SILVIO FUNTOWICZ

AI and the Promise of Organized Knowledge

Institutions that seek certainty from artificial intelligence risk losing the habits of self-correction and accountability that have made them resilient.

An emergency physician who is struggling to understand a patient's deteriorating lab results asks an artificial intelligence system for help. The answer arrives in a definitive form, with seeming fluency and confidence. Where a human doctor, faced with the same numbers, might admit uncertainty, order more tests, and proceed with caution, clinical AI tools do not reliably communicate whether a question exceeds their competence or whether the evidence underlying their recommendations is sufficient. Even when excellent data are available, weakly supported and strongly supported answers can be delivered in the same authoritative form. Clinicians are expected to acknowledge the limits of their knowledge in their quest to "do no harm." An AI given similar authority may fail to make that boundary equally visible.

In Louisiana, a prisoner can lose the opportunity for a parole hearing because of a risk score determined by the TIGER system, an algorithmic model for predicting recidivism developed to guide correctional interventions. Under a 2024 change in the law, only those with a low TIGER risk score can go before the parole board; everyone with a moderate or high score is automatically disqualified. However, the score relies heavily on pre-incarceration characteristics and does not credit subsequent rehabilitation efforts. *ProPublica* and *Verite News* documented multiple cases of prisoners whose hearings were canceled on this basis. Among them was a nearly blind 70-year-old man in a wheelchair who had taken anger management classes, learned a trade, and enrolled in drug treatment. He was deemed ineligible for parole because TIGER judged him a moderate risk of reoffending. The problem with TIGER is not only its predictive accuracy; it's that the score it generates was designed to guide rehabilitation programming—not to assess suitability for parole. Even so, it has been given authority to foreclose institutional judgment on prisoners.

In scientific publishing, the start-up QED Science developed an AI system to assess the originality and validity of life-science manuscripts before publication. In November 2025, the major preprint service bioRxiv launched a pilot integration with QED Science that lets authors send papers directly to the service for AI review. This may be useful quality control. It also raises a key question: Is an assessment of the relation between claims, evidence, and other papers sufficient to determine whether an inquiry is original? And what if it is not papers that need examination but the standards of a field itself? Is AI the kind of peer that can carry out that sort of review?

These cases differ in consequences and technological details, but they share a pattern. Essential institutions, including those we rely on for health care, justice, public policy, and the evaluation of research, are increasingly delegating judgment, in whole or in part, to automated systems. And the outputs of these

systems can acquire authority more quickly than the tools needed to scrutinize them can be created.

In transferring judgment to algorithms, these institutions aggressively pursue one of the foundational visions of Enlightenment and post-Enlightenment science. From philosopher Francis Bacon's novel *New Atlantis* to science administrator Vannevar Bush's 1945 report *Science, the Endless Frontier*, modern societies, even those arrayed along very different cultural and political lines, have presumed that knowledge is what I call *organized promise*—that scientific and technical power, properly institutionalized, secures collective progress. A key element of this promise is progress toward certainty: the reduction of error, ambiguity, and dispute.

Artificial intelligence inherits and intensifies this promissory tradition. By generating immediate, confident answers, AI offers the impression that uncertainty itself can be managed away. Yet many of the uncertainties inherent in AI—as well as those surrounding climate, pandemics, financial systems, and biotechnology—are permanent features of complex societies.

This is one of the insights of post-normal science (PNS), which fellow philosopher of science Jerry Ravetz and I described decades ago: PNS is not a method but a name for the class of problems in which closure is not available. We showed that where facts are uncertain, values in dispute, stakes high, and decisions urgent, knowledge relevant to public decisions cannot be judged solely through disciplinary or professional expertise—much less through computational pattern matching and conformity with existing knowledge. Quality, in that setting, means fitness for the purpose at hand, and it depends not only on method but on the arrangements through which claims are scrutinized and the institutions offering them are made accountable. Accountability provides the grounds on which knowledge claims acquire authority. Institutions that mistake the appearance of certainty for authority risk losing the habits of self-correction and accountability that make them resilient.

Artificial intelligence is commonly treated as a technological breakthrough, a source of economic disruption, or an existential threat. Those descriptions obscure the institutional form it already takes. AI, as it is presently constituted, is a diverse set of computational architectures and applications that is transforming the social conditions under which knowledge is assessed and governed.

Many of AI's most important institutional implications are already here. Automated systems will continue to develop, but the arrangements by which institutions decide what to rely on have changed within a few years, without a defined process, expert mandate, or democratic consultation. PNS helps identify how policymakers and institutional leaders might protect or build those capacities and how institutional resilience might be strengthened.

