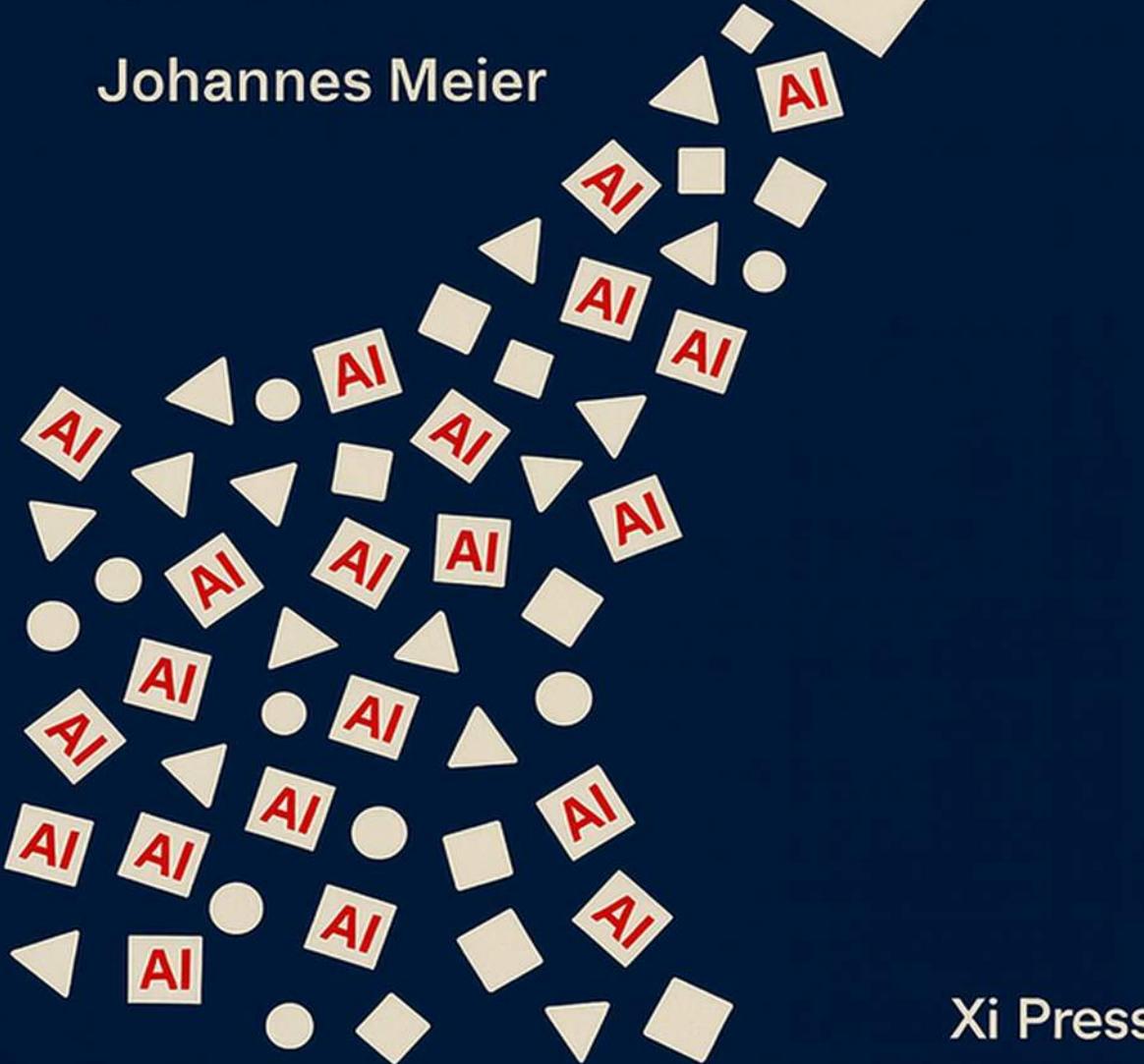


# CHANGE MANAGEMENT IN THE AGE OF AI

Johannes Meier



Xi Press

# Change Management in the Age of AI

Navigating Transformation Across Individual,  
Organizational, and Societal Levels

Johannes Meier

Xi Press

December 2025



# **Copyright**

© 2025 Xi GmbH, Gütersloh, Germany

Published by *Xi Press*, an imprint of Xi GmbH

Print Edition ISBN 978-3-00-085630-3

No part of this publication may be reproduced without prior written permission of the publisher.



