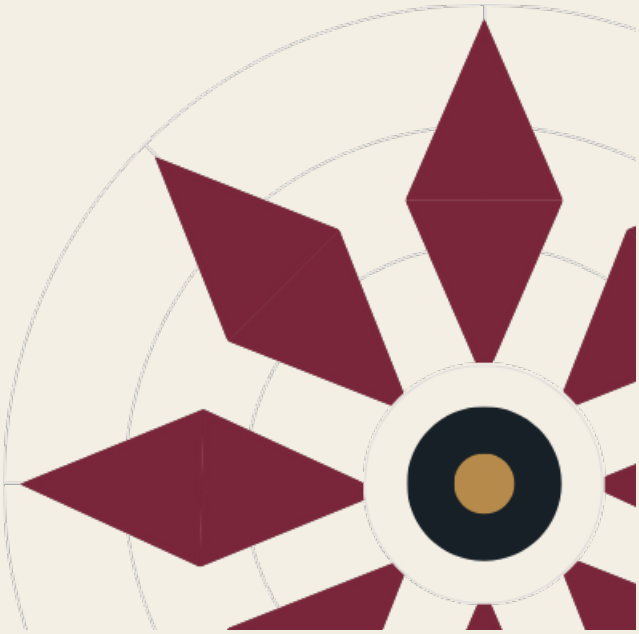
THE MIRANDOLA INSTITUTE

HumanitAI

An Oration for the Age of Machines



VOLUME 1

MIRANDOLA INSTITUTE · 2026

Table of Contents

An Oration for the Age of Machines

In 1487, a young scholar proposed to defend nine hundred theses in Rome, in public, against any Doctor of Theology willing to argue them. He wrote, as the opening speech for a disputation that would never take place, an account of why human beings are distinguished not by any fixed nature but by the capacity to determine, through inquiry, what they become. The disputation was suspended before it could occur. The oration survived it.

This is the oration this Institute owed before it could open a disputation of its own.

The essays collected here do not argue that artificial intelligence should be built differently, or regulated differently, though both questions matter. They argue that the more consequential question is what human beings and human institutions become as machines capable of reasoning, generating, and acting become part of ordinary life — and that this question has not yet been asked with the specificity it deserves. Six essays state that argument from six directions.

"Doctrine Without a Church" asks why this question is urgent now: convergent AI output, we argue, produces a form of epistemic influence with no historical precedent in its combination of scale, speed, and invisibility — a doctrine with no church to argue with.

"Capability as the Dependent Variable" asks what, precisely, this Institute studies: not whether artificial intelligence improves performance, but under what conditions it leaves a more capable person or institution behind it, a question three adjacent literatures have approached in pieces but not asked whole.

"The Complement Becomes the Target" asks when, and why now rather than at any other point in the history of automation: prior general-purpose technologies substituted for routine work and complemented reasoning and communication; this one, the evidence suggests, does the reverse.

"The Second Door" asks where this question bears most unevenly: on the developing economies for whom a services-led path to convergence, opened by digital connectivity, may be closed by the same technology that opened it.

"Who Is the AI Humanist?" asks who is being asked to take this up, and what capacities that requires: not fluency with a tool, but the retained authorship of one's own conclusions, formed in a tradition that was civic before it was private.

"How a Claim Earns Trust" asks how an institution making claims of this kind ought to conduct itself, including in its own use of the tools it studies — a question this Institute answers by adopting, from a field that already had to answer it under pressure, the discipline of committing in advance to what would prove it wrong.

This introduction, and the essays it precedes, carry no byline. That is deliberate, not an oversight. What follows is offered as the Institute's own position, not any individual's, in the manner of an argument made to be judged on its merits rather than its author's standing — a convention with its own precedent, and one this Institute intends to keep for founding statements even as it comes, in time, to publish many names.

Every claim made in this volume is stated so that it can be challenged. The terms of that challenge, the Institute's use of artificial intelligence in preparing these essays, and the disclosures that accompany both are set out together at the close of this edition, indexed thesis by thesis, rather than repeated after each argument. We ask readers to treat that closing section as part of the volume, not an appendix to it.

The question this Institute exists to keep asking has one form, repeated in what follows from several directions:

What can human beings become in an age of capable machines?

This edition begins the answer. It does not end it.

Capability as the Dependent Variable

Five theses, and an invitation to prove them wrong

Under what conditions does an increase in machine capability produce a durable increase in human or institutional capability?

This is the question the Mirandola Institute exists to answer. It is not a new worry dressed in new language. Three established literatures already study pieces of it — cognitive offloading, automation complacency, and the labor economics of human-AI complementarity. Each has real findings. None of them, on its own terms, asks what we are asking.

This essay states that argument as five theses and invites anyone, without credential or introduction, to challenge them. This is, deliberately, an old form. The Institute's namesake opened his own century with nine hundred theses and an offer to defend every one of them, in public, against any scholar in Europe who wished to try. We are not claiming his ambition. We are borrowing his method.

Three literatures, three dependent variables

Cognitive offloading asks what happens to memory when retrieval is externalized. The founding result is Sparrow, Liu, and Wegner's 2011 study in ***Science***, which found that people who expect information to remain accessible online encode the information itself less effectively, while encoding ***where to find it*** more effectively — the "Google effect." A 2025 MIT Media Lab study (Kosmyna et al., released as a preprint and not yet through full peer review) extends this to generative AI directly: writers using an LLM showed weaker neural connectivity than those using a search engine or working unaided, and struggled to recall or quote from essays they had just written — "cognitive debt." The dependent variable throughout is narrow and specific: what a person retains in memory after using a tool.

Automation complacency asks what happens to vigilance when a system can be trusted to act on its own. Parasuraman and Riley's 1997 framework — use, misuse, disuse, and abuse of automation — has organized decades of aviation and clinical research on why skilled operators stop monitoring systems that usually work, and what happens when those systems fail. The dependent variable is

behavioral: does the human keep watching. The remedies proposed tend to live at the level of interface design and individual training. The literature diagnoses an individual failure mode inside an otherwise unexamined institutional structure.

Complementarity economics asks what happens to output and wages when workers gain access to AI. Noy and Zhang (2023), in a randomized experiment with 453 professionals, found that ChatGPT raised average output quality on writing tasks and ***reduced*** the gap between stronger and weaker performers. But the mechanism, in the authors' own words, was that the tool "mostly substitutes for worker effort rather than complementing worker skills." Dell'Acqua and colleagues' 2023 field experiment with 758 BCG consultants found large gains on tasks inside GPT-4's capability range and a 19-percentage-point ***degradation*** on tasks outside it, driven partly by consultants misjudging where that boundary sat. Both studies measure output. Neither was designed to determine whether the humans involved became more or less capable, independent of the tool, by the end of the study.

Five theses

Thesis 1. Cognitive offloading research measures what a person retains in memory after using a tool. It does not measure, and is not designed to measure, what happens to that person's judgment, expertise, or institutional capacity.

Thesis 2. Automation complacency research diagnoses a failure of individual vigilance and proposes remedies at the level of the individual and the interface. It does not, as a rule, ask whether institutional structure — mandate, bounded authority, evidence, escalation — could make the outcome independent of any one person's sustained attention.

Thesis 3. Complementarity economics measures output, quality, and wages, and treats skill largely as a fixed variable that determines who benefits. It does not, as a rule, measure whether the underlying skill itself grew, shrank, or was substituted for.

Thesis 4. No literature currently in wide circulation tracks the same person or institution across individual, institutional, epistemic, and governance levels at once, longitudinally, to ask a developmental question: not ***did performance improve***, but ***what remains, and under what conditions would more remain***.