Accountability, trust, and the moral crumple zone

Trustworthy institutional authority depends on accountability. Human knowledge, however imperfect, can be made accountable because a claimant can be identified, questioned about evidence and reasons, challenged by others, and held answerable for judgment. A large language model (LLM), one prominent form of AI today, stands in a different relation to accountability. An output may have identifiable sources and a recorded prompt, yet the system is not an author in the institutional sense. It cannot be questioned as a responsible claimant or held answerable for the consequences of a judgment. These unaccountable systems are increasingly used to draft reports, summarize research, navigate vast bodies of information, and support professional decisionmaking.

To be sure, the challenge of establishing trustworthy knowledge under conditions of uncertainty predates AI. It arises because evidence is incomplete, experts disagree, values conflict, and decisions cannot wait. Traditional scientific procedures have never been sufficient on their own. Better facts cannot entirely resolve these conditions. Yet governments, corporations, and other organizations adopting knowledge automation routinely treat AI output as the fulfillment of Leibniz's dream of a *calculus ratiocinator*, a means of resolving disputes through calculation and abstract reasoning. Today's AI systems carry this dream into domains once thought resistant to formalization, such as language and social decisionmaking.

How did AI so rapidly come to appear an authoritative source of knowledge, even in these novel arenas? A key factor is investment. The prospect of revolutionary innovation helps secure investment, shape policy, and transform institutions, but expectations about the technology outpace the evidence needed to assess whether those expectations have been met. Long before results are in, resources are committed and authority reorganized, so that path-dependency is quickly established.

Institutions increasingly transfer work that once required human judgment (interpreting information, classifying cases, making recommendations, drafting documents, evaluating evidence) to AI, acting on the assumption that these systems can replace human expertise. Trust moves with the work. When institutions treat AI outputs as reliable contributions to knowledge without requiring careful checking, the systems acquire authority: the recognized power to shape what counts as robust knowledge, sound judgment, or valid evidence.

Language contributes to this ascription of authority. Terms like "reasoning," "understanding," and "hallucination" are borrowed from human mental life. They can be useful shorthand, but they import expectations associated with judgment and responsibility. An AI system can generate a plausible explanation, classification, or image

without becoming a claimant who has reasons for believing in its own output or who can be held answerable for asserting it. Yet the language of mental agency turns a functional capacity into an apparent claim of authority. As metaphors, these words make AI seem to do what humans do. They carry prescriptive force alongside descriptive force, shaping the expectation that AI-generated knowledge can be authoritative in the same way as human-generated knowledge.

Fluent presentation can also outrun calibrated reliability. Experiments have found that users can overestimate LLM accuracy when answers are accompanied by explanations. In medicine, even physicians trained in AI literacy have shown automation bias—overreliance on automated outputs—when an LLM supplied erroneous diagnostic recommendations. Fluency can create an appearance of confirmation, and lend it authority, even when the underlying answer is wrong. The danger lies in the ease with which statistical output can appear and then be treated as reliable knowledge.

Institutional settings amplify these risks. Much discussion of AI safety emphasizes keeping a "human in the loop." The practical question is whether meaningful oversight is possible in the context of deployment. Pressures to cut costs and standardize procedures are pushing organizations toward heavy reliance on AI outputs to demonstrate efficiency, sidelining human review. A clinician may be required to approve an AI-generated summary, for example, but be subject to time and resource constraints that make careful scrutiny difficult. Under such conditions, human oversight becomes a procedural routine with little checking power.

When failures occur, blame often falls on the frontline professional who signed off on the output, if indeed anyone is held responsible. The institutional decisions that structured the failure, together with the limits of the technology as a sociotechnical system, escape scrutiny.

Medicine provides a particularly clear case. Physician and researcher David Toro-Tobon and colleagues have argued that "clinician in the loop" arrangements can shift responsibility for AI safety from AI developers and clinical institutions to doctors. The legal structure already points in that direction. In a June 2026 report on UK law, the Medical Protection Society, a medical indemnity organization, warned that clinicians and the National Health Service could become the default targets of claims arising from defective AI systems because developers may fall outside existing product-liability rules. Where that boundary falls is a legislative choice rather than a property of the technology. The European Union's recently revised product-liability directive, which member states must transpose into national law by December 2026, expressly brings software, including AI systems, within the definition of a product. Similarly, the Medical Protection Society is asking the UK government to "classify AI systems as products subject to strict liability." Without such protections, a clinician may be held answerable for following a faulty recommendation and also for overriding one that later proves correct, even where time, information, or institutional authority make independent scrutiny unrealistic. A human signature from an "operator" in the loop merely locates liability at the point where control is weakest.

