Introduction for "The Economics of Transformative AI"

By Ajay Agrawal, Erik Brynjolfsson, and Anton Korinek

Introduction[1]

We stand at the threshold of perhaps the most consequential economic transformation since the Industrial Revolution. Leading artificial intelligence (AI) laboratories and researchers are converging on a remarkable prediction: artificial intelligence systems that match or exceed human cognitive capabilities across virtually all domains may emerge within the coming decade. The leaders at frontier AI companies speak not in terms of decades but years—with some suggesting that AI matching human intelligence could arrive in less than five years. While such predictions carry inherent uncertainty, the recent acceleration in AI capabilities suggests we are approaching an inflection point in human history. The transition from an economy where human intelligence is the bottleneck factor to one where intelligence becomes easily reproducible and abundant represents perhaps the most significant economic transformation that humanity has experienced thus far. This possibility—no longer confined to science fiction but actively discussed in corporate boardrooms and policy circles—demands rigorous economic analysis to help society navigate the profound changes ahead.

This volume presents a set of papers and their discussants' comments presented at the NBER Conference on the Economics of Transformative AI at Stanford in September 2025. The assignment given to the authors in this volume was deliberately ambitious and provocative. Erik opened the Conference with our charge to the participants:

> Ajay, Anton and I believe that economists need to radically increase our attention to AI. If the technologists achieve what they're setting out to achieve, or anything close to it, there's no way the economy is going to be the same. Existing AI is amazing enough but … we're going to ask you to suspend disbelief and think about the kind of AI that may come into the world in the next few years. We're calling it *Transformative AI*. One description is Dario Amodei's (2024) "country of geniuses in a data center"[2]. We've asked you to start with the assumption that a

[1] This volume and the associated NBER conference are supported through the NBER Digital Economics and Artificial Intelligence Initiative, originally founded by Shane Greenstein with support from the Alfred P. Sloan Foundation. We thank Shane for his leadership in helping to build this community. Today, the initiative is supported by the Sloan Foundation and Microsoft and is led by Catherine Tucker and Avi Goldfarb. We are grateful for organizational support by Christie Ko and Susan Young at the Stanford Digital Economy Lab and Editorial work by Helena Fitz-Patrick at the NBER.

[2] That is, millions of systems that perform at or beyond the level of the world's brightest minds in virtually all cognitive domains, each working 10 to 100 times faster than any human.

> technology like that exists and then think through what that means for the particular area you're diving into.

While we appreciate that there is significant uncertainty about both the timeline and feasibility of such advances, we believe it is both useful and essential to rigorously analyze this non-zero probability scenario given its potentially transformative implications. We believe the result is a forward-looking volume that helps us prepare for a radically different future, providing the analytical frameworks necessary to navigate what could be humanity's most significant economic transformation.

Our volume builds upon and extends the foundation laid by the 2019 NBER volume "The Economics of Artificial Intelligence: An Agenda" co-edited by Agrawal, Gans, and Goldfarb (2019), which provided economists with crucial frameworks for understanding AI and established foundational insights about how machine learning would affect firms, markets, and economic growth. That volume emerged when AI's commercial applications were just beginning to demonstrate their potential, focusing primarily on narrow AI systems that excelled at specific tasks. In the brief period since, we have witnessed the emergence of general-purpose large language models that can engage in sophisticated reasoning and AI agents that can both augment and substitute for human capabilities across many professional domains. The questions have evolved from whether AI would have economic impact to how we might manage a transformation where AI systems approach and potentially exceed human cognitive abilities across all domains. This rapid evolution—from specialized prediction machines to potentially general-purpose cognitive systems—necessitates a renewed research agenda that grapples with the possibility of "transformative" AI.

At the start of our 2025 conference, Ajay recited excerpts from the presentation delivered by the late Daniel Kahneman at the inaugural NBER 2017 Economics of AI conference (included as the final chapter in the 2019 NBER volume). Kahneman offered observations that now appear remarkably prescient about AI's transformative potential. He argued that there are no fundamental limits to what artificial intelligence might achieve: "We have in our heads a wonderful computer…It is made of meat, but it's a computer. It's extremely noisy, but it does parallel processing. It is extraordinarily efficient, but there is no magic there." This fundamental insight—that human intelligence operates through physical processes that can, in principle, be replicated and improved upon—underlies the transformative potential of AI. Kahneman emphasized AI's crucial advantages: "One of the major limitations on human performance is not bias, it is just noise." AI systems can achieve much higher consistency. Moreover, he noted that "robots will be able to predict emotions and development in emotions far better than people can" and ultimately "will be

endowed with broad framing" that gives them the wisdom humans struggle to achieve. His conclusion was unequivocal: "I do not think that there is very much that we can do that computers will not eventually be programmed to do." These observations, grounded in understanding intelligence as computation rather than magic, make a compelling case for why AI may become transformative—not merely augmenting human capabilities but potentially exceeding them across all cognitive domains.