Thesis 5. A research program that makes capability formation itself the central, cross-level, longitudinal dependent variable — rather than an incidental finding inside a study built to measure something else — is therefore a distinct contribution, not a restatement of existing work.

None of these theses claims the adjacent literatures are wrong. Thesis 4 in particular is a claim about absence, which is the easiest kind of claim to falsify: a single existing study that does track capability formation longitudinally, across levels, as its primary design, would defeat it outright. That is by design. A thesis that cannot be defeated by evidence is not worth defending.

A narrower gap than it may first appear

The three literatures above are not the only ones that study how capability forms. Expertise-acquisition research, following Ericsson, Krampe, and Tesch-Römer's (1993) account of deliberate practice, studies exactly the developmental question Thesis 4 asks — what conditions convert sustained effort into durable skill — in detail this essay does not attempt to reproduce. Organizational-learning research, following Argyris and Schön (1978), studies how institutions convert experience into altered practice, which is a close cousin of what this Institute calls Institutional Integrity. Skill-formation economics, following Cunha and Heckman's (2007) model of dynamic complementarity — the finding that skills acquired early make later skill acquisition more productive, not merely additive — studies capability formation as a rigorous, longitudinal, quantitative object in its own right.

Thesis 4 would be false if any of these literatures already did what this essay claims none of them does: track a person or institution across individual, institutional, epistemic, and governance levels at once, in relation to AI-mediated delegation specifically, as a primary research design. None of them does this, because none of them was built to. Deliberate practice studies skill formation through sustained, structured effort, not through the specific dynamics of delegating a task to a system capable of performing it unassisted. Organizational learning studies institutions adapting to experience, not to a form of capability that can be borrowed from outside the institution entirely, on demand, at near-zero marginal cost. Dynamic complementarity models how a person's own skills build on one another; it does not model what happens to that sequence when a machine can supply, at any stage, the skill the sequence assumes the person is building.

The gap this essay identifies is accordingly narrower than "capability formation is unstudied." It is not. The gap is that no existing framework asks how AI-mediated delegation interacts with the capability-formation processes these adjacent literatures have already established — whether

delegation accelerates the sequence Cunha and Heckman describe, substitutes for the practice Ericsson's tradition treats as necessary, or bypasses the institutional learning Argyris and Schön describe entirely. That is a real and, we think, still-open question. It is also a considerably more precise one than the essay's first draft posed, and the precision is owed to those who pointed out the omission.

What this is not

This is not a claim that AI Humanism supersedes offloading, complacency, or complementarity research. The Capability-Conversion Problem is built directly on what those literatures have already established; it could not be asked without them. The claim is narrower: that none of them, individually or together, currently treats capability formation as the thing to be measured, cultivated, and designed for, across the levels at which people actually act.

An invitation to prove these wrong

The Institute takes its name, and this format, from Giovanni Pico della Mirandola, who at twenty-three proposed to defend nine hundred theses in Rome against any doctor of theology who would come argue them, regardless of that doctor's rank or his own. Any of the five theses above may be challenged in writing, by anyone, without credential.

This essay's claims are subject to challenge under the terms of the Institute's Disputation Protocol. Editorial disclosure, including this publication's AI review process, appears in the closing section of this edition.

References

Argyris, C., & Schön, D. A. (1978). ***Organizational Learning: A Theory of Action Perspective***. Addison-Wesley.

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. ***American Economic Review***, 97(2), 31–47.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. ***Harvard Business School Technology & Operations Management Unit Working Paper No. 24-013.*** Published in ***Organization Science*** (2025).

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. ***Psychological Review***, 100(3), 363–406.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. ***arXiv preprint*** arXiv:2506.08872.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. ***Science***, 381(6654), 187–192.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. ***Human Factors***, 39(2), 230–253.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. ***Science***, 333(6043), 776–778.

Doctrine Without a Church

Five theses on algorithmic convergence and the case for AI Humanism

In 1487, a twenty-three-year-old scholar proposed to defend nine hundred theses in Rome, in public, against any doctor of theology willing to argue them. Giovanni Pico della Mirandola's ***Oration on the Dignity of Man*** — written as the opening speech for a disputation that was suspended before it could occur — argued that human beings possess no fixed nature, only the capacity to shape themselves through inquiry. The claim was radical because the alternative, the one the Church offered, was a fixed and given order: what to believe, in what sequence, on whose authority, with dissent named, tried, and punished.

Historians of the period have long located the deeper mechanism of that century's transformation not in any single confrontation but in a shift in method. Lorenzo Valla's 1440 philological demonstration that the ***Donation of Constantine*** — the document underwriting centuries of papal temporal authority — was a forgery is the example most often reached for here, and it deserves a more careful telling than the usual one. Valla did not defeat the Church by force, and he did not defeat it quickly: the demonstration was itself entangled with his employment under Alfonso V of Naples, then in dispute with the Papacy over territorial claims the ***Donation*** helped underwrite, and the document's authenticity continued to be asserted and defended for generations after Valla's philology had, in fact, settled the question. What Valla actually accomplished was narrower than "method defeated authority": he showed that a specific claim could be tested against evidence, and shown false, by anyone equipped to read the Latin against its purported era — a capability, not an outcome. Whether that capability changes what an institution does, and how long that takes, is a separate question, and the gap between the two is part of what this essay's own case has to reckon with. Vernacular translation of scripture, from Wycliffe through Tyndale, extended the same logic from philology to access — if the text is available in a language people can read, doctrine becomes something to evaluate rather than something to receive. The Renaissance did not abolish authority. It insisted that authority answer to method, evidence, and an individual capacity for judgment that could, in principle, be exercised by anyone who developed it — while leaving open, then as now, how long it takes for that capacity to change what authority actually does.

We think a structurally similar problem is forming now, and that the parallel is closer than metaphor. But the differences matter more than the similarity, and stating them precisely is the only way this argument earns the comparison rather than merely borrowing its drama.

What is actually converging

The claim that AI systems produce convergent output is not speculative. Kleinberg and Raghavan (2021), writing in ***PNAS***, formalized what they termed algorithmic monoculture: when many decision-makers rely on the same or similar algorithms, their outputs and errors become correlated in ways that can reduce social welfare even as each individual system's accuracy improves — a risk they compare to the vulnerability of biological monocultures to a single pathogen. Bommasani and colleagues' widely cited 2021 report on foundation models, from Stanford's Center for Research on Foundation Models, extended the concern directly to the current generation of AI: because so many downstream applications are built on a small number of shared base models, whatever patterns, gaps, or biases exist in those base models propagate outward to everything built on top of them.

The effect is measurable, not merely theoretical. A 2025 study evaluating twenty-one state-of-the-art language models against a dataset of genuinely diverse human preferences found that the models' collective outputs aligned with only 41 percent of the variation present in actual human preferences — the models agreed with each other far more than humans agree with each other, regardless of which model or combination of models a person consulted.

This is a different phenomenon from the one that concerned an earlier generation of technology critics, and the difference is precise, not rhetorical. Pariser's 2011 account of the "filter bubble" and Sunstein's work on echo chambers described a curation problem: algorithms selecting which existing, human-authored content a person encounters, narrowing exposure without narrowing the underlying diversity of what had been written. The dependent variable in that literature is exposure — what you are shown from among things other people wrote. The dependent variable in algorithmic monoculture, and in what generative AI now does at far greater scale, is different in kind: the system does not select among existing positions, it generates the position itself, in the voice of a reasoning interlocutor, with no visible author and no indication that the same underlying convergence produced a substantial share of everyone else's answer as well. Curation narrowed what people saw. Generation now supplies what people think they arrived at independently.

