Research Proposal

Amir Konigsberg, PhD

Epistemic Risk and the Trajectory of Model Scaling

The dominant trajectory of frontier AI development is one of scale: larger models, trained on ever-larger corpora, deployed to ever-larger publics. This research examines a class of risk that this trajectory generates but that current safety and ethics frameworks largely overlook, risk that is epistemic in nature, cumulative rather than acute, and located not in any single model but in the information environment as a whole.

As successive model generations are trained on corpora increasingly composed of machine-generated text, the referential link between what systems assert and what is independently true begins to attenuate. The dynamic is adjacent to what the technical literature calls model collapse, but the more consequential version operates one level up, at the scale of the shared information ecosystem rather than the individual model. Beyond some horizon of iterations, what degrades is the epistemic substrate itself: the common, independently verifiable ground against which knowledge claims are checked. My aim is to articulate this risk and to work through its ethical and philosophical implications.

I am guided by three questions. The first is conceptual: what exactly is being lost when we speak of a degraded epistemic environment, and how should we characterise it in terms philosophy already possesses, testimony, justification, the social epistemology of knowledge, without collapsing the new phenomenon into older ones? The second concerns human judgment: if the materials from which people form beliefs are progressively shaped by systems whose outputs are decoupled from truth, then the very inputs to human reasoning are compromised, and with them the capacity to arrive at and stand behind one's own understanding rather than inherit it ready-made. What would safeguarding that capacity require? The third is a question of responsibility: where does obligation sit for a harm that is diffuse, gradual, and the product of many hands, and what would an account of epistemic integrity adequate to this moment actually demand of those who build and deploy these systems?

The intended output is a sequence of papers and a framework piece suitable for both philosophical and policy audiences, connecting the empirical reality of scaling dynamics to a normative account of epistemic responsibility. The work draws on my background at the intersection of interactive epistemology and cognitive psychology, and on twenty-five years in AI, including direct work on information integrity and adversarial threats to the information environment in a national-security context, which has shown me concretely how quickly epistemic degradation becomes a societal and security problem.

The Institute is the natural setting for this. Its interdisciplinary anchoring in philosophy, alongside law, social science and computer science, is precisely what a problem of this shape requires: technically grounded, but ultimately a question about knowledge, judgment, and the conditions of a functioning democratic society.