This volume presents sixteen creative and careful analyses organized in a logical sequence that builds from fundamental theoretical foundations through practical applications to long-term societal implications. The journey begins with Part I establishing the economic foundations of transformative AI, starting with modeling AI as "genius on demand" and examining how it transforms innovation while challenging our measurement frameworks. Part II explores how markets and organizations restructure when intelligence becomes commodified, revealing both concentration risks and new organizational possibilities. Part III grounds these abstractions in human reality, examining which workers face displacement, how labor markets evolve, and where humans find meaning beyond economic productivity. Part IV investigates systemic transformations in knowledge creation and information ecosystems, including risks of centralized decision-making and informational collapse. Finally, Part V provides concrete policy frameworks (but, in accordance with NBER guidelines, no policy recommendations), from redesigning public finance for an AI economy to quantifying appropriate investments in AI safety. This progression—from micro-level production functions through market dynamics to macro-level governance—provides readers with a comprehensive framework for understanding AI's economic transformation.

Part I: Foundations of Transformative AI Economics

Agrawal, Gans, and Goldfarb (2025) establish a theoretical bedrock of the volume by modeling transformative AI as "genius on demand," showing how AI fundamentally alters the economics of knowledge work. Their framework distinguishes between routine knowledge workers who apply existing knowledge and genius workers who create new knowledge, demonstrating that AI's introduction will push human geniuses to increasingly novel domains while potentially displacing routine workers entirely. This foundational model provides the analytical lens through which subsequent chapters examine specific economic transformations. The authors operationalize Amodei's vision of a "country of geniuses in a datacenter," creating an economic framework that reveals how the commodification of genius-level intelligence reshapes comparative advantage across the knowledge economy. In the short run, geniuses are underutilized because managers are slow to adapt workflows, as the primary benefit of geniuses is solving non-routine

problems but most workflows are designed to operate without requiring the resolution of such problems. In the long run, however, workflows are adjusted and take advantage of the new supply of geniuses. Their insight that the value of an increase in the supply of geniuses depends on the demand for complex problem solving offers a framework for understanding the labor market dynamics explored throughout the volume.

Building on this foundation, Ben Jones (2025) examines how AI transforms the innovation process itself, introducing the crucial concept of bottlenecks that constrain AI's transformative potential. While Agrawal et al. focus on the supply of genius, Jones explores what happens when that genius is applied to research and development. He demonstrates that even with unlimited AI intelligence, physical experimentation, regulatory approval, and human adoption create binding constraints on innovation speed. This bottleneck framework becomes a recurring theme throughout the volume, tempering more explosive growth predictions with practical realities. Jones's analysis is particularly valuable in showing why broad task coverage matters more than narrow superintelligence—a finding that shapes how we think about AI development priorities. His work provides the theoretical bridge between AI capabilities and actual economic transformation, explaining why the path from laboratory to marketplace remains fraught with friction even in an age of artificial genius.

Coyle and Poquiz (2025) address a fundamental challenge that underlies all economic analysis of TAI: our statistical frameworks cannot adequately measure what they cannot see. They demonstrate how current GDP metrics, designed for an industrial economy, systematically miss AI's transformative impacts—from zero-priced digital goods to quality improvements in personalized services. This measurement blindness threatens both research and policy, as we risk making decisions based on statistics that reflect yesterday's economy rather than tomorrow's transformation. Their proposed solutions, including broader welfare metrics and real-time data collection, provide the empirical foundation necessary for validating the theoretical claims made throughout the volume. Without addressing what Coyle and Poquiz identify as the "AI measurement gap," policymakers risk fighting tomorrow's economic battles with yesterday's data—a theme that resonates strongly with Korinek and Lockwood's (2025) later analysis of how traditional tax bases may become unmeasurable.