What this is not

This is not a claim that a Church exists, or that anything resembling one is coordinating the outcome. That distinction is the whole argument, not a footnote to it.

The Church Pico confronted was a single, doctrinally unified institution with an explicit canon, a named authority, and the coercive power to try, excommunicate, and execute dissent. The convergence described above has none of these properties. It arises from competing firms with different owners and different commercial incentives, converging anyway — because the available

training data is drawn from a largely overlapping internet, because alignment processes optimize toward similarly cautious and inoffensive output, and because competitive markets reward resembling a safe consensus more than they reward visible divergence from it. No memo produced this outcome. No one needs to have intended it for it to occur, which is a harder problem to name and resist than one with an address in Rome, not a lesser one. And no AI system today can compel a person's belief the way ecclesiastical courts once could compel conduct: there is no equivalent of trial, exile, or execution for the person who consults a different source or reasons differently. The influence here is structural, not coercive — closer to the lineage of McLuhan and Postman's arguments about how a medium shapes thought by its form, or to Zuboff's account of instrumentarian power operating through default and design rather than command.

Nor is this a claim that the convergence is evenly distributed in its effects, or that no pattern can be traced in what it converges toward. Noble (2018) documented that search systems trained on unrepresentative data reproduce and amplify racist associations already present in that data; Buolamwini and Gebru (2018) found that commercial facial-analysis systems misclassified darker-skinned women at rates dramatically higher than light-skinned men, because the benchmark datasets those systems were evaluated against were overwhelmingly composed of lighter-skinned subjects. Neither finding required a coordinating authority to produce; both are convergence toward a pattern, in the sense this essay uses the term, and the pattern is not neutral. The absence of a deliberate coordinating actor does not mean the convergence this essay describes is authorless in every sense. It means its authorship is diffuse and structural — traceable to whose language, whose faces, and whose labeled judgments the training and evaluation data privilege — rather than institutional and nameable in the way a Church's was. That makes the convergence harder to hold any single party responsible for. It does not make the convergence pattern-free.

Nor is the core difference from the Renaissance case a matter of the public being more credulous now than then. Most people under Church doctrine also experienced it as simply how the world was, not as an imposed position to be weighed — awareness of doctrine as doctrine was always the province of disputants like Pico, not the general condition of the age. The sharper difference is speed and scale. Doctrinal convergence took centuries to establish and was repeatedly, visibly contested along the way, by philology, by translation, by print. Algorithmic convergence is being established across a global population within a handful of years of consumer deployment, largely without an equivalent visible act of contestation, because there is no canon anyone perceives themselves as being asked to accept. There is only an answer, arrived at, apparently, on one's own.

Five theses

Thesis 1. Convergent AI output across competing systems is a documented phenomenon, not a speculative one, evidenced in formal models of algorithmic monoculture, in analysis of shared foundation-model components, and in direct empirical measurement of output homogeneity across current systems.

Thesis 2. This convergence differs in kind from earlier concerns about algorithmic curation and media concentration, because generative systems produce the position itself, in the form of independent reasoning, rather than selecting among visibly authored human content.

Thesis 3. Unlike centralized doctrinal authority, this convergence has no single coordinating actor and no coercive power over the individual; it is an emergent property of shared data, shared alignment methods, and converging commercial incentive, not a deliberate policy imposed by a nameable institution.

Thesis 4. The absence of a nameable authority makes this form of convergence harder to identify and resist than doctrinal control was, not because it is more coercive, but because it presents as a neutral, authorless answer rather than as an assertable position that invites a response.

Thesis 5. Because the deficit here is in the individual and institutional capacity to recognize convergence and exercise independent judgment against it, the appropriate response is closer to the humanist project of cultivating that capacity than to either a purely technical fix — mandating model diversity — or a regulatory one that risks recreating a single centralized authority over what counts as acceptable output.

Thesis 5 is the most contestable of the five, deliberately. It is a claim about what follows from the first four, not a claim that follows from them by necessity, and it is the one most directly at stake in whether AI Humanism, as an institutional response, is the right response at all.

Why the parallel still holds, precisely because the disanalogies are real

Pico's actual wager was not that the Church was wrong about any particular doctrine. It was that human beings retained, and should exercise, the capacity to examine any claim on its merits rather than receive it on authority — and that this capacity had to be ***cultivated***, through the ***studia humanitatis***, not simply asserted as a natural right. The absence of a Church to argue with today does not remove the need for that capacity. It may increase it. A doctrine with a name can be identified, translated, and disputed — and, eventually, however slowly, revised when it fails the argument, as happened to the ***Donation of Constantine***, even if the document's political usefulness

outlasted its philological credibility for generations. A convergence with no name and no author cannot be challenged in the same way, because there is no single claim to put to the test — only a pattern, distributed across systems that no one built to agree with each other, that agree anyway.

That is the argument for an institution organized, explicitly, around the cultivation of independent human judgment in an age of increasingly capable machines. Not because the machines are doctrinal authorities. Because they may be producing the effects of one without anyone having built one, and the only remedy on record for that condition is the same one the Renaissance proposed for the version it faced: people capable of examining a claim on its own terms, whether or not anyone can be named as having made it.

This essay's claims are subject to challenge under the terms of the Institute's Disputation Protocol. Editorial disclosure, including this publication's AI review process, appears in the closing section of this edition.

References

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. ***arXiv preprint*** arXiv:2108.07258. Stanford Center for Research on Foundation Models.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. ***Proceedings of Machine Learning Research***, 81, 77–91.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. ***Proceedings of the National Academy of Sciences***, 118(22), e2018340118.

Noble, S. U. (2018). ***Algorithms of Oppression: How Search Engines Reinforce Racism***. NYU Press.

Pariser, E. (2011). ***The Filter Bubble: What the Internet Is Hiding from You***. Penguin Press.

Sunstein, C. R. (2017). ***#Republic: Divided Democracy in the Age of Social Media***. Princeton University Press.

Zuboff, S. (2019). ***The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power***. PublicAffairs.

Cultivating pluralism in algorithmic monoculture: The Community Alignment Dataset. (2025). ***arXiv preprint*** arXiv:2507.09650.

Who Is the AI Humanist?

The scholar-citizen, reconsidered under conditions of machine testimony

The question of who the ***studia humanitatis*** was meant to produce has a precise answer in the sources, and it is not "the well-read." Pier Paolo Vergerio's ***De ingenuis moribus et liberalibus adulescentiae studiis*** (c. 1402), among the founding statements of humanist pedagogy, frames the educational aim not as the accumulation of learning but as the formation of a person capable of exercising judgment and contributing to the life of the city. Leonardo Bruni's letter ***De studiis et litteris*** (c. 1424) makes the same claim by a different route, treating grammar, rhetoric, history, poetry, and moral philosophy — the content Kristeller (1961) identifies as the substance of the ***studia humanitatis*** — not as separate subjects but as the integrated formation of a single capacity: the ability to understand, to argue, and to act.

That formation had, from its origin, a civic cast rather than a purely contemplative one. Baron's (1955) account of Florentine "civic humanism" locates the movement's most consequential early exponents — Salutati, Bruni himself — not in the cloister but in the chancery, as scholars who held public office and understood their learning as an instrument of the republic's political life, not an alternative to it. The person the ***studia humanitatis*** aimed to produce was, by design, a scholar-citizen: formed by private study, answerable to public life.

We take up the same question under different conditions. What should the person who has undergone a modern ***studia humanitatis*** — as set out in the Institute's companion framework of that name — actually be able to do, and for whom is this an achievable ideal? The answer requires attending to a problem the Renaissance humanists did not face in its current form: the growing share of what a person comes to believe that arrives not from another person's testimony, but from a machine's.