Madeleine Clare Elish, a cultural anthropologist who leads evaluations and responsible development at the AI research lab Google DeepMind, calls this the *moral crumple zone*: Frontline humans absorb blame for failures built into sociotechnical systems while the deploying institutions and broader governance structures remain largely insulated from responsibility. The term borrows its logic from automotive engineering. A car's crumple zone absorbs the force of a collision so that the frame around the passenger remains intact. In Elish's institutional analogy, the absorbed force is moral and legal liability. The human closest to the system takes responsibility for outcomes substantially determined upstream. When something goes wrong, attention converges on that person, and the decisions that shaped the failure recede from view.

This arrangement suits institutions, whether or not anyone intends it that way. Moral crumple zones allow an organization to claim the efficiency of automation while concentrating accountability on individuals with limited power to question or correct the system. Genuine oversight requires time and standing to interrogate outputs before acting. It also requires authority and responsibility to remain coupled: If the human is answerable for the decision, the institution must give that human real capacity to challenge or reject the AI output.

Living with uncertainty, establishing pedigree

Computers are often imagined as machines of perfect precision, yet every computational system depends on approximation. Engineers manage these limits through tolerance and validation. Reliability depends on carefully designed technical and institutional practices for managing those approximations.

AI systems further complicate the assessment of reliability. Typically, reliability is established by benchmarks. But benchmark performance may not transfer to the institutional setting in which a system is used. A further challenge is that the basis of a particular model output can be difficult to reconstruct, and therefore examine, at the point of decision. The speed and scale at which AI systems operate compound the problem: An error or hidden assumption can be reproduced across thousands of cases before it is recognized.

The original Epic Sepsis Model shows why external validation matters. The electronic health records company Epic designed this algorithm to detect likely cases of sepsis on the basis of patients' health information. Epic claimed that its model accurately discriminated between septic and nonseptic patients between 76% and 83% of the time. But when

researchers at the University of Michigan's medical center tested the model—which was already running in hundreds of US hospitals—on nearly 40,000 hospitalizations, it accurately predicted sepsis only 63% of the time, performing just a bit better than one would expect on the basis of pure chance. It was also prone to false negatives, missing around two-thirds of sepsis cases. A later validation in two hospital emergency departments also found poor performance. Epic subsequently released a revised model designed for local fine-tuning; multicenter external validation in 2026 found some improvement, but performance was uneven across sites. Most alerts were still false alarms, and the overall alert burden was high. Notice the sequence here: External testing exposed limitations that internal assessment had not, and the model changed in response. Credibility depends on testing in context and making arrangements for challenge and correction.

Institutions that depend on algorithmic recommendations and numerical scores must decide which outputs deserve trust, and on whose authority. AI systems cannot themselves supply the validation on which the authority of their outputs depends, so institutions need external practices of testing and revision that remain open to challenge. That requires judgment and transparency, sustained by accountability. An institution that reaches for AI to manage away uncertainty trades the deliberation and scrutiny that maintain the institution's trustworthiness under pressure for an illusion of closure.

The institution's task, I would argue, is not to produce certainty where none exists. It is to build tolerance for uncertainty into its operations, acknowledging what is unknown and preserving the capacity to revise judgment as conditions change. Outsourcing that tolerance to an algorithm hides uncertainty and erodes the capacities the institution will need when uncertainty surfaces.

The late social scientist Steve Rayner used the term *uncomfortable knowledge* to describe information that institutions possess but organize themselves not to act on because acknowledging it would threaten legitimacy or existing commitments. An institution may know, for example, that a model performs unevenly across settings, or that human review has become largely nominal, yet find that acting on this knowledge would disrupt established workflows, investments, or claims of efficiency. The knowledge is available, but organizational commitments make it difficult to use, and so it remains hidden.

Pedigree, a category Jerry Ravetz and I developed as part of NUSAP, our notational scheme for improving scientifically informed policy debates, brings such uncomfortable knowledge, including uncertainty, into the open. Pedigree is a tool for assessing the quality of knowledge. Accuracy remains important. But pedigree also asks how knowledge was produced, under what assumptions, by whom, for what purpose, and with what recognized limitations. It helps evaluate whether a system's output performs reliably under actual conditions of use as well as under ideal or benchmark conditions. The example of Louisiana prisoners shows what is lost when the question of purpose goes unasked. TIGER was built to guide correctional interventions, but when used as a gate on parole eligibility, it carries an authority its pedigree does not support.

Assessing a particular claim and its evidential support is the work of *quality pedigree*. A second level of analysis, *epistemic pedigree*, concerns the character and limits of the field or mode of inquiry itself: what its methods can reveal, what its framing excludes, and for what purposes its results are adequate. AI may become very useful at the first level, but success there does not confer authority at the second level.

Moral crumple zones allow an organization to claim the efficiency of automation while concentrating accountability on individuals with limited power to question or correct the system.

