The Forgotten Specification

Human Imitation, Continuation Gain, and the Geometry of Technological Progress

Flyxion
Independent Researcher

September 2026

Abstract

Daron Acemoglu's argument against the artificial general intelligence program, developed in *What Happened to Liberal Democracy?* (2026) and in a September 2026 conversation with Aaron Bastani on Novara Media's *Downstream*, is usually summarized as a plea for augmentation over replacement. This essay argues that the argument is more radical than that summary suggests. Its target is the choice of the human being as the specification against which machine progress is measured. Once that choice is made visible, a chain that appears to run from progress to imitation to equivalence to supersession loses its first link. The essay formalizes the alternative as a difference between two optimization objectives: overlap with human task space, and enlargement of the continuation set available to a human operating with a machine. It then states a compact separation, that aggregate capability gain does not imply continuation gain for participants, and connects it to Acemoglu's account of industrial organization as a social topology whose fragmentation removed adjacency as well as employment. Timothy Shenk's *Left Adrift* (2024) supplies a third instance in party politics, where the loss of unions and local organization was answered by polling and modeling: a party that could no longer reach its voters through edges began estimating them instead. A final section traces how a fictional object becomes a research target, a benchmark, a capital allocation, and finally an apparent inevitability, and defines a forgotten specification as a contingent objective that institutional commitment has reclassified from the design space into the constraint set. Shenk's history of the argument between two strategists over whether cultural polarization was fixed shows the same reclassification in electoral strategy, and suggests that the classification depends on the time horizon over which a choice is judged. A generalization, specification by organization, treats a specification as a map that enters the dynamics of what it maps, so that its later accuracy cannot establish its original necessity. Acemoglu's account of liberal democracy's promises is then read as a case of persistence of form combined with failure of continuation. The essay closes with the limits of the framework.

1 A Design Decision Presented as Destiny

The familiar version of the concern about artificial intelligence and labor is distributive. Machines will do tasks formerly done by people; the gains will accrue to owners; policy should cushion the transition or redirect investment toward tools that help workers. Acemoglu has argued all of this, with David Autor and Simon Johnson, under the banner of pro-worker AI. But the more interesting move in his conversation with Bastani, which accompanied his new book on liberal democracy and shared prosperity, lies underneath the distributive claim. Asked how pro-worker AI would be built, he pointed to nurses, electricians, teachers, and scientists, and

to giving them better tools to expand their expertise rather than attempting to mimic human talent. The obstacle, in his account, is the AGI agenda, which has persuaded nearly everyone that making machines ever more humanlike, until they surpass humans, is both the best and the inevitable use of the technology.

The book states the historical root plainly. From the 1956 Dartmouth conference onward, the stated aspiration of the field was machines with humanlike capabilities, able to perform most or all of the work people do. Acemoglu's own definition of AGI follows the same template: a single system matching the best humans across every cognitive task, from board games to legal argument to diagnosis. The human being was the target from the first day, and the target has never been seriously revisited.

The AGI framing presents a sequence as though it were a single fact:

$$\text{AI progress} \implies \text{human imitation} \implies \text{human equivalence} \implies \text{human supersession.}$$

Each later arrow is plausible given the one before it. A system that imitates humans well enough becomes equivalent on the imitated tasks, and a cheaper equivalent supersedes the original in any market that prices those tasks. The contestable arrow is the first. Nothing about computation requires that its improvement be measured by resemblance to a person. That measurement is a choice of target, and once it is recognized as a choice, the rest of the chain becomes a consequence of the choice rather than a law of technological history.

This reframing changes what kind of disagreement is available. Arguments about whether AGI is near, whether it will be safe, or whether its gains will be shared all accept the target and dispute the trajectory. Acemoglu's objection disputes the target.

2 Instruments That Do Not Resemble Their Users

Most valuable tools are profoundly unlike the people who use them. An electrician's circuit tracer does not become more useful by becoming more electrician-like; its value lies in detecting a signal through drywall that no human sense can register. A telescope does not approach the astronomer. A spreadsheet does not converge on an accountant. In each case the instrument's contribution is measured by what becomes possible for its user that was impossible before, and that contribution is often largest precisely where the instrument is least human: in sensing ranges, speeds, memory capacities, and forms of indifference that no person possesses.

Acemoglu's own illustrations of pro-worker AI have exactly this character. His leading case is the electrician, whose services he identifies as a bottleneck likely to tighten as equipment grows more sophisticated and grids require new designs. The tool he describes draws on engineering knowledge, equipment specifications, and past cases in which experienced electricians solved similar problems. It reads photographs, sensor data, and manual inputs, proposes candidate faults, and then supports an iterative process in which the electrician experiments under its guidance, with remote access to an expert when needed. Nothing in the design aims at an electrician-shaped machine. The tool enlarges what an electrician can diagnose. His classroom example is sharper still: a system that reads frequent short quizzes, identifies where students are struggling, and proposes regrouping and new material to the teacher. Where automated testing tries to reduce the need for teachers, this tool would increase it, because smaller groups need more teaching and the tasks involved become more sophisticated. Nurses, in his account, could take a larger part in diagnosis and care decisions, check drug interactions, and get guidance on when to escalate.