Part II: Markets, Competition, and Organization

Athey and Morton (2025) shift focus from theoretical foundations to market realities, analyzing how AI's unique economic characteristics—high fixed costs, network effects, and data advantages—create natural tendencies toward concentration. Their general equilibrium analysis reveals a troubling "double harm" scenario where workers in AI-

importing countries face both job displacement and reduced purchasing power as monopolistic AI providers extract surplus. This market structure analysis provides essential context for understanding the distributional concerns that permeate the volume. Building on Ben Jones's bottleneck framework, they show how market power itself becomes a bottleneck to AI's beneficial diffusion. Their finding that even productivity-enhancing AI can reduce welfare in importing countries challenges simplistic narratives about technological progress. This work establishes why the organizational and institutional questions addressed in subsequent chapters matter: market structure shapes not just who captures AI's benefits, but whether those benefits materialize at all for most of humanity.

Hadfield and Koh (2025) extend the market analysis by envisioning an economy populated by AI agents as autonomous economic actors, fundamentally challenging our assumptions about market participants. They show how AI agents differ qualitatively from human actors—processing information at superhuman speeds, coordinating perfectly, and potentially circumventing regulations designed for human psychology. This analysis builds directly on Athey and Morton's concentration concerns, as AI agents might facilitate algorithmic collusion at unprecedented scales. Yet Hadfield and Koh also identify opportunities, showing how AI agents could enable new forms of economic organization that transcend traditional firm boundaries. Their framework for "AI-native" institutions provides a blueprint for the governance challenges explored later by Korinek and Lockwood. The authors make a compelling case that our economic infrastructure—from contract law to market oversight—requires fundamental redesign for an economy where machines negotiate with machines at the speed of light.

Shahidi et al. (2025) push the AI agent analysis to its logical extreme with their concept of the "Coasean Singularity"—a moment when transaction costs approach zero and traditional firm boundaries dissolve. Similar to Hadfield and Koh's institutional framework, they explore what happens when AI agents eliminate the frictions that have historically shaped economic organization. Their analysis reveals both utopian possibilities—perfectly efficient markets, frictionless matching, optimal resource allocation—and dystopian risks of manipulation, exploitation, and human irrelevance. The tension between platform-provided and user-controlled agents echoes Athey and Morton's concentration concerns while adding new dimensions of algorithmic governance. Their work suggests that Coase's fundamental insight about why firms exist must be reconsidered when artificial agents can coordinate without the transaction costs that plague human exchange. This chapter provides the theoretical capstone to Part II's exploration of how transformative AI impacts the very foundations of economic organization.

Chatterji, Rock, and Talamás (2025) bring Part II to a close by examining how firms must reorganize when intelligence—once scarce and embodied in human experts—becomes an abundant commodity. They show how the traditional advantages of large corporations (attracting top talent, coordinating complex projects) erode when any startup can access genius-level AI. Building on the market dynamics explored by Athey and Morton, the organizational possibilities outlined by Shahidi et al., and the supply shock of geniuses described by Agrawal et al., they demonstrate how firms must shift from hoarding intelligence to orchestrating it. Their analysis of "knowledge hierarchies" flattening into "execution networks" provides concrete illustrations of the abstract transformations theorized in earlier chapters. The authors' insight that competitive advantage shifts from intelligence to data, relationships, and physical assets helps explain why some sectors may concentrate further while others fragment—a nuanced view that enriches our understanding of transformative AI's differential impacts across the economy.

Part III: Labor, Distribution, and Human Welfare

Manning and Aguirre (2025) ground the volume's theoretical insights in empirical reality by examining which American workers face the highest risk from AI transformation. Their innovative "adaptive capacity index" reveals a crucial finding that challenges conventional wisdom: many highly-exposed workers possess strong adaptation capabilities, while 7.2 million workers combine high exposure with low adaptive capacity—concentrated in clerical and administrative roles. This empirical foundation validates concerns raised by Agrawal et al. about routine knowledge workers while adding nuance about which workers can successfully transition. Their positive correlation between AI exposure and adaptive capacity offers hope that the disruption may be less severe than feared for many workers. However, their identification of specific vulnerable populations—those in routine cognitive work—provides potential targets for policy intervention. This chapter exemplifies how careful empirical work can transform abstract concerns about AI displacement into actionable policy insights.

Restrepo (2025) provides the theoretical complement to Manning and Aguirre's empirical analysis, modeling the long-term trajectory of labor markets under AGI. His distinction between "essential work" and "accessory work" (non-essential) offers a more nuanced view than simple replacement narratives. Building on Agrawal et al.'s framework of genius specialization, Restrepo shows how human wages may converge to the computational cost of replicating human capabilities. This sobering finding validates concerns about labor's declining share of income. His mathematical demonstration that labor's share approaches zero even when humans retain some economic role provides a stark illustration of the distributional challenges ahead. Yet Restrepo's analysis also suggests

transition paths, showing how the sequence of automation matters for adjustment costs. This work provides essential theoretical grounding for why the fiscal innovations proposed by Korinek and Lockwood become necessary: in Restrepo's long-run equilibrium, traditional labor income taxation simply cannot function.