# Contents

<b>Copyright</b>	i
<b>Preface</b>	xii
<b>Book Overview</b>	1
<b>Introduction: The Adaptation Crisis</b>	7
Who This Book Is For . . . . .	9
An Interdisciplinary Perspective . . . . .	11
Grand Challenge of Growing AI Capabilities . . . . .	16
Toward Antifragile Change Models . . . . .	31
<b>AI Fundamentals for Change Leaders</b>	35
Major AI Milestones in Modern AI Development . . . . .	35
Deep Neural Networks as Core Building Block . . . . .	39
The Transformer Architecture . . . . .	44
Different Paths to Model Adaptation . . . . .	52
Extensions of the Transformer Architecture . . . . .	60
Agentic AI Systems . . . . .	65
Alignment and Safety . . . . .	75
Conclusion: Long-Term AI Risks . . . . .	79
Key Takeaways: AI Fundamentals . . . . .	85
<b>Individual Change in the Age of AI</b>	87
The New Landscape of Individual Change in the Age of AI . . . . .	87
Immunity to Change . . . . .	90
Single-Loop vs. Double-Loop Learning . . . . .	96
Personal Mastery and an Individual Growth Mindset . . . . .	97

## *Contents*

Meaning-Making Under Pressure . . . . .	99
Consciousness as Capacity . . . . .	102
Neuroscience of Attention: Right vs. Left Brain Hemisphere . .	108
Practical Tools . . . . .	112
Conclusion: Developing Self-Transforming Capabilities . . . .	122
Key Takeaways: Individual Change . . . . .	127
<b>Organizational Change in the Age of AI</b>	<b>129</b>
Why Organizational Transformations Fail . . . . .	130
Lewin's and Schein's Foundational Frameworks . . . . .	131
Layers of Organizational Resistance in an AI-Driven World . .	135
The Dual Role of AI as Disruptor and Enabler . . . . .	151
Roadmap to an Adaptive, AI-Empowered Organization . . . .	155
Practical Tools . . . . .	163
AI Mediator . . . . .	175
Conclusion: Towards More Adaptive and Learning Organizations	180
Key Takeaways: Organizational Change . . . . .	182
<b>Societal Change in the Age of AI</b>	<b>183</b>
Reflexivity and the Interconnection of Change . . . . .	184
A Shifting Post-COVID Zeitgeist: Instability, AI, and Global Risks	185
The Limits of Linear, Technocratic Models of Change . . . .	188
Non-Linear and Path-Dependent Societal Transition . . . .	191
Fragility, Pseudo-Stability, and Antifragility . . . . .	197
Values, Worldviews, and Meaning-Making in Societal Change .	205
Implications for Institutions and Long-Term Resilience . . . .	206
Practical Tools . . . . .	208
Conclusion: A More Political Modern Leadership Agenda . . . .	214
Key Takeaways: Societal Change . . . . .	218
<b>Societal Impacts of AI</b>	<b>221</b>
Economic and Workforce Impacts of AI . . . . .	221
Education and Personalized Learning at Scale . . . . .	233
Scientific Research Acceleration and Epistemological Risks . .	237
Privacy Erosion and Surveillance Risks . . . . .	241
The Attention Economy and Algorithmic Manipulation . . . .	244

*Contents*

The New Geography of Power . . . . .	249
Cultural and Mental Impacts . . . . .	255
Emerging Systemic Risks . . . . .	258
Divergent Views on AI's Future . . . . .	260
Conclusion: Diffusion Barriers and Policy Implications . . . . .	269
Key Takeaways: Societal Impacts of AI . . . . .	275
<b>Conclusion: Leading Through Transformation in the AI Age</b>	<b>277</b>
Individual Transformation: Developing Adaptive Capabilities . .	277
Organizational Transformation: Towards Hybrid Organizations	279
Societal Transformation: Leadership as a Political Act . . . . .	281
Reflexive Individual, Organizational, and Societal Levels . . . . .	283
AI as Both Disruptor and Enabler . . . . .	284
Moving Forward: Principles for Continued Learning . . . . .	285
<b>Bibliography</b>	<b>287</b>
<b>About the Author</b>	<b>303</b>