The inherited problem: dependence is not new

No epistemology of the AI Humanist can proceed as though dependence on others' knowledge is itself the novelty. Hardwig's (1985) "Epistemic Dependence" established, in terms that remain the reference point for this literature, that no individual possesses the time or capacity to independently verify more than a fraction of what they justifiably believe — that a physicist's confidence in a claim outside her subfield rests, like anyone else's, on trust in colleagues she has

not personally checked. Epistemic dependence on testimony is not a corruption of rational belief; Hardwig argued it is a condition of it, for experts as much as for laypeople.

The question this raises — on whose testimony, and why — has its own literature. Goldman (2001) proposed practical criteria a layperson can use to identify trustworthy experts without possessing the expertise to evaluate their claims directly: track record, credentials assessed by relevant peer communities, and the disclosed interests and biases that might distort an expert's testimony. These criteria work, when they work, because they are anchored in structures external to the testimony itself — a reputation accumulated over a career, a professional community with its own standards, consequences for being wrong that the expert bears personally.

Why the inherited answer does not transfer

None of Goldman's anchors is available in the same form when the testimony comes from a machine. A model has no career and accrues no personal reputational cost for error. Its "peer community," if the term applies at all, is a benchmark leaderboard and a set of internal evaluations largely invisible to the person receiving its output. Its disclosed interests, to the extent they exist, are the commercial and design choices of the organization that built it — several steps removed from the specific claim being tested, and not addressed to the individual relying on it.

This is not a claim that machine testimony is therefore untrustworthy. It is a claim that the ordinary machinery by which a person calibrates trust in human expert testimony does not carry over intact, and that a capacity formed to navigate human testimonial dependence is not, without further formation, adequate to navigate this one. The risk this creates has been named precisely, if not for this exact case. Nguyen (2020) distinguishes an epistemic bubble, in which relevant information is simply missing, from an echo chamber, in which a person has been given active reasons to discount outside sources — and argues the second is far harder to escape because it is structurally self-reinforcing. Fluent, confident, endlessly available machine testimony creates a related risk by a third route: not bubble, not chamber, but the erosion of the impulse to question at all. Nguyen (2022) calls the resulting posture "trust as an unquestioning attitude" — trust that has stopped operating as a considered judgment and become a default.

The distinguishing capacity

The AI Humanist, as this Institute uses the term, is the person formed to resist that default — not by refusing machine testimony, which would forfeit its genuine value, but by retaining the specific capacities that ordinary testimonial trust, transplanted uncritically from the human case, does not supply on its own: the ability to ask what would count as evidence for a claim, to notice when a fluent answer has substituted for a supported one, and to know when a question exceeds what any current system's testimony can settle.

This is a narrower claim than "AI literacy," and the distinction is the substance of the definition, not a footnote to it. Fluency with a tool — knowing how to prompt it, how to chain its outputs, how to use it quickly — is a technical skill, genuinely useful, and orthogonal to the capacity in question. A highly fluent user can still fully surrender judgment to the tool's output; a comparatively unpracticed one can retain it. The Institute's guiding formulation for its Human Agency pillar — delegate execution while preserving human authorship — names the distinguishing mark precisely: not whether a person uses machine testimony, but whether the person using it remains the author of the conclusion, in the sense of being able to say what grounds it and to revise it when the grounds change.

This essay, and this issue of the publication, appear without individual bylines — a convention explained elsewhere in this edition and worth reconciling here directly, since retained authorship is the very capacity this section defines. The two senses of authorship are not the same one. Public attribution names who wrote a sentence. Retained authorship, in the sense this essay uses the term, names whether someone — named or not — can say what grounds a conclusion and revise it when the grounds change. An unsigned institutional statement satisfies this second sense exactly to the degree that a named human editor stands behind it, can account for its claims, and is bound to correct them when they fail; it fails this sense only if no one, in fact, could do so. The convention adopted here rests on the first condition holding, not on its absence being a virtue. A reader entitled to ask "who can say what grounds this, and would revise it if shown wrong" should always have an answer, whether or not that answer appears on the page as a name.

A note on the humanist premise

This essay, and the Institute it serves, take for granted that a specific capacity — the ability to form purposes, weigh evidence, revise a conclusion, and bear the consequences of a choice — is worth cultivating and protecting. That is a narrower claim than it might first appear, and worth distinguishing carefully from two things it is not.

It is not the claim, often associated with transhumanism, that human beings should merge with or be superseded by more capable machines in order to keep pace with them. Wolfe (2010) draws this distinction sharply: much of what calls itself posthumanist thought, in his account, is transhumanist rather than posthumanist, because it remains committed to the very ideals of rational agency and self-perfection that Renaissance humanism first articulated, merely relocating them into engineered bodies. Wolfe traces this genealogy, following Bostrom's own account of transhumanism's intellectual history, back through Pico della Mirandola specifically — the same ***Oration*** this Institute draws its founding proposition from is also claimed as an ancestor by the tradition arguing humans should engineer their way past their present limits. Pico's argument that human beings have no fixed nature, only the capacity to shape what they become, does not by itself settle which shaping counts as fidelity to that capacity and which counts as its abandonment. This essay takes a position on that question — that cultivating the capacity itself is the more

faithful reading than replacing it — without claiming Pico's text forces that reading on any careful interpreter.

What distinguishes the two readings is not, in the end, a matter of textual interpretation alone. It is a matter of what Pico did with the argument. The ***Oration*** was not written as a contribution to a literature; it was written to open a disputation, in public, in which an unlicensed twenty-three-year-old proposed to defend his theses against the established theological authorities of his age, before any of them had granted him standing to do so — a posture of contesting entrenched authority on the merits, at real risk, rather than a claim about what human nature ultimately permits. That posture, not only the argument it introduced, is the inheritance this Institute claims directly: the Disputation Protocol that governs this publication, open to challenge from anyone without credential, is a working instance of it. Transhumanism's own development has, on the whole, proceeded through and alongside existing centers of technological and economic power rather than against them — a generalization with exceptions, but a fair one — extending what power can already do more often than contesting it. Two things, not one, are named by this Institute: a rebirth of the capacity for human thought under conditions artificial intelligence has changed, and the contest against entrenched power that a rebirth of that kind has always required to become more than a private conviction.

Nor is this the claim that human interests should override every other consideration as a matter of species privilege. Academic posthumanism, in the tradition running from Haraway through Wolfe, is substantially a critique of exactly that kind of privilege. What this essay defends is narrower: not that humans matter more, but that the capacity for reflective, revisable, consequence-bearing choice is worth cultivating wherever it is found, and that as far as can currently be established, that capacity resides most fully in human beings. This essay does not take up, and does not need to resolve, what humans owe to animals, ecosystems, or the machines increasingly built among them; recognizing that human welfare is entangled with all three is compatible with everything argued here and is a separate claim, better made by the literatures built to make it.

That formulation has its own vulnerability, and it deserves to be stated rather than left for a challenger to find. Grounding the case for cultivating this capacity in the ability to participate in a debate about it risks quietly excluding everyone who cannot argue their own case — infants, people with severe cognitive disabilities, and, on most accounts, animals — from the moral consideration such an argument is meant to protect. Nussbaum (2006) built the capabilities approach substantially to correct this problem in the philosophical tradition this essay draws on, grounding moral standing in the capacity for flourishing rather than the capacity for rational self-defense, precisely so that beings who cannot debate their own standing are not thereby excluded from it. This essay's commitment to cultivating human deliberative capacity does not require denying that other beings warrant moral consideration on different grounds; it claims only that the particular capacity this Institute studies — deliberation under conditions of increasingly capable

machines — is a human one, and that studying and protecting it does not, in itself, require deciding the broader question Nussbaum's tradition poses.