Stevenson (2025) shifts focus from economic mechanics to human meaning, addressing what may be the deepest challenge of the AI transformation. Drawing parallels to how women found purpose beyond household productivity after domestic automation, she explores how humanity might find meaning when machines handle most economically valuable work. Her three-phase framework—job transformation, economic restructuring, and existential reimagining—provides a temporal structure that complements the economic analyses of previous chapters. Stevenson's emphasis on human capacities for care, creativity, and connection that transcend economic value offers hope amid the disruption documented by Manning and Aguirre and theorized by Restrepo. Her historical perspective reminds us that humanity has navigated similar transitions before, though perhaps never at this scale or speed. This chapter provides essential balance to the volume, insisting that human worth cannot be reduced to economic productivity—a message that gains urgency as the economic logic of previous chapters points toward human obsolescence.

Ludwig et al. (2025) offer a more optimistic vision of human-AI collaboration through their framework of algorithmic choice architecture. Rather than seeing AI as simply replacing human decision-making, they show how personalized algorithms can help individuals reach their own "reflective equilibrium" making choices aligned with their true preferences rather than behavioral biases. This vision builds on the human-centric values emphasized by Stevenson while acknowledging the economic realities documented by Manning and Aguirre. Their five algorithmic functionalities—personalized diagnostics, choice translation, search assistance, outcome mapping, and implementation support—demonstrate how AI can enhance rather than replace human agency. The authors' emphasis on "first-person" rather than "third-person" paternalism resonates with concerns about concentrated decision-making power raised by Hadfield and Koh. By showing how AI can serve human flourishing rather than simply economic efficiency, Ludwig et al. provide a crucial counterweight to displacement narratives while acknowledging that realizing this positive vision requires careful institutional design.

Part IV: Information, Knowledge, and Systemic Risks

Brynjolfsson and Hitzig (2025) examine how TAI challenges Hayek's fundamental insight about the superiority of decentralized market decisions over central planning. They demonstrate that AI's ability to process vast amounts of information and simulate complex

interactions can reverse this logic, making centralized decision-making more efficient than distributed markets for the first time in history. This profound challenge to economic orthodoxy extends on Hadfield and Koh's analysis of AI agents while raising deeper questions about economic organization. Their concept of "alienable" local knowledge—information that AI can extract and process at scale—suggests that transformative AI might succeed where earlier efforts at central planning failed. Yet the authors also warn about concentration of power, echoing Athey and Morton's concerns about market structure while adding political dimensions. Their framework helps us understand why the information ecosystem questions raised by Stiglitz and Ventura-Bolet matter so deeply: if AI enables efficient central planning, then who controls AI systems becomes the central question of political economy.

Mullainathan and Rambachan (2025) explore how algorithms transform scientific practice itself, moving beyond Ben Jones's analysis of R&D bottlenecks to examine how AI changes the fundamental methods of knowledge creation. They show how algorithms can formalize the informal—turning hunches into hypotheses, intuitions into tests, and experience into experiments. This transformation of science from static theories to dynamic algorithmic representations promises to accelerate discovery in ways that Jones's bottleneck analysis might not fully capture. Their vision of "algorithmic science" where AI systems continuously update models based on new data represents a paradigm shift comparable to the scientific revolution itself. Building on Brynjolfsson and Hitzig's themes of centralized knowledge processing, they show how AI might not just accelerate existing science but create entirely new modes of understanding. This chapter provides crucial context for why measurement challenges identified by Coyle and Poquiz matter: if science itself becomes algorithmic, then understanding and governing these algorithms becomes essential for human agency in knowledge creation.

Stiglitz and Ventura-Bolet (2025) provide a sobering counterpoint to optimistic visions of AI-enhanced knowledge systems by demonstrating how AI could paradoxically degrade the information ecosystem even while improving information processing. Their formal model, building on the Grossman-Stiglitz paradox, shows how perfect AI transmission of information reduces incentives to produce original information, potentially leading to "informational collapse." This analysis gains urgency in light of Brynjolfsson and Hitzig's vision of centralized knowledge processing and Mullainathan and Rambachan's algorithmic science: if AI systems train on increasingly synthetic data while human information production declines, the foundations of knowledge itself become unstable. Their warning about misinformation becoming both cheaper to produce and harder to detect provides a crucial systemic risk that previous chapters largely overlooked.