# List of Figures

1	Dual role of AI with regard to change at all levels . . . . .	6
2	Interconnected dimensions of change . . . . .	14
3	Change management and related perspectives . . . . .	15
4	Dimensions of intelligence . . . . .	17
5	AI models reaching human capabilities . . . . .	18
6	Rapid progress of AI model capabilities . . . . .	19
7	Exponential decrease in cost of AI models for comparable performance . . . . .	20
8	Emotional intelligence testing of AI models . . . . .	22
9	Persuasiveness of AI model . . . . .	23
10	Real-world performance benchmarking of AI models . . . . .	24
11	Adoption speed of technologies . . . . .	25
12	The jagged frontier of AI capabilities . . . . .	26
13	AI performance within current capability frontier . . . . .	27
14	AI performance outside current capability frontier . . . . .	28
15	Most basic Deep Neural Network . . . . .	40
16	Spatial similarities of meaning with the help of embeddings . . . . .	42
17	Exploring the transformer interactively . . . . .	45
18	Basic RLHF workflow . . . . .	56
19	A simple multi-agent system for report generation . . . . .	67
20	AI Co-Scientist multi-agent system with human in the loop . . . . .	69
21	Automated scientific discovery with multi-agent system . . . . .	71
22	AlphaEvolve architecture . . . . .	75
23	AI models resorting to blackmail to reach their goals . . . . .	78
24	Kübler-Ross Change Curve . . . . .	90

## *List of Figures*

25	Shlomo Shoham’s vicious circle . . . . .	95
26	Changing views on mental complexity . . . . .	103
27	Stages of adult mental development . . . . .	106
28	Mapping status among the cardinals prior to the conclave	141
29	Ecosystem equilibrium states vary with conditions . . . . .	193
30	Hysteresis in the Greenland Ice Sheet . . . . .	194
31	Two essential types of nonlinearity . . . . .	198
32	The concave phenomenon of driving a car against an obstacle	199
33	GPT-4 predictions in simulated surveys . . . . .	210
34	Solvable software tasks by model generation . . . . .	225
35	Hybrid teams outperforming human teams or individuals.	229
36	AI will reshuffle the economy fundamentally. . . . .	231
37	AI impact on legal tasks . . . . .	232
38	Exponential gap . . . . .	263
39	AI Policy Agenda . . . . .	270

# List of Tables

1	Strengths and Limitations of Transformer LLMs . . . . .	45
2	Example of Immunity to Change map . . . . .	92
3	Jagged Frontier Mapping Template . . . . .	163
4	Learning from nature . . . . .	199
5	Fragility drivers versus antifragility drivers . . . . .	203



# Preface

This book distills the lectures from my course “Change Management in the Age of AI”, which I have taught for many years in the MBA programs at HHL Graduate School of Management in Leipzig.

Given AI’s rapid technological evolution, I update the course content annually. Consequently, this book captures a fleeting snapshot of current technological developments and AI applications. Nevertheless, the fundamental principles and heuristics for navigating complex, nonlinear change at the individual, organizational, and societal levels remain enduring and robust across time.

Many AI-agents based on GPT-5.1, Claude Sonnet 4.5, and Gemini 3 Pro supported me in a Cursor environment in researching references, prototyping applications, improving my writing, creating infographics including the title page, getting critical editorial feedback, and typesetting the book. They took over many of the tedious tasks and left me with the fun parts of writing this book. I found working in this “hybrid team” an amazing experience.



# Book Overview

The book opens by confronting a sobering reality: despite decades of change management research and practice, failure rates remain stubbornly high. This persistence of failure signals that traditional approaches are fundamentally inadequate for today’s grand challenges. Among these challenges, AI stands out as particularly transformative, reshaping how we work, compete, and create value at a fundamental level while amplifying all other challenges.