The citizen half of the formation

Consistent with the civic cast of the tradition it draws on, this formation is not complete as a private discipline. Baron's civic humanists did not stop at cultivating their own judgment; they brought it into chancery and council. The contemporary parallel is participation in the institutions through which AI-mediated claims now travel — asking, in a professional or civic setting, who authorized this system's use, on what evidence its output rests, and who remains accountable if it is wrong. These are the questions this Institute's Institutional Integrity and Governed Intelligence pillars pose at the organizational level; the scholar-citizen is the person who knows to ask them at the individual one, in the meetings and decisions where such questions are otherwise least likely to be raised.

Four theses

Thesis 1. The historical formation the ***studia humanitatis*** aimed at was explicitly civic, not merely contemplative; an account of the AI Humanist that stops at private epistemic discipline, without the civic-humanist component of institutional participation, is incomplete relative to the tradition it claims.

Thesis 2. The criteria by which a layperson ordinarily calibrates trust in human expert testimony — track record, peer standing, disclosed interest — do not transfer to machine testimony in their existing form, and no adequate substitute has yet been established in either the philosophical or the practical literature.

Thesis 3. Fluency in using AI tools is necessary but not sufficient for the capacity this essay describes; the distinguishing mark is retained authorship of the resulting belief or claim, not skill in producing the output.

Thesis 4. The ideal described here is available to any person regardless of professional background or technical specialization, consistent with a tradition whose most consequential disputant held no advanced standing among those he proposed to argue with.

This essay's claims are subject to challenge under the terms of the Institute's Disputation Protocol. Editorial disclosure, including the essay's AI review process, appears in the closing section of this edition.

References

Baron, H. (1955). ***The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny***. Princeton University Press.

Bruni, L. (c. 1424). ***De studiis et litteris*** [Letter to Battista Malatesta].

Goldman, A. I. (2001). Experts: Which ones should you trust? ***Philosophy and Phenomenological Research***, 63(1), 85–110.

Haraway, D. (1985). A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. ***Socialist Review***, 80, 65–108.

Hardwig, J. (1985). Epistemic dependence. ***The Journal of Philosophy***, 82(7), 335–349.

Kristeller, P. O. (1961). ***Renaissance Thought: The Classic, Scholastic, and Humanist Strains***. Harper & Row.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. ***Episteme***, 17(2), 141–161.

Nguyen, C. T. (2022). Trust as an unquestioning attitude. ***Oxford Studies in Epistemology***, 7, 214–244.

Nussbaum, M. C. (2006). ***Frontiers of Justice: Disability, Nationality, Species Membership***. Belknap Press of Harvard University Press.

Vergerio, P. P. (c. 1402). ***De ingenuis moribus et liberalibus adulescentiae studiis***.

Wolfe, C. (2010). ***What Is Posthumanism?*** University of Minnesota Press.

How a Claim Earns Trust

What the replication crisis teaches an institute that studies machines

An institute that asks what human beings and institutions should become as machines grow more capable owes its reader an account of its own method, not only its subject. This essay is, in that sense, unlike the three that precede it. It does not argue primarily about artificial intelligence. It argues about how any claim — including the ones this Institute makes — comes to deserve belief, and it takes that argument from a field that has already been forced to answer the question under pressure: experimental psychology, over the past two decades.

A field discovers it cannot trust its own literature

Ioannidis (2005) made the theoretical case before the empirical one existed: given small sample sizes, flexibility in study design and analysis, financial and other interests, and a research culture that rewards novel, striking findings over careful null results, most published research findings in many fields were more likely than not to be false — not through fraud, in the typical case, but through the ordinary operation of incentives that reward being interesting over being right.

A decade later, the Open Science Collaboration (2015) supplied the empirical test, for psychology specifically. Researchers attempted to replicate 100 studies published in three leading psychology journals, using the original materials and high-powered designs wherever possible. Ninety-seven percent of the original studies had reported statistically significant results. Thirty-six percent of the replications did. Where an effect did replicate, its magnitude was, on average, roughly half the size originally reported. The finding was not that psychology's published literature was fraudulent. It was that a field's collective confidence in its own findings had been running well ahead of what those findings could actually support — and that this had happened without any single actor doing anything most researchers, at the time, would have called wrong.

What actually fixed it

The response that followed did not consist of exhortations to greater rigor or improved peer review in the abstract. It consisted of specific, adopted mechanisms, each sharing one structural feature: a public commitment, made before the outcome of a study is known, to what result would count as support for a hypothesis and what would count against it.

Chambers (2013) introduced Registered Reports at the journal ***Cortex***: a submission format in which a study's introduction, hypotheses, and analysis plan are peer reviewed and can receive in-principle acceptance for publication ***before data collection begins*** — meaning the decision to publish is made on the quality of the question and the method, not on whether the result turned out to be interesting. The format has since been adopted, in some form, by several hundred journals. Nosek, Ebersole, DeHaven, and Mellor (2018) documented the broader adoption of pre-registration across the discipline as part of what they term the "preregistration revolution," and Nosek and colleagues (2015), writing in ***Science*** as the Center for Open Science, proposed the Transparency and Openness Promotion guidelines now used by many journals to specify, at the level of editorial policy rather than individual author discretion, what data, materials, and pre-registration a published study is expected to disclose.

The common mechanism across all three is worth stating precisely, because it is the part that transfers. None of these reforms rests on trusting researchers to be more honest. Each removes, structurally, the opportunity to decide after the fact what the hypothesis was, which analyses to report, or what would count as confirmation — by requiring that commitment be made publicly, in advance, while the outcome is still unknown to everyone, including the person making the claim.

The same vulnerability, a faster mechanism

The vulnerability Ioannidis identified was structural before it was ever about any particular technology: credibility asserted or defended only after an outcome is known, by the party with the greatest interest in that outcome being believed. Generative artificial intelligence does not create this vulnerability. It lowers the cost of exploiting it, at a scale the reforms above were never built to anticipate — a separate concern this publication has examined directly, in the growth of AI-assisted paper mills producing plausible-seeming research output faster than existing review processes can evaluate it. The reform lineage that rebuilt credibility in experimental psychology is the relevant one here not because artificial intelligence resembles a psychology experiment, but because the underlying failure is the same failure: a claim protected from disconfirmation by being evaluated only in retrospect.

This is the method this Institute has adopted, and the reason it takes the form it does. The Standards of Inquiry set out in the Institute's Charter and Editorial Charter — evidence, historical seriousness, institutional realism, pluralism, revision, intellectual independence — are published once, in advance, as the standard against which any claim this Institute makes, including this essay's, may be judged; they are not assembled after a claim has already been made to justify it. The Disputation Protocol asks the same discipline of a live exchange that Registered Reports ask of a study: before the session, each side states what evidence or argument would change its position, so that the outcome cannot be redefined after the fact to declare victory regardless of what occurred. Governed Intelligence, at the institutional level this publication studies in others, asks that delegated machine execution remain observable and revisable rather than judged only by

whether its output looks right after the work is done — the same principle, applied to institutions using AI rather than to researchers studying human subjects.

Four theses

Thesis 1. Credibility in a knowledge-producing field is restored not by appeals to improved intentions or peer prestige, but by specific, adopted, prospective methodological commitments — a claim for which the psychology replication crisis and its reforms constitute direct evidence, not merely an illustrative analogy.