The book also notes an architectural consequence. Troubleshooting electrical equipment does not require a model that can compose sonnets or discourse on Greek philosophy; it requires high-quality, domain-specific records of past repairs that worked. The AGI agenda pushes investment toward large general-purpose foundation models instead, so the imitation target shapes not only what systems are asked to do but what data gets collected to build them.

This is an old observation in computing. J. C. R. Licklider's 1960 account of man-computer symbiosis and Douglas Engelbart's 1962 program for augmenting human intellect both took the enlargement of human capability as the design objective. Erik Brynjolfsson's 2022 essay on the "Turing Trap" made the economic version explicit: a benchmark defined by indistinguishability from humans steers investment toward substitution and away from complementarity, with consequences for both wages and political power. Acemoglu opens his chapter on algorithms with Licklider, and his reading is that Licklider was early rather than wrong: the computers of 1960 and the following decades could not retrieve context-specific knowledge fast enough to couple with human work, and generative models may finally be able to. What Acemoglu adds to this lineage is the insistence that the steering away from Licklider's program is not incidental. It is built into the definition of the goal.

3 Two Objectives

The distinction can be stated as a difference between optimization targets.

Definition 1 (Task overlap)**.** *Let $\mathcal{T}_H$ be the set of tasks a human agent H can perform and $\mathcal{T}_M$ the set a machine system M can perform unaided. The* imitation objective *selects*

$$M^*_{\text{imit}} = \arg\max_M \ \mu\big(\mathcal{T}_H \cap \mathcal{T}_M\big),$$

for some measure μ on task space.

Definition 2 (Continuation set and continuation gain)**.** *Let $\Gamma(H)$ denote the set of actions and futures reachable by H given current capabilities and institutions, and let $H \oplus M$ denote H operating with access to M. The* continuation gain *of M for H is the newly reachable region*

$$\Delta\Gamma_H(M) = \Gamma(H \oplus M) \setminus \Gamma(H),$$

and the augmentation objective *selects*

$$M^*_{\text{aug}} = \arg\max_M \ \mu\big(\Delta\Gamma_H(M)\big).$$

The two objectives can be pursued with the same underlying methods and described with the same word, yet they reward different technologies. The imitation objective is maximized by a system that can do what H already does, which makes H optional. The augmentation objective is indifferent to overlap and is maximized by a system that opens regions of action H could not previously reach. A system can score near zero on the first and very high on the second. The most valuable instrument under the augmentation objective might resemble a person in no respect at all.

A resemblance metric of the form $S(M) = \text{sim}(M, H)$ therefore measures the wrong object. It records how far the machine has travelled toward the human, not how far the human can now travel.

4 Capability Gain Is Not Continuation Gain

The sharpest consequence of the formalization concerns the relation between aggregate and individual measures.

Definition 3 (System capability)**.** *Let C_{sys} be the total set of tasks performable within an economy by any composition of its agents and machines.*

Proposition 1 (Separation)**.** *It is possible that $\Delta C_{\text{sys}} > 0$ while $\mu\big(\Gamma(H_w)\big)$ decreases for a large class of workers H_w.*

Sketch. Suppose a system M, owned by a firm F, comes to perform a set of tasks $A \subset \mathcal{T}_{H_w}$ at lower cost. System capability does not fall, since A is still performed, and it rises if M also performs tasks outside the prior union. But the portion of $\Gamma(H_w)$ that consisted of selling A on a labor market is removed, because the market no longer admits H_w as a supplier of A. Unless M is also accessible to H_w in a way that opens a compensating region $\Delta\Gamma_{H_w}(M)$, the worker's continuation set contracts while the system's capability expands. □

There is no paradox in the separation. The two quantities concern different objects. The first measures what the whole can do. The second measures what a participant can do, including what the participant can offer, refuse, bargain over, and build on. A society can become far more capable in aggregate while the admissible futures of many of its members narrow.

The separation also explains why the usual compensating proposals miss the target. Broad ownership of AI tools or a universal basic income could in principle restore consumption to displaced workers. Acemoglu doubts that the firms capturing the gains would accept such redistribution, but his deeper objection is that it would not restore what was lost. A society in which a large share of people do not contribute to production produces status gaps between makers and takers, and communities, in his account, are most vibrant when their members feel they contribute to society at large.