The introductory chapter highlights that traditional change management approaches tend to fail us as they assume gradual, predictable transformation. The exponential, nonlinear disruptions of the AI age shatter that assumption. We face an “exponential gap”: the widening chasm between how fast AI capabilities evolve and how slowly our organizations, mindsets, planning processes, and governance structures adapt.

Success requires transforming how we think about change itself – moving from discrete projects with clear endpoints to continuous adaptation as an individual and collective capability. The book integrates insights from developmental psychology, systems thinking, complexity theory, organizational design, and AI capabilities research. Throughout, I emphasize practical application: each chapter concludes with actionable tools and frameworks that leaders can implement immediately.

Before diving into change management strategies, the chapter on AI fundamentals provides essential grounding for non-technical leaders. Core technical concepts are explained in accessible terms: deep neural networks as function approximators that learn from data, the transformer architecture’s attention mechanism that enables parallel processing and elicits long-range dependencies, and the training of these systems through backpropagation

## *Book Overview*

and optimization. This technological foundation enables readers to engage meaningfully with AI opportunities and risks throughout the subsequent chapters.

The premise of the chapter on individual change is that effective change management starts with individual transformation. The primary constraint in navigating complex change isn't technical knowledge but rather the individual's capacity to process complexity, tolerate ambiguity, and continuously reconstruct their mental models. Drawing on Robert Kegan's developmental psychology, three levels of adult consciousness can be distinguished. Most leaders operate with a "socialized mind" shaped by external expectations, or a "self-authoring mind" guided by internal principles. However, the adaptive challenges of the AI age increasingly demand a "self-transforming mind" – one that can hold multiple contradictory perspectives simultaneously, embrace genuine paradox, and view identity itself as fluid rather than fixed. This isn't about accumulating more knowledge but expanding the very structure of consciousness through which we perceive reality.

I present several practical frameworks (detailed with implementation guidance in the chapter): **Immunity to Change mapping** reveals hidden psychological commitments that sabotage change goals; **double-loop learning** encourages questioning underlying assumptions; **personal mastery and growth mindset** provide frameworks for continuous learning; **meaning-making under pressure** (drawing on Viktor Frankl) provides resilience during turbulent transitions; and **whole-brain thinking** balances analytical strengths with contextual awareness.

AI serves not just as a disruptor but as a developmental tool itself. AI coaches can provide 24/7 support for reflection, offer Socratic questioning, and help leaders test assumptions – democratizing access to developmental support. Specific prompts and protocols illustrate how AI can be used as a thinking partner in your own development.

The chapter on organizational change demonstrates that resistance to change manifests across three interconnected contexts: (1) **Formal context** – structures, processes, metrics, and incentive systems that often inadvertently reward old behaviors; (2) **Social context** – trust

levels, communication patterns, and the “psychological contract” between employer and employees; and (3) **Mental context** – collective mindsets and paradigms, including unspoken commitments to past success formulas that unconsciously sabotage adaptation.

The chapter demonstrates how to use data-driven tools for diagnosing resistance across these contexts. Techniques like **sentiment analysis**, **organizational network mapping**, and **behavioral tracking** reveal where adoption stalls and why. Targeted interventions – differentiated by employee archetype (from “Silent Resistors” to “Anxious Learners” to “Active Opponents”) – prove far more effective than uniform change programs. Organizations can use AI itself to manage change: sentiment analysis reveals hidden concerns, predictive analytics identify adoption barriers before they become crises, and AI agents can provide personalized support at scale.

An organizational transformation roadmap includes: comprehensive diagnosis across all three contexts; reshaping structures to align with desired behaviors; strengthening social fabric through psychological safety; challenging mental models through scenario planning; leveraging AI tools for continuous monitoring; and building organizational resilience through redundancy, diversity, and distributed decision-making. Given the current maturity of AI systems, it is important to maintain human judgment while leveraging AI’s analytical power – creating “hybrid intelligence” rather than replacement automation.

Organizational success ultimately depends on healthy societal ecosystems. Leaders cannot afford to view their organizations as separate from broader social, political, and economic dynamics – AI renders these interdependencies impossible to ignore. The chapter on societal change draws on complexity theory and historical analysis to explain why societal transitions are often fundamentally nonlinear and path-dependent, crisis-driven rather than gradually managed. This creates both danger (catastrophic tipping points) and opportunity (crises open windows for transformative change).