Thesis 2. The mechanism shared by pre-registration, Registered Reports, and this Institute's Disputation Protocol is identical in structure: a public commitment, made before an outcome is known, to what would count as confirming or defeating a claim, which removes the opportunity to redefine that standard after the fact.

Thesis 3. The vulnerability generative artificial intelligence introduces into research and publication is not new in kind — it is the same vulnerability Ioannidis (2005) identified — but is compounded in degree by the collapse in the marginal cost of producing plausible-seeming material, which is why the appropriate response is an existing, tested reform lineage rather than a novel one built from first principles.

Thesis 4. An institution that asks others to submit their claims to prospective, falsifiable commitment is obligated to have already done the same with its own; this essay treats that obligation as satisfied only to the extent that the Standards of Inquiry, the Disputation Protocol, and this publication's corrections record remain checkable by any reader, at any time, against what was actually published.

This essay's claims are subject to challenge under the terms of the Institute's Disputation Protocol. Editorial disclosure, including this publication's AI review process, appears in the closing section of this edition.

References

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at *Cortex*. *Cortex*, 49(3), 609–610.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

The Complement Becomes the Target

Four theses on why this technological moment differs from the last one

The question of urgency has already been argued in this edition on epistemic grounds: convergent AI output produces a form of influence with no historical precedent in its combination of scale, speed, and invisibility. The question of method has been argued on grounds of institutional practice: what makes a claim, including this Institute's, worth believing. This essay argues a narrower and more specific claim, one that does not depend on either of the other two: that the current moment is technologically distinct from prior waves of automation in a way that can be stated precisely, and that the distinction bears directly on the tradition this Institute draws from.

What makes an era-defining technology

Bresnahan and Trajtenberg (1995) proposed the concept of a general-purpose technology to explain why whole eras of economic growth appear to be driven by a small number of technologies — the steam engine, the electric motor, the semiconductor — rather than by innovation spread evenly across the economy. What distinguishes a general-purpose technology from an ordinary one, in their account, is pervasiveness across sectors, continued potential for improvement, and what they call innovational complementarities: the property that improvements in the technology make innovation in the many downstream activities that depend on it more productive, not merely more automated.

That artificial intelligence plausibly qualifies as a general-purpose technology in this sense is not, at this point, a controversial claim, and this essay does not attempt to establish it. The more specific and more useful question is what kind of general-purpose technology it is — which activities it substitutes for and which it complements — because the answer to that question has, for the last several decades, been remarkably stable, and it has just changed.

What the last wave actually did

Autor, Levy, and Murnane (2003) gave that stability its most precise empirical account. Studying the computerization of the American workplace between 1960 and 1998, they found that computer capital substituted for labor specifically in routine tasks — those that could be accomplished by following explicit, codifiable rules — while it complemented labor in non-routine cognitive and interpersonal tasks: problem-solving, communication, and judgment exercised under conditions too variable to reduce to a fixed procedure. Computerization did not merely fail to threaten these

capacities. It increased the market's demand for them, accounting, by the authors' estimate, for a substantial share of the shift in relative demand toward workers who could exercise them.

This finding did more than describe a labor market. It supplied, for the length of a generation, an implicit answer to the question of where human comparative advantage would continue to reside as automation advanced: not in calculation or procedure, which machines would increasingly perform, but in reasoning, judgment, and communication — the capacities a machine following explicit rules could not substitute for because they could not be reduced to explicit rules in the first place. A great deal of educational and economic policy, over the same decades, was built on some version of this assumption, often without stating it as one.

What changed

Eloundou, Manning, Mishkin, and Rock (2024), in a large-scale study of task exposure to large language models published in ***Science***, found a pattern that inverts the one Autor, Levy, and Murnane documented rather than extending it. Exposure to language-model capability, in their analysis, is concentrated not in routine, codifiable work but in higher-wage occupations characterized by exactly the non-routine cognitive and communicative tasks that the prior computerization wave had left to humans, and had made more valuable by leaving to them.

This is the precise sense in which the current moment differs from the last one, and the difference can be stated without appeal to speculation about future capability: the technology that previously served as a general-purpose complement to reasoning and communication has been followed by one whose most direct effects fall on reasoning and communication themselves. The distinction is not that machines have never touched cognitive work before — calculators, spreadsheets, and word processors did so for decades, as instruments a person used to execute an argument or a calculation they had already formed. The distinction is that a system capable of originating the argument, not merely executing one already specified, is a categorically different kind of instrument, and it is this capacity — to produce the position, not merely process it — that the ***studia humanitatis*** was built specifically to cultivate in the human being who would otherwise have to originate it alone.

Four theses

Thesis 1. Prior general-purpose technologies, including the computerization wave documented by Autor, Levy, and Murnane (2003), substituted for routine tasks and increased the relative economic value of non-routine cognitive and communicative work, establishing reasoning and communication as the durable human complement to automation for roughly four decades.

Thesis 2. Generative artificial intelligence is the first widely deployed general-purpose technology to directly target non-routine cognitive and communicative tasks, a pattern documented empirically in large-scale task-exposure research (Eloundou et al., 2024) that inverts, rather than extends, the labor-market pattern established by the prior computerization wave.

Thesis 3. This inversion is a difference in kind, not merely in degree, from previous waves of automation, because it removes the tacit assumption — embedded in several decades of labor-market adaptation and educational policy — that reasoning and communication constituted stable ground on which human comparative advantage would continue to rest as automation advanced.

Thesis 4. Because the ***studia humanitatis*** was built specifically to cultivate reasoning and rhetorical communication as capacities requiring deliberate formation, this technological inversion is not incidental to the humanist tradition's present relevance; it is the precise condition under which the tradition's founding claim — that these capacities must be cultivated, not assumed — becomes newly testable rather than merely of historical interest.

This essay's claims are subject to challenge under the terms of the Institute's Disputation Protocol. Editorial disclosure, including this publication's AI review process, appears in the closing section of this edition.

References

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. ***The Quarterly Journal of Economics***, 118(4), 1279–1333.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? ***Journal of Econometrics***, 65(1), 83–108.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. ***Science***, 384(6702), 1306–1308.

The Second Door

Four theses on artificial intelligence and the developing world's remaining path to convergence

The essays preceding this one in the edition have argued that artificial intelligence raises a question — what human beings and institutions become as machines grow more capable — that applies wherever the technology reaches. It does not reach everywhere the same way. The consequences of generative AI for a management consultancy in Boston and for a call-center operator in Manila are not the same consequences, and a framework that treats "human flourishing in the age of artificial intelligence" as a single, undifferentiated condition will miss what is, for a large share of the world's population, the more consequential question: not whether reasoning and communication are being automated, but whether the specific economic path many developing countries were counting on to close the gap with wealthier ones is still open.

The first door was already closing

For most of the twentieth century, industrialization was the standard route by which poorer countries converged with richer ones: build a manufacturing base, absorb agricultural labor into higher-productivity factory work, and grow. Rodrik (2016) documented that this route has been narrowing for several decades, independent of any AI-specific cause. Manufacturing's share of employment and output now peaks earlier and at substantially lower income levels than it did for the economies that industrialized first — a pattern Rodrik terms premature deindustrialization, driven by a combination of globalization and labor-saving technological progress in manufacturing itself. The effect has fallen unevenly: economies in Asia with strong manufactured-export sectors have been comparatively insulated, while Latin American economies in particular have been hit hard. For much of the developing world, the traditional route to convergence was already less available than it had been for the countries that walked it first.