Remark 1 (Transfers restore consumption, not continuation)**.** *A transfer τ can hold a displaced worker's consumption constant while leaving $\Gamma(H_w)$ contracted, because the removed region consisted of actions (offering a skill, refusing a bargain, building on expertise, contributing to a shared undertaking) and not of goods. Compensation measured in consumption can be complete while compensation measured in continuation is zero.*

The operative variable is therefore not automation as such but *where capability resides* after automation. Expertise that was distributed across millions of workers, and available to each of them as a basis for bargaining and for independent action, can be captured in systems owned by a small number of firms. The expertise still exists. It has moved.

Acemoglu frames this as centralization. The early promise of digital technology was decentralization; its enthusiasts regarded the IBM of the 1960s as a bureaucratic monolith that personal computing would dissolve. IBM's dominance did fade, but it was succeeded by Microsoft and Oracle, then by Google and Facebook, each more powerful than the last. AI developed along the imitation objective continues that succession rather than reversing it.

The book gives this a Hayekian form that sharpens the separation into a dynamic claim. Hayek argued in 1945 that central planning fails because the knowledge an economy runs on is dispersed, partial, and often contradictory, held by the people closest to particular circumstances. A naive reading suggests that sufficiently powerful computation could now gather what planners

could not. Acemoglu's reply is that the relevant knowledge is not a stock waiting to be collected. It is produced continually by human experimentation, and much of it is tacit, social, and tied to situations that keep changing. He illustrates the gap with a language model that can explain how to repair a broken garage door but has no grasp of the social setting of the repair, and treats it as no different from summarizing an ancient text. He also points to early evidence that centralization is already eroding the process it depends on: people post fewer questions and answers on discussion forums once a chatbot supplies answers, and Wikipedia views have fallen on topics the chatbot covers.

Remark 2 (Absorption erodes its own source)**.** *Let the flow of new dispersed knowledge be generated by the experimentation contained in the continuation sets of participants,* $\dot{K} = f\big(\{\Gamma(H_i)\}\big)$*, with* f *increasing. A system trained on* K *and optimized under the imitation objective contracts the* $\Gamma(H_i)$ *whose tasks it absorbs, and so lowers* $\dot{K}$*. The imitation objective is therefore not only distributively costly but epistemically self-limiting: it consumes the experimentation that produces the knowledge it absorbs. The augmentation objective, by enlarging* $\Gamma(H_i)$*, raises* $\dot{K}$*.*

5 The Factory as Topology

Acemoglu's account of industrial history supplies a second instance of the same separation. The mid-century factory did not merely pay wages. A workplace on the scale of General Motors placed mechanical engineers, workers on the factory floor, accountants, and personnel staff in overlapping institutional space, a multiclass arrangement of people with different educations, aspirations, and ideological commitments. Its supply chain reached further still: in the auto towns around Flint, a large share of the population worked in auto-related activity, and those towns supported high wages alongside a real sense of community before they were shut down. The book's emblem is Lancaster, Ohio, which *Forbes* celebrated in 1947 as the all-American town. Its glassmakers provided stable jobs, and workers and managers lived in the same neighborhoods, attended the same churches, and sent their children to the same schools; the spouses of glassworkers organized vaccination drives and pressed for better sidewalks. After a raid, a restructuring, and a conglomerate purchase in the 1980s, office and factory workers were laid off, machinery replaced experienced workers, and the town's cohesion went with the jobs. The same pattern appears in the aggregate. In the 1960s about a fifth of a typical college graduate's neighbors held a degree; today nearly half do. The chance that a college graduate marries another rose from roughly a quarter in 1970 to over two fifths. The classic manufacturing firm had mixed manual workers with technicians, engineers, accountants, and managers; surviving firms are far more education-heavy, and sectors such as retail and fast food are staffed almost entirely by workers without degrees.

Acemoglu adds a pointed consequence: leaders of center-left parties who have never worked in a firm with a labor movement find labor solidarity hard to understand, because they never occupied the topology that produced it.

Represent this as a graph $G = (V, E)$ whose vertices are workers and whose edges are relations of routine contact, mutual dependence, and shared stake. Acemoglu's description of community as the conduit of information, trust, and mutual understanding is, in these terms, a description of what edges carry. Solidarity, on this reading, is not a sentiment but a structural property: the degree to which the graph is connected across occupational and class boundaries, so that a grievance in one region can propagate to others and a bargaining position can be held collectively.

Deindustrialization removed vertices, but more consequentially it removed edges. Outsourcing split production across firms and countries. Occupational specialization partitioned the remaining workforce into clusters with few cross-links. Geographic sorting separated those clusters physically. Each process is locally rational and each lowers the graph's cross-boundary connectivity. Online community, whatever its other merits, has generally failed to replace the lost edges, because shared stake in a common workplace is not reconstructed by shared interest in a common topic.

