022) The Algorithmic Imprint. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, USA, 1305–1317. https://doi.org/10.48550/arXiv.2206.03275

Perrow C (1984) *Normal Accidents*. New Jersey: Princeton University Press.

Santoni de Sio F, Mecacci, G (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 34(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Walker MU (2006) *Moral Repair*. Cambridge: Cambridge University Press..