A second door opened

Baldwin (2019) and Baldwin and Forslid (2020) describe an alternative that digital connectivity has made newly available: services-led development through what Baldwin calls telemigration — workers in lower-wage countries performing remote work for employers in higher-wage ones, without physically relocating. Baldwin and Forslid argue explicitly that this route could extend the "emerging market miracle" geographically, resembling India's services-led growth more than China's manufacturing-led growth, and reaching economies that missed the earlier industrialization window Rodrik describes. The pattern is not speculative. Horton, Kerr, and Stanton's (2017) analysis of remote-work contracts on a major online labor platform found the

largest source countries for this kind of work to be the Philippines, India, and Bangladesh — economies for which the services door, unlike the manufacturing one, has been opening rather than closing.

The same key may close it

The technology that makes telemigration possible — digital connectivity paired with software capable of performing analytical, administrative, and communicative work remotely — is not separate from the technology now capable of performing a growing share of that same work directly. Baldwin and Forslid's own account does not treat this as a settled matter in the services-led path's favor: they note directly that "white-collar robots may displace some offshore humans," meaning the same globotics transformation that opens the second door for a labor-cost arbitrage business model can, at the same time, erode the labor-cost advantage that model depends on. A telemigrant customer-service worker, translator, or data processor competes not only against workers in other countries but increasingly against a system that performs the same task at a lower marginal cost than any worker, anywhere, can offer.

This is no longer only a theoretical tension. Brynjolfsson, Chandar, and Chen (2025), using high-frequency payroll data covering millions of U.S. workers, found a 13 percent relative decline in employment for early-career workers in occupations most exposed to generative AI since its widespread adoption, concentrated specifically in roles where AI automates tasks rather than augments them — the pattern the services-led development path depends on not occurring at scale. The evidence for the developing economies this essay is actually about is thinner and less rigorous than that: the International Labour Organization has estimated that roughly one in four Philippine workers, some 12.7 million people, hold jobs in occupations that could be affected by generative AI, and industry reporting from the Philippine and Indian outsourcing sectors describes slowing entry-level hiring alongside continued aggregate growth — a mixed and still-unsettled picture, not a documented collapse. The clearest realized-effects evidence available concerns a wealthy labor market, not the developing ones whose access to the second door is actually in question. That gap in the evidence is itself worth stating plainly, since it means the strongest data currently available speaks to the risk this essay describes without yet confirming or resolving it for the population the essay is actually about.

This is not a claim that the second door is closing. It is a claim that whether it closes, stays open, or opens further is presently unresolved for the developing economies in question, that the clearest available evidence so far concerns a different population than the one at stake, and that the answer will not be determined by the technology's trajectory alone.

What determines the outcome

The dynamics described above are not only economic. Mohamed, Png, and Isaac (2020) argue that AI development and deployment can reproduce a colonial structure of power even where no colonial intention is present: the technology is designed, trained, and governed overwhelmingly by institutions in wealthy economies, while its consequences — beneficial or displacing — are absorbed by workers and economies with comparatively little voice in the decisions that produce them. On this reading, telemigration is not simply a neutral opportunity that technology happened to open. It is a labor-cost arbitrage whose terms — which tasks are profitable to offshore, which are profitable to automate instead, and on what timeline — are set almost entirely outside the economies whose workers depend on it.

This sharpens what "durable capability" has to mean in Thesis 4 below. The Institute's Capability-Conversion Problem, argued elsewhere in this edition, offers a relevant distinction here. A country whose workers perform outsourced tasks using AI tools without those tools building durable local capability — expertise, institutional capacity, the ability to move up the value chain rather than compete indefinitely on cost — occupies a fundamentally weaker position than one where the same technology use converts into lasting capacity. Read alongside Mohamed, Png, and Isaac, that capacity cannot be only technical or educational. Durable capability plausibly also requires some degree of control over the terms of the arbitrage itself — the capacity to shape AI governance, data, and labor standards, or to invest in AI development locally, rather than only absorbing decisions made elsewhere. Whether telemigration functions as a genuine development path or a temporary arbitrage that closes as the cost of automation falls is likely to depend less on which countries currently supply the most remote labor and more on which countries convert that period of participation into capability of this fuller kind — technical, institutional, and, in Mohamed, Png, and Isaac's sense, a degree of agency over the arrangement itself.

Four theses

Thesis 1. The manufacturing-led development path is closing earlier and at lower income levels for most developing economies than it did for the countries that industrialized first (Rodrik, 2016), a trend independent of and prior to any AI-specific cause, which increases the significance of alternative growth paths.

Thesis 2. Digitally enabled telemigration has been proposed, and is already empirically observed, as a services-led alternative capable of extending economic convergence to countries the manufacturing-led path left behind (Baldwin, 2019; Baldwin & Forslid, 2020; Horton, Kerr, & Stanton, 2017).

Thesis 3. The same technological capability that enables this services-led path also directly threatens to automate many of the tasks telemigrant workers currently perform; early realized-effects evidence documents this pattern in a wealthy labor market (Brynjolfsson, Chandar, & Chen, 2025), but comparably rigorous evidence for the developing economies this essay concerns remains thin, leaving the question genuinely open rather than resolved in either direction.

Thesis 4. Whether artificial intelligence accelerates or forecloses developing-economy convergence through this second path is not determined by the technology's trajectory alone, but by whether participating countries convert the resulting economic activity into durable institutional and human capability rather than a cost advantage that automation itself will eventually erase.

This essay's claims are subject to challenge under the terms of the Institute's Disputation Protocol. Editorial disclosure, including this publication's AI review process, appears in the closing section of this edition.

References

Baldwin, R. (2019). ***The Globotics Upheaval: Globalization, Robotics, and the Future of Work***. Weidenfeld & Nicolson.

Baldwin, R., & Forslid, R. (2020). Globotics and development: When manufacturing is jobless and services are tradeable. ***NBER Working Paper No. 26731***. Also published in ***World Trade Review***.

Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. ***Stanford Digital Economy Lab Working Paper***.

Horton, J. J., Kerr, W. R., & Stanton, C. (2017). Digital labor markets and global talent flows. ***NBER Working Paper No. 23398***.

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. ***Philosophy & Technology***, 33(4), 659–684.

Rodrik, D. (2016). Premature deindustrialization. ***Journal of Economic Growth***, 21(1), 1–33.

Disclosure and the Terms of Challenge

How every claim in this edition may be tested

This section applies to every essay in this edition — "Doctrine Without a Church," "Capability as the Dependent Variable," "The Complement Becomes the Target," "The Second Door," "Who Is the AI Humanist?," and "How a Claim Earns Trust" — rather than being repeated within each. It does three things: it discloses how artificial intelligence was used in preparing these essays, it indexes every disputable thesis this edition has advanced so that a challenge can cite one unambiguously, and it states, briefly, how to bring that challenge. The full rules governing what happens next are set out in the Institute's Disputation Protocol; this section does not restate them.

Disclosure: AI Critique Panel

Before publication, each essay in this edition was reviewed by a panel of large language models, given the draft and instructed to identify factual errors, weak arguments, unsupported claims, and points where an essay's characterization of the cited literature or history might be unfair to either. Their critiques informed revisions to these drafts. A human editor made the final determination, for each essay, of what to change and what to publish.

This is disclosed, not claimed as equivalent to independent peer review. Several models critiquing the same draft are not independent the way a panel of human reviewers with different training and incentives would be; they may share correlated blind spots, and a claim that survived this panel has not thereby survived scrutiny from a genuinely different vantage point. Each model's willingness to dissent is also bounded by decisions made by its provider, in ways the Institute cannot fully see or control. One essay in this edition argues that convergence among AI systems is real and measurable; readers should weigh the irony, noted here rather than hidden, that this edition's own review process draws on exactly the kind of systems that essay is about.