The loss is therefore larger than the count of jobs. It is a loss of adjacency, and adjacency is part of what made individual continuation sets collectively defensible.

6 The Party as Topology

Timothy Shenk's *Left Adrift* tells the political half of this story from inside the campaign room. His opening charts compare the Democratic coalition of 1948 with that of 2016, divided by education. In 1948 support fell steadily as education rose, a diagonal line. By 2016 it formed a U, high among voters without a high-school diploma, lower in the middle, and highest among the most educated: a fishhook. The same reshaping appears, with local variation, in France, Italy, Denmark, and well beyond Europe. In Shenk's summary, the left was once a bottom-up coalition in which organizations representing workers set the tone, and it now owes more to universities than to unions.

The organizational change is the part that matters here. Shenk describes state and local party machines atrophying and being replaced by a permanent political class concentrated in Washington, funded by a small number of large donors and employed across an overlapping network of foundations, think tanks, and advocacy groups. At the same time the decline of organized labor removed the institution most responsible for linking blue-collar voters to the Democratic Party. In graph terms, the edges that ran from workers through union halls and ward organizations to the party were cut, and the party's remaining edges ran mostly to donors, professionals, and staff.

Shenk's account of Greenberg's early scholarship explains why the edges mattered. Greenberg's dissertation studied poor communities in Appalachia, San Jose, Atlanta, Detroit, and Philadelphia and asked why shared economic hardship failed to produce a common class consciousness. He found his answer in Gramsci, who treated politics as something other than a mechanical reflection of economic interest: workers would have to forge a collective consciousness out of whatever cultural materials were at hand. Stated compactly,

$$\text{shared condition} \not\Rightarrow \text{shared political identity}.$$

Something has to perform the mapping from the first to the second, and in the twentieth century that something was organization: the union, the local party, the workplace that made dispersed people mutually legible as participants in the same undertaking. The causal sequence is therefore not class structure producing class politics directly but material difference passing through organization, adjacency, and shared interpretation before it becomes a political relation.

What replaced the edges was measurement. The book follows two consultants, Stanley Greenberg and Douglas Schoen, who entered politics as academics studying the breakup of the New Deal coalition and became global strategists. Both held that democracy translates the views of ordinary people into policy and that those views can be discovered through polls and

focus groups. Around them, campaigns turned to big data and quantitative models in the hope of making electioneering a science.

The substitution has a precise structure. A party connected to its voters by organizational edges can alter what those voters are able to do together: strike, bargain, turn out, press a demand across a region. It acts on their continuation sets. A party that has lost those edges can still estimate what its voters currently think and fit its messages to the estimate. It acts on a model of the electorate. The first is analogous to the augmentation objective, since it enlarges what participants can reach. The second is analogous to the imitation objective, since it optimizes the fit between the party's output and a representation of the people it no longer reaches directly. Polling is not the villain of this account, and Shenk defends the craft at its best, but a party that relates to its voters mainly through their representation in a survey has the same relation to them that a humanlike model has to the worker it is trained on.

7 One Transformation, Three Times

Read together, the three accounts describe the same movement in three domains. In political economy,

$$\text{shared production} \longrightarrow \text{distributed capability}$$

has increasingly given way to

$$\text{centralized production} \longrightarrow \text{concentrated capability.}$$

In artificial intelligence,

$$\text{tool} \longrightarrow \text{expanded human capability}$$

is being displaced by

$$\text{model} \longrightarrow \text{synthetic human} \longrightarrow \text{replacement.}$$

In party politics,

$$\text{organization} \longrightarrow \text{collective capability of members}$$

has given way to

$$\text{estimation} \longrightarrow \text{a model of members} \longrightarrow \text{messages fitted to the model.}$$

In each, the question of how much capability exists is less decisive than the question of where it lives and who can reach it. The imitation objective, the fragmented workplace, and the consultant-managed party converge on the same outcome: capability accumulates in places that individual participants cannot access, bargain with, or organize around.

8 From Fiction to Infrastructure

Acemoglu's closing observation, that humanlike machine intelligence was first a science-fiction object, is more consequential than it first appears. Asked why so many have accepted the AGI agenda, he gave two answers. One is industrial history: the largest technology firms grew by selling automation tools to companies, so the direction was set before the current models existed. The other is cultural: the field has been shaped since its inception by science fiction, whose staple

is the humanlike machine that takes over, and by imagining that machine for seventy years the field has been willing it closer to reality. The book extends both answers. On the economic side, the established revenue models of the leading firms, digital advertising and automation software sold to companies, give no natural push toward new worker tasks, whose development paths are unpredictable and whose gains often accrue to firms other than the one that invested. Tax codes in many industrialized countries subsidize capital spending relative to labor, adding a further bias toward automation. On the ideological side, the field's metrics of success center on what the industry calls "reaching human parity," and its highest honors go to models that act and sound human. Acemoglu argues that belief in AGI can be self-fulfilling in a specific sense: even if AGI remains out of reach, its pursuit continues as ever more sophisticated-looking systems are marketed and adopted for automation, and the alternative directions go unexplored. Concentration closes the loop. In a sector where start-ups face a choice between being crushed and being acquired by incumbents, they converge on the incumbents' model, and the diversity of experimentation that might have broken the pattern is dulled. Together these suggest a path-dependence argument with a recognizable sequence:

fiction → research object → benchmark → capital allocation → infrastructure → apparent inevitability.