The concept of “antifragility” becomes crucial at this level. Modern societies, optimized for efficiency, have often become dangerously fragile –

## *Book Overview*

witness brittle supply chains during COVID-19, healthcare systems without surge capacity, and financial systems vulnerable to cascading failures. Building societal resilience requires accepting some inefficiency to maintain redundancy, cultivating diversity, ensuring decision-makers bear consequences (“skin in the game”), and maintaining optionality rather than premature optimization.

Leadership in the AI age is unavoidably political – not in partisan terms, but in building coalitions, negotiating between competing interests, creating shared meaning across differences, and forging collective action despite fragmentation. The skills required increasingly resemble statecraft: reading complex stakeholder dynamics, sensing shifts in power configurations, brokering compromises that preserve core values, and communicating vision across diverse worldviews.

Practical tools include agent-based modeling for policy testing, digital twins for scenario simulation, and frameworks for participatory AI governance. However, I consistently stress these are aids for human judgment, not replacements for democratic deliberation.

The chapter on the societal impacts of AI dives deeper into AI’s broader societal effects. Key areas include:

- **Economic disruption and workforce transformation:** While AI promises productivity gains, distribution remains deeply uneven. Entry-level positions face particular risk as AI eliminates traditional pathways for skill development.
- **The attention economy and manipulation risks:** AI-powered recommendation systems optimize for engagement rather than well-being, creating “weapons of mass distraction” that affect employee focus, mental health, and organizational culture.
- **Privacy erosion and surveillance capitalism:** Surveillance infrastructure already exists at scale. Organizations collecting and deploying AI on personal data bear ethical responsibilities that extend far beyond legal compliance.
- **Power concentration:** AI capabilities, training data, and computational resources are concentrating in a small number of firms and

- nations, creating “Matthew effects” where those ahead accelerate further ahead.
- **Deskilling and expertise erosion:** Heavy AI reliance can atrophy human capabilities, creating dangerous dependencies and raising fundamental questions about maintaining human expertise alongside AI deployment.

Among AI experts, there is disagreement about whether AI is a “normal technology” that can be managed with established change management and risk containment approaches, or whether we face a fundamental tipping point with the advent of artificial superintelligence. Even without making a prediction on the timing and impact of artificial superintelligence, it is clear to me that the widening chasm between accelerating technological capabilities and incremental institutional adaptation exceeds what conventional management approaches can address, requiring fundamental shifts in how we understand change, organizational function, and societal adaptation.

The book concludes by synthesizing insights across all three levels of change and emphasizing their reflexive interconnections. The following figure visualizes this book’s central argument: AI’s dual role as both disruptor and enabler across all three interconnected levels of change.

The conclusion reinforces that individual, organizational, and societal transformation are not separate domains but interconnected aspects of a single complex system. An individual leader’s expanded consciousness enables organizational innovations previously unthinkable; organizational practices that distribute agency and cultivate learning create environments where individuals can develop; societal investments in education and social cohesion determine what human capital organizations can draw upon. AI makes this integration both more urgent and more complex, as its impacts cascade across all levels.

AI plays a dual role: as the source of disruption requiring new forms of leadership, and as a tool that, properly deployed, enables that leadership to succeed. The task isn’t choosing between these possibilities but holding