We are asking readers to do what the panel did, more rigorously, and with fewer constraints than any model operates under.

The Theses of This Edition

Every thesis below may be challenged individually. Cite it by the code given.

Doctrine Without a Church (WHY)

WHY-1. Convergent AI output across competing systems is a documented phenomenon, not a speculative one, evidenced in formal models of algorithmic monoculture, in analysis of shared foundation-model components, and in direct empirical measurement of output homogeneity across current systems.

WHY-2. This convergence differs in kind from earlier concerns about algorithmic curation and media concentration, because generative systems produce the position itself, in the form of independent reasoning, rather than selecting among visibly authored human content.

WHY-3. Unlike centralized doctrinal authority, this convergence has no single coordinating actor and no coercive power over the individual; it is an emergent property of shared data, shared alignment methods, and converging commercial incentive, not a deliberate policy imposed by a nameable institution.

WHY-4. The absence of a nameable authority makes this form of convergence harder to identify and resist than doctrinal control was, not because it is more coercive, but because it presents as a neutral, authorless answer rather than as an assertable position that invites a response.

WHY-5. Because the deficit here is in the individual and institutional capacity to recognize convergence and exercise independent judgment against it, the appropriate response is closer to the humanist project of cultivating that capacity than to either a purely technical fix or a regulatory one that risks recreating a single centralized authority over acceptable output.

Capability as the Dependent Variable (WHT)

WHT-1. Cognitive offloading research measures what a person retains in memory after using a tool. It does not measure, and is not designed to measure, what happens to that person's judgment, expertise, or institutional capacity.

WHT-2. Automation complacency research diagnoses a failure of individual vigilance and proposes remedies at the level of the individual and the interface. It does not, as a rule, ask whether institutional structure could make the outcome independent of any one person's sustained attention.

WHT-3. Complementarity economics measures output, quality, and wages, and treats skill largely as a fixed variable that determines who benefits. It does not, as a rule, measure whether the underlying skill itself grew, shrank, or was substituted for.

WHT-4. No literature currently in wide circulation tracks the same person or institution across individual, institutional, epistemic, and governance levels at once, longitudinally, to ask a developmental question: not ***did performance improve***, but ***what remains, and under what conditions would more remain***.

WHT-5. A research program that makes capability formation itself the central, cross-level, longitudinal dependent variable — rather than an incidental finding inside a study built to measure something else — is therefore a distinct contribution, not a restatement of existing work.

The Complement Becomes the Target (WHN)

WHN-1. Prior general-purpose technologies, including the computerization wave documented by Autor, Levy, and Murnane (2003), substituted for routine tasks and increased the relative economic value of non-routine cognitive and communicative work, establishing reasoning and communication as the durable human complement to automation for roughly four decades.

WHN-2. Generative artificial intelligence is the first widely deployed general-purpose technology to directly target non-routine cognitive and communicative tasks, a pattern documented empirically in large-scale task-exposure research (Eloundou et al., 2024) that inverts, rather than extends, the labor-market pattern established by the prior computerization wave.

WHN-3. This inversion is a difference in kind, not merely in degree, from previous waves of automation, because it removes the tacit assumption — embedded in several decades of labor-market adaptation and educational policy — that reasoning and communication constituted stable ground on which human comparative advantage would continue to rest as automation advanced.

WHN-4. Because the ***studia humanitatis*** was built specifically to cultivate reasoning and rhetorical communication as capacities requiring deliberate formation, this technological inversion is not incidental to the humanist tradition's present relevance; it is the precise condition under which the tradition's founding claim — that these capacities must be cultivated, not assumed — becomes newly testable rather than merely of historical interest.

The Second Door (WHR)

WHR-1. The manufacturing-led development path is closing earlier and at lower income levels for most developing economies than it did for the countries that industrialized first (Rodrik, 2016), a trend independent of and prior to any AI-specific cause, which increases the significance of alternative growth paths.

WHR-2. Digitally enabled telemigration has been proposed, and is already empirically observed, as a services-led alternative capable of extending economic convergence to countries the manufacturing-led path left behind (Baldwin, 2019; Baldwin & Forslid, 2020; Horton, Kerr, & Stanton, 2017).

WHR-3. The same technological capability that enables this services-led path also directly threatens to automate many of the tasks telemigrant workers currently perform; early realized-effects evidence documents this pattern in a wealthy labor market (Brynjolfsson, Chandar, & Chen, 2025), but comparably rigorous evidence for the developing economies this essay concerns remains thin, leaving the question genuinely open rather than resolved in either direction.

WHR-4. Whether artificial intelligence accelerates or forecloses developing-economy convergence through this second path is not determined by the technology's trajectory alone, but by whether participating countries convert the resulting economic activity into durable institutional and human capability rather than a cost advantage that automation itself will eventually erase.

Who Is the AI Humanist? (WHO)

WHO-1. The historical formation the ***studia humanitatis*** aimed at was explicitly civic, not merely contemplative; an account of the AI Humanist that stops at private epistemic discipline, without the civic-humanist component of institutional participation, is incomplete relative to the tradition it claims.

WHO-2. The criteria by which a layperson ordinarily calibrates trust in human expert testimony — track record, peer standing, disclosed interest — do not transfer to machine testimony in their existing form, and no adequate substitute has yet been established in either the philosophical or the practical literature.

WHO-3. Fluency in using AI tools is necessary but not sufficient for the capacity this essay describes; the distinguishing mark is retained authorship of the resulting belief or claim, not skill in producing the output.

WHO-4. The ideal described here is available to any person regardless of professional background or technical specialization, consistent with a tradition whose most consequential disputant held no advanced standing among those he proposed to argue with.

How a Claim Earns Trust (HOW)

HOW-1. Credibility in a knowledge-producing field is restored not by appeals to improved intentions or peer prestige, but by specific, adopted, prospective methodological commitments — a claim for which the psychology replication crisis and its reforms constitute direct evidence, not merely an illustrative analogy.

HOW-2. The mechanism shared by pre-registration, Registered Reports, and this Institute's Disputation Protocol is identical in structure: a public commitment, made before an outcome is known, to what would count as confirming or defeating a claim, which removes the opportunity to redefine that standard after the fact.

HOW-3. The vulnerability generative artificial intelligence introduces into research and publication is not new in kind — it is the same vulnerability documented in the pre-reform psychology literature — but is compounded in degree by the collapse in the marginal cost of producing plausible-seeming material, which is why the appropriate response is an existing, tested reform lineage rather than a novel one built from first principles.

HOW-4. An institution that asks others to submit their claims to prospective, falsifiable commitment is obligated to have already done the same with its own; this obligation is satisfied only to the extent that the Standards of Inquiry, the Disputation Protocol, and this publication's corrections record remain checkable by any reader, at any time, against what was actually published.

How to Bring a Challenge

Any person may challenge any thesis above, in writing, without credential or introduction, by citing its code (for example, "WHY-3") and stating the objection plainly: a factual error, a misreading of a cited source, a counter-example, or an argument the thesis does not survive. Submit challenges through the channel published alongside this edition. Every challenge is reviewed; the strongest are invited to a live, recorded exchange, published unedited regardless of outcome. A thesis that does not survive challenge is revised or retracted in public, with a visible record of what changed and why. Full rules governing triage, review, and live sessions are set out in the Institute's Disputation Protocol.

The goal is not to win an argument. It is to find out which of these twenty-six theses are wrong, and to say so plainly when they are.

The humanist project continues through challenge.

The humanist project continues.