At each stage the target acquires more institutional weight. Benchmarks define progress in its terms. Investment follows the benchmarks. Data centers, energy contracts, hiring pipelines, and academic prestige are organized around the target. By the final stage, the scale of the commitments is itself offered as evidence that the destination was natural.

The causal direction is at least partly reversed. The destination looks inevitable because resources were organized around making it the destination. This is a pattern that recurs wherever infrastructure accumulates around a contingent choice: after enough capital has been sunk, the choice begins to present itself as an external constraint, and alternatives look not merely different but unrealistic.

The compact statement is that AGI may be less a prediction about where computation is going than a *specification that has forgotten it was a specification.*

The phrase can be made precise. Distinguish two sets that any account of a technical trajectory implicitly uses: a *design space* $\mathcal{D}$ of objectives regarded as open to choice, and a *constraint set* $\mathcal{K}$ of conditions regarded as given, whether by physics, by logic, or by fixed features of the world.

Definition 4 (Forgotten specification)**.** *Let O be an objective that was originally selected from among alternatives, so that $O \in \mathcal{D}$ at the time of its adoption. Let $I(O)$ denote the accumulated institutional commitment to O: sunk capital, benchmarks, contracts, trained personnel, prestige, and narrative. O is a* forgotten specification *when, after sufficient $I(O)$, discourse about the trajectory places O in $\mathcal{K}$ rather than $\mathcal{D}$,*

$$O \in \mathcal{D} \quad \xrightarrow{I(O)} \quad O \in \mathcal{K},$$

while the system continues to optimize O as before.

Nothing about the optimization changes at the moment of forgetting. What changes is the classification. A variable that could be set differently is now described as a boundary condition, and proposals to set it differently are heard as proposals to violate a constraint. They sound unrealistic rather than merely unpopular.

Remark 3 (Cost is not impossibility)**.** *The reclassification is not pure illusion. The commitments that make up $I(O)$ are real, and so are the costs of redirecting them. The error lies in converting a cost into an impossibility. A genuine constraint in $\mathcal{K}$ cannot be relaxed at any price; a forgotten specification can be relaxed at a price that is high, finite, and falls on identifiable parties. A diagnostic follows: O is a specification rather than a constraint whenever some alternative objective O' is technically feasible and is blocked only by the reversal cost of $I(O)$. For AGI, the augmentation objective of Section 3 is such an O'. It uses the same methods, and what stands between it and the current trajectory is commitment, not physics.*

The mechanism is general. It describes how any chosen destination, once enough has been built around it, comes to be perceived as the shape of the road. On this reading, the title names a misclassification rather than a lapse of memory: the forgotten specification is a design variable that has been filed among the constraints.

Concentration is the mechanism that keeps the file closed. When every new entrant must either adopt the incumbent objective or be absorbed, $I(O)$ grows even from actors who might have chosen otherwise, and the apparent absence of alternatives becomes an artifact of market structure rather than evidence about what computation can do.

The political instance

Shenk's history supplies a second case of the same reclassification, in electoral strategy rather than engineering. The argument between Greenberg and Schoen was, at bottom, an argument about whether cultural polarization belonged to $\mathcal{D}$ or $\mathcal{K}$. Schoen read the electorate as permanently tilted toward Republicans by cultural division and advised Democrats to follow voters toward the center on both economics and culture. Greenberg held that a mix of economic populism and respect for the cultural concerns of blue-collar voters could split the electorate along economic lines again and rebuild a multiracial working-class majority. For Schoen, polarization was a constraint to be managed. For Greenberg, it was a variable to be reset.

Both men treated public opinion as real and discoverable; they disagreed about what follows from measuring it. Let P_t be the state of opinion, M_t its measurement, and $\mathcal{A}_t$ the space of politically admissible positions. Schoen's practice approximates a *measurement regime*, in which measurement informs a campaign and the space it searches is treated as given:

$$P_t \to M_t \to C_t, \qquad \mathcal{A}_{t+1} \approx \mathcal{A}_t.$$

Greenberg's ambition was a *constructive regime*, in which measurement informs organization and organization changes the space itself:

$$P_t \to M_t \to O_t \to P_{t+1}, \qquad \mathcal{A}_{t+1} = F(\mathcal{A}_t, O_t).$$

Neither regime assumes voters are infinitely malleable. The disagreement concerns how plastic the admissible space is, and over what period.