The distinction matters because a technically strong assessment may still inherit the limitations of the data, methods, or framing on which it depends. Consider a simple illustration: An AI system generates a medical recommendation from patterns in its training data. The output looks precise and confident. Without knowing the pedigree of the underlying data—how they were collected, which populations they represent, what biases or systematic errors they contain, and how complete they are—the recommendation cannot be properly evaluated. A system trained primarily on data from one demographic group may perform poorly on another. Historical data may reproduce historical inequities. Systematic errors in training data may also be reproduced or amplified. Pedigree makes these limitations visible enough to be weighed in decisions and opens sources of uncertainty to examination and debate.

By making these limits explicit, pedigree could help to prevent the uncritical adoption of AI systems and support more responsible forms of governance. Placing AI outputs within a broader process of scrutiny and judgment opens their authority up to challenge.

Extended peer governance

Post-normal science locates pedigree within governance through what my colleagues and I have called the *extended peer community*. In this community, credentialed experts are joined by people affected by decisions, those responsible for implementing them, and those whose practical or experiential knowledge may reveal consequences that formal expertise misses. The point of an extended peer community is to build a plurality of human perspectives to widen the tests to which claims are exposed and make blind spots easier to detect.

The value of such plurality is that it can prevent failures of expert-only governance. Technology users often understand failure modes that designers do not anticipate, because they must live with the consequences. And during public health emergencies, patients, practitioners, and ordinary people sometimes recognize on-the-ground realities before experts can. In Flint, Michigan, residents' observations and sampling of their water, undertaken with researchers at Virginia Tech, helped expose lead contamination after officials had given assurances of safety. Another example occurred during the early months of the COVID-19 pandemic, when health authorities had to issue guidance before clinical trials could settle the underlying science. While debating that guidance in real time, epidemiologists, frontline clinicians, and the public were performing the function of an extended peer community, albeit under conditions no one would have chosen. The point is not how well this improvised peer community performed, but that its function proved unavoidable even though no institution had made provisions for it.

AI governance has the characteristics PNS was developed to address. AI systems have failure modes that are uncertain and context-dependent; the systems are deployed rapidly, often before their risks are understood; and the consequences of AI use are unevenly distributed. Leaving governance solely to technical experts or corporate insiders risks leaving crucial harms and vulnerabilities unexamined. The teachers, clinicians, public servants, workers, students, and patients most affected by AI will have insights into how these systems behave in practice.

Some of the extended peer communities relevant for AI governance are already visible. Writers and illustrators have organized lawsuits against the use of their work as training data; communities near Memphis have contested the siting and emissions of the AI company xAI's data centers; data labelers in Kenya have organized over working conditions; and families and state authorities have brought the harms of chatbots before the courts. Each case identifies a concrete deployment and parties with something demonstrably at stake. It is only when deployments and affected parties become identifiable that such extended peer communities can form around concrete controversies.

Governing knowledge responsibly

AI therefore presents institutions with a paradox. It intensifies an older promise of organized, frictionless knowledge—immediate answers, standardized judgment, and fewer visible disputes. Yet institutions that would use AI responsibly have to cultivate precisely the capacities that this promise obscures: They must keep uncertainty visible, preserve room for contestation of their criteria as well as their conclusions, and do the harder work of determining how knowledge gains authority and who is accountable for its use. Some friction is a condition of reliable knowledge.

Extended peer communities provide one way to build those capacities into institutions. As Flint's story shows, such communities often form without invitation; the institutional task is less to convene them than to make room for them and to take seriously what they find. They place expertise within a broader framework of accountability and democratic legitimacy, and they create the space in which tools such as pedigree can be meaningfully applied. Their assessments include technical performance, the purposes for which a system is being used, the standards of the field itself, the consequences of error, and the experience of those who must live with its effects. For institutions, this matters because challenge and correction are part of reliability itself. Procedures that appear inefficient from the standpoint of frictionless automation may be precisely what makes uncertainty visible.

Undoubtedly, AI systems will evolve to be more sophisticated and more accurate, but that will not settle the essential institutional questions the technology poses. How do AI outputs acquire authority? When does computer-generated text become a reliable contribution to knowledge, a sound basis for judgment, or a legitimate ground for decisions? Accuracy, usefulness, and fluency do not confer authority. Authority rests on the institutional capacity to verify claims, expose them to challenge, and revise them when warranted.

Institutions frequently treat AI adoption as a technical problem, but it is a political one. As the production and certification of knowledge are transformed, power determines how societies govern it: who may confer authority, who can challenge it, whose experience counts, and who bears the consequences when judgment fails. Authority in these new domains requires institutions that acknowledge uncertainty and make their values and procedures explicit, so that technical capability remains subject to open democratic oversight. Those are the grounds on which authority can be earned.

Silvio Funtowicz *is a philosopher of science and the codirector of the European Centre for Governance in Complexity. He is also a guest researcher at the Centre for the Study of the Sciences and the Humanities, University of Bergen, Norway.*