Part V: Policy Responses and Long-term Considerations

Korinek and Lockwood (2025) tackle the fundamental challenge of maintaining public finance when AI erodes traditional tax bases, providing concrete policy solutions to the economic transformations documented throughout the volume. Their two-stage framework—first shifting from labor to consumption taxes, then taxing autonomous AI systems directly—offers a pragmatic path through the fiscal crisis created by AI displacement. Building on Restrepo's demonstration that labor income approaches zero, they show how governments must fundamentally reimagine revenue generation. Their optimal AI tax formula elegantly demonstrates that the appropriate tax rate depends on human patience, not AI productivity—a profound insight that reframes how we think about sharing AI's benefits. The authors' analysis of specific proposals (robot taxes, compute taxes, windfall clauses) provides the practical tools policymakers need to navigate the transition. This chapter exemplifies how economic theory can generate actionable policy insights, transforming abstract concerns about inequality and displacement into concrete fiscal mechanisms that ensure AI benefits are broadly shared rather than narrowly captured.

Chad Jones (2025) concludes the volume by quantifying humanity's most consequential investment decision: how much to spend reducing AI's existential risks. Using the statistical value of life as a benchmark, he calculates that optimal safety investment likely exceeds 1% of GDP annually, with baseline estimates around 16%. This economic approach to existential risk provides a rational framework for a debate often dominated by speculation and fear. Jones's analysis gains credibility from the systemic risks identified throughout the volume: Stiglitz and Ventura-Bolet's information collapse, Brynjolfsson and Hitzig's concentration of decision-making power, and Athey and Morton's market concentration all suggest ways transformative AI could go catastrophically wrong. His willingness-to-pay methodology transforms abstract existential concerns into concrete budget allocations, showing how standard economic tools can address even civilization-scale challenges. By grounding AI safety in the familiar framework of cost-benefit analysis, Jones provides policymakers with analytically rigorous justification for substantial safety investments. This final chapter brings the volume full circle: from Agrawal et al.'s vision of the long run adaptation to a genius supply shock to Jones's calculation of what we should pay to ensure that transformation enhances rather than eliminates human flourishing.

Advancing the Research Agenda

This volume represents an important first step in developing the economic frameworks necessary to understand and shape the AI transformation, yet the research agenda extends far beyond what these chapters can address. In our article "A Research Agenda

for the Economics of Transformative AI"(Brynjolfsson, Korinek, and Agrawal, 2025), we offer a systematic research agenda organized around nine Grand Challenges spanning economic growth, innovation, income distribution, decision-making power, geoeconomics, information flows, AI safety, human well-being, and transition dynamics. We suggest a set of research methodologies, tools, and frameworks that can be applied to address these questions. This framework emphasizes the urgency of proactive research, noting that while technological capabilities advance rapidly, our institutions and policies evolve slowly—creating a dangerous gap that economic research must help bridge. In addition, Korinek (2024) provides a research roadmap emphasizing the profound paradigm shift as human-level intelligence becomes reproducible and loses its scarcity value. His work also highlights the benefits of employing increasingly capable AI systems to accelerate economic research.

The stakes of this research agenda could not be higher. We stand at a unique moment where economic analysis can shape humanity's most consequential technological transition. Unlike previous technological revolutions that enhanced human physical capabilities or processed information faster, transformative AI promises to replicate and exceed human cognitive abilities. This raises fundamental questions that economics is uniquely positioned to address: How do we maintain broad prosperity when machine intelligence can substitute for human intelligence? What market structures and institutions can harness AI's benefits while mitigating concentration of power? How do we redesign social contracts and fiscal systems for an economy where traditional employment may become less broadly distributed or even obsolete?

The analyses in this volume demonstrate that economic research provides essential tools for navigating these challenges. Through rigorous modeling, empirical analysis, and institutional design, economists can help ensure that the AI transformation enhances rather than undermines human flourishing. The path ahead requires not just understanding AI's technical capabilities but designing the economic frameworks, policies, and institutions that channel those capabilities toward broadly shared prosperity. The chapters that follow begin mapping that journey, but the full research agenda will require the sustained effort of a broad and diverse community of economists. The future we create with transformative AI tomorrow will be shaped by the economic insights we develop today.