Shenk's own framing sides with the second classification at the level of principle. Social divisions, he writes in effect, are constant; which of them get politicized, and how, is open to choice, and in a democracy no choice matters more. Greenberg's later reflections describe the reclassification from the inside. After campaign after campaign won on populist platforms, he watched his clients meet with the managers of the economy and turn toward austerity, and he came to believe that forces not always visible worked to rule out messages about class and

fairness even when those messages tested well in his surveys. Shenk's verdict on the consultants who told leaders that ideology and class conflict were finished is that, while claiming realism, they were "making politics smaller than it had to be." That is a description of a forgotten specification at work: an electoral strategy chosen for its fit to present opinion, filed among the facts about the electorate, and then offered as the reason no other strategy was realistic.

Remark 4 (Classification depends on horizon)**.** *Shenk also records the strongest argument for treating public opinion as a constraint. He holds that voters' views can be known, that they rarely move much during a single campaign, and that candidates who ignore the limits set by the public are quickly punished. All three claims can be true together with Greenberg's. Index the constraint set by the horizon over which a choice is evaluated, $\mathcal{K}_\tau$. Over one election, the distribution of opinion is close to fixed and belongs in $\mathcal{K}_\tau$ for small τ. Over decades, which divisions are politicized is shaped by organizations, parties, and movements, and it belongs in $\mathcal{D}$. Shenk's own account of realignment makes the same distinction: a campaign can shape how a race is seen, but durable change requires a movement that alters how an entire party is perceived, delivered over many cycles. A forgotten specification can then be restated as a horizon error: a short-horizon constraint applied to a long-horizon decision.*

The same restatement fits the technical case. Over the next funding cycle, the frontier-model trajectory is effectively a constraint on any single firm. Over the span in which a society decides what its technology is for, it is a choice.

9 Specification by Organization

The two cases share a structure that is worth stating once in general form. A technological field selects which capability counts as progress; a political system selects which distinction counts as politically meaningful. Neither merely discovers the object it measures.

Definition 5 (Specification by organization)**.** *Let X_t be a heterogeneous domain (a population, a field of possible technologies) and $S_t : X_t \to Y_t$ a specification that renders it actionable, such as a benchmark, a survey category, or an administrative classification. Action A_t is allocated through S_t, and the domain evolves under that action:*

$$S_t(X_t) \to A_t \to X_{t+1} = F(X_t, A_t) \to S_{t+1}(X_{t+1}).$$

The specification is organizational *when F depends materially on A_t, so that the map enters the dynamics of what it maps.*

Acemoglu supplies the political theory of the first step. His critique of social engineering draws on James C. Scott's account of legibility: control from above requires simplifying a varied population into uniform, administratively readable categories, and uniform rules and classifications in turn reinforce the demand for uniformity. Written as a projection $\pi : X \to L$ from a social state to its legible representation, the difficulty is not only that $\pi(x) \neq x$ but that the admissible continuations available to the people represented change once governance acts on $\pi(x)$ rather than on x. The humanlike benchmark, the poll-fitted electorate, and the administrative category are three instances of the same move. Each is a legible projection, and each becomes a site of intervention.

Two different pathologies can then follow, and they are easily confused. Let G be the underlying phenomenon and S its institutional signal, initially informative, so that $I(G; S) > 0$.

- *Decoupling.* Rewards are made conditional on S, agents optimize the signal, and S rises while $I(G;S)$ falls. The signal becomes false. This is the familiar pattern associated with Goodhart's law.
- *Recursive construction.* Optimization around S changes G itself, $S_t \to A_t \to G_{t+1}$, so that S and G come to correspond because the world was rebuilt around the signal. The signal becomes true because it was enforced. This is closer to what Ian Hacking called the looping effects of human kinds and to what Donald MacKenzie described as the performativity of economic models.

Proposition 2 (Accuracy does not establish necessity)**.** *Two histories, $G \to S$ (the specification tracked a prior structure) and $S \to G$ (the structure was produced by acting on the specification), can yield the same present correspondence $S \approx G$. Observation of the present surface cannot distinguish them. Distinguishing them requires intervention on the mapping: perturbation, counterfactual comparison, or institutional genealogy.*

This is the deepest reason a specification can be forgotten. By the time anyone asks whether the target was a choice, the world may have been reorganized enough that the target fits it well, and the fit is offered as proof that it was never a choice. A distinction is selected, institutions organize around it, resources follow it, people adapt to it, infrastructure preserves it, and success becomes measurable through it. At the start it was one possible specification. Later it looks like a fact about the world.

Shenk's work imposes an important limit on this line of argument. His democratic commitments depend on public opinion having a reality independent of its measurement. Voters have jobs, rents, families, health-care costs, and memories, and the concerns that Greenberg and Schoen kept encountering (jobs, education, housing, health care) were not artifacts of their surveys. Organization can change which of these concerns becomes a political identity; it does not conjure the concerns from nothing. The same holds for the technical case. A benchmark can shape AI development without making computational capability fictitious. The accurate formulation is

$$\text{constructed legibility} \neq \text{constructed reality}.$$

The error being diagnosed is not that measurement fabricates. It is that measurement becomes intervention, and the intervention is then forgotten.

10 Persistence of Form, Failure of Continuation

Acemoglu's larger book places the argument inside a theory of institutional decay. He defines liberal democracy's practical promises as three: sharing, meaning that growth yields broadly distributed gains; drains, his deliberately mundane shorthand for competent public services; and voice, meaning real participation in self-government rather than formal rights alone. For much of the industrial period the mapping from liberal principles to these outcomes worked, not automatically but because labor demand, worker organization, and democratic regulation made growth distributive.

Then the environment changed. Automation, globalization, and the shift toward services altered the production function, and the institutional configuration did not adapt with it. Acemoglu calls the resulting failures sins of omission and commission. The omission was the failure to redesign institutions for a postindustrial economy whose costs fell heavily on the less educated.

The commission, in his account, was that parts of the educated postindustrial elite gave up commitments to community, self-government, and shared prosperity in favor of top-down cultural transformation. He also refuses a two-class picture, distinguishing highly rewarded elite graduates from a much larger group he calls the mass college educated, whose credentials promised a status the postindustrial economy could not deliver to most of them. Education becomes a sorting boundary as well as a ladder.

In the terms of this essay, the crisis has a definite shape. The observable institutional surface persists: elections are held, rights are enumerated, agencies and universities operate. What decays is the mapping from that surface to the goods that gave it meaning. If S is the democratic surface and G the promised outcomes, then S persists and G remains the declared objective while $I(S;G)$ falls. Distrust, on this reading, need not signal hostility to democratic ideals. It can be the rational response to a surface that no longer predicts what it promises. The compact statement, which is this essay's abstraction rather than Acemoglu's wording, is

$$\text{liberal-democratic crisis} = \text{persistence of form} + \text{failure of adaptive continuation}.$$

Shenk's conclusion reaches the same distinction from the side of strategy. A candidate can change how one race is perceived, but realignment requires a movement that changes how an entire party is perceived, sustained across many offices and delivered through policy over many cycles. A campaign victory is a state. A coalition is a continuation structure, and the question is not whether it occupies a winning position now but whether it keeps producing viable successors.

Acemoglu's four pillars of liberalism (experimentation, nondomination and self-government, economic growth, and soft communitarianism) can be read, without strain, as conditions on continuation sets. Experimentation enlarges what is reachable. Nondomination prevents any actor from arbitrarily collapsing another's reachable set. Self-government gives people standing in changes to their own admissible space. Growth supplies resources for continuation, subject to the separation proposition, since aggregate gains need not reach individual sets. Soft communitarianism acknowledges that many capabilities are relational: what a person can reach depends on the associations, workplaces, and neighborhoods around them, which is the topology of Section 5 in another vocabulary. That is why the book treats community as a natural nexus of self-government and experimentation, and notes that labor unions themselves depend on solidarity rooted in community relations: the edge structure comes first, and the bargaining agent is built on it.

11 What Pro-Worker AI Means When Formalized

Against this background, the phrase "pro-worker AI" carries more theoretical content than its slogan form suggests. It does not mean AI that avoids layoffs, which would be a constraint on deployment. It means technology whose primary optimization object is the expansion of another agent's admissible action space rather than the absorption of that agent's functions: the augmentation objective of Section 3 rather than the imitation objective.

Two further conditions follow from the separation proposition. First, the continuation gain must accrue to the worker and not only to the owner of the system, which is a question of access, ownership, and interface rather than of model architecture. Second, the gain must be defensible, which returns the argument to topology: individual continuation sets are more durable when the graph connecting participants retains enough cross-boundary edges to hold them collectively.

Neither condition requires opposition to AI. Both require a change in what counts as its success.

The two books converge here. Acemoglu's political program, which he calls working-class liberalism, joins liberal values to working-class priorities: shared prosperity, jobs, and public services, rooted in community and wary of top-down social engineering. Shenk ends on the claim that realignment requires a movement that raises living standards in the short run while using government to build worker power over the long run. Stated in the terms of this essay, both call for policies that enlarge participants' continuation sets and restore the edges through which those sets are defended, in technology and in politics alike.

12 Limits

Substitution has historically raised living standards. The mechanization of agriculture replaced most farm labor and was, over the long run, among the largest welfare gains in history. Substitution can enlarge the continuation sets of consumers even while contracting those of displaced producers, and a complete account must net these against each other across time rather than treating displacement as the whole ledger.

The distinction lives in deployment as much as in design. The same system can function as substitute or complement depending on who holds it, how it is priced, and what institutions surround it. The imitation and augmentation objectives describe targets, but outcomes also depend on ownership and bargaining power, which no training objective fixes.

Human benchmarks have honest uses. The built world is designed for human bodies and human interfaces, and human-level performance on a task is a convenient, legible evaluation point. The objection is not to measuring against humans but to treating the measurement as the destination.

Edges can bind as well as connect. Acemoglu is explicit that community is a natural seat of self-government and information sharing but can also become a cage that limits its members' freedom. A denser graph is not automatically better. The claim here is narrower: cross-boundary edges carry information and bargaining capacity, and their loss is a real cost, not that every tie is a good one.

The argument does not depend on AGI arriving. Acemoglu doubts that AGI is near and expects the medium-term productivity gains from AI-based automation to be real but modest. The forgotten-specification argument holds either way. If AGI is distant, the pursuit still steers investment and data collection toward substitution; if it arrives, the separation proposition applies at full force.

Reclassification is not conspiracy. Shenk warns against the caricature in which a neoliberal elite simply abandoned workers. The transformation he traces was a contested battle, pushed forward at decisive moments by politicians trying to hold workers and suburban professionals at once and then failing to deliver. A forgotten specification needs no author. It arises from many locally reasonable commitments that together close off a choice no one decided to close.

Continuation sets are hard to measure. Task overlap is comparatively easy to benchmark; newly reachable action space is diffuse, delayed, and partly unforeseeable, since many of its regions consist of tasks that do not yet exist. The Acemoglu and Restrepo emphasis on the creation of new tasks points at exactly this region. Its difficulty of measurement is itself part of why the imitation target wins by default: it is the one that fits on a leaderboard.

13 Conclusion

The usual debate about artificial general intelligence asks whether it will arrive, whether it will be safe, and whether its gains will be shared. Each question accepts the target. The prior question is why human replacement was selected as the geometry of progress in the first place, whether that selection still sits in the space of things a society believes it can choose, and whether a society that measured its machines by what they made reachable for people, rather than by how closely they resembled people, would have built different machines.

The answer the formalization suggests is that it would, and that the difference is not one of sentiment. An economy can grow more capable while its members grow less free to act, and the measure that hides this is the one that asks how human the machine has become. A party can grow more precise about its voters while losing the power to act with them, and the measure that hides that is the one that asks how well its messages fit the poll.

References

[1] D. Acemoglu, interviewed by A. Bastani, "Nobel Economist Warns: AI Will Destroy Democracy," *Downstream*, Novara Media, 20 September 2026. `https://www.youtube.com/watch?v=aNYDMfdD6pg`

[2] D. Acemoglu, *What Happened to Liberal Democracy? Remaking a Politics of Shared Prosperity*. London: Profile Books; New York: Dutton, 2026.

[3] T. Shenk, *Left Adrift: What Happened to Liberal Politics*. New York: Columbia Global Reports, 2024.

[4] F. A. Hayek, "The Use of Knowledge in Society," *American Economic Review* 35(4), 1945, 519–530.

[5] J. C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press, 1998.

[6] C. A. E. Goodhart, "Problems of Monetary Management: The U.K. Experience," *Papers in Monetary Economics*, Reserve Bank of Australia, 1975.

[7] I. Hacking, "The Looping Effects of Human Kinds," in D. Sperber, D. Premack, and A. J. Premack (eds.), *Causal Cognition*. Oxford: Clarendon Press, 1995.

[8] D. MacKenzie, *An Engine, Not a Camera: How Financial Models Shape Markets*. Cambridge, MA: MIT Press, 2006.

[9] D. Acemoglu, D. Autor, and S. Johnson, essay on building pro-worker AI, 2026; discussed in J. Cassidy, "Can A.I. Be Pro-Worker?", *The New Yorker*, March 2026.

[10] D. Acemoglu and S. Johnson, *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. New York: PublicAffairs, 2023.

[11] D. Acemoglu and P. Restrepo, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives* 33(2), 2019, 3–30.

[12] E. Brynjolfsson, "The Turing Trap: The Promise and Peril of Human-Like Artificial Intelligence," *Dædalus* 151(2), 2022, 272–287.

[13] J. C. R. Licklider, "Man-Computer Symbiosis," *IRE Transactions on Human Factors in Electronics* HFE-1, 1960, 4–11.

[14] D. C. Engelbart, *Augmenting Human Intellect: A Conceptual Framework*. Stanford Research Institute, 1962.