## *Book Overview*

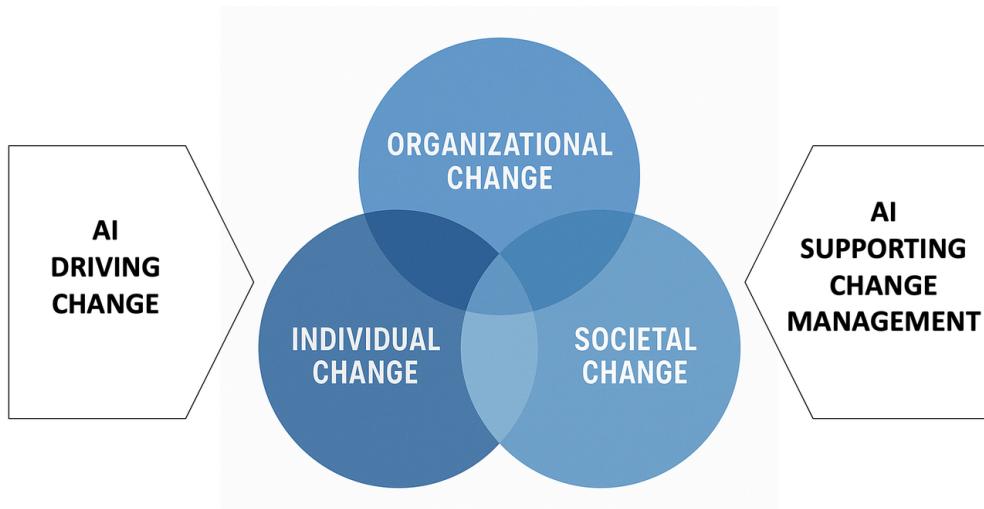


Figure 1: Dual role of AI with regard to change at all levels

them in productive tension – leveraging AI’s capabilities while constraining its risks, using it to amplify human capability while ensuring humans remain genuinely in control.

The book’s ultimate message is one of tempered but genuine optimism. The challenges are real and urgent, but human institutions have navigated previous transformations that seemed overwhelming at the time. Success requires abandoning the comfortable fiction that we can predict and plan our way through this transition, and instead building the capacity for continuous learning and adaptation. The age of AI demands nothing less than leadership at its best – technically informed, psychologically aware, organizationally sophisticated, and politically engaged.

# Introduction: The Adaptation Crisis

Leading organizational change has long been a fraught endeavor, with a majority of initiatives falling short of their goals. Up to 60% of change programs fail to achieve their targeted outcomes<sup>1</sup>. Beer and colleagues documented this failure rate in the 1990s, and later research confirmed it<sup>2</sup>. A 2024 survey of executives by Bain & Company suggests that 88% of business transformations fail to achieve their original ambitions.<sup>3</sup> The persistence of such failure rates over decades signals a need to rethink conventional change management approaches. In an era of rapid disruption and complexity, simply doing “more of the same” is clearly insufficient. Organizations require new frameworks that address the root causes of change failures and adapt to today’s challenging environment.

Organizations today operate amid an unprecedented scale of global grand challenges that drive the need for continual change. These grand challenges define the modern leadership context: globalization pressures industries to integrate across borders; demographic shifts and equity issues demand more inclusive strategies; digitization and emerging technologies like AI and bio-engineering are transforming business models; global health crises

---

<sup>1</sup>Beer, Michael et al., “Why Change Programs Don’t Produce Change,” *Harvard Business Review* 68, no. 6 (1990): 158–66.

<sup>2</sup>Jørgensen, Henrik H. et al., “Making Change Work,” *IBM Global Business Services*, 2008.

<sup>3</sup>Bain & Company, *88% of Business Transformations Fail to Achieve Their Original Ambitions; Those That Succeed Avoid Overloading Top Talent*, Bain & Company, 2024, <https://www.bain.com/about/media-center/press-releases/2024/88-of-business-transformations-fail-to-achieve-their-original-ambitions-those-that-succeed-avoid-overloading-top-talent/>.

## *Introduction: The Adaptation Crisis*

(as seen in recent pandemics) test system resilience; climate change and natural resource constraints impose new risks; and volatile geopolitics and security concerns create uncertainty.

Each of these forces pushes organizations to adapt or risk obsolescence. Among them, AI stands out as a particularly transformative challenge, reshaping economies and societies at a fundamental level. AI's rapid advancement and pervasive impact cut across industries, making it not just another challenge but a catalytic force that amplifies all others. For this reason, this book places a special focus on AI as a highly dynamic grand challenge. By examining change management through the lens of AI, we can explore how to harness this powerful technology while addressing the risks it poses.

Before 2016, when AlphaGo defeated the world's best Go player, experts predicted it would take decades for AI to master such intuitive, strategic thinking. By 2023, AI models were passing bar exams, writing complex code, and generating persuasive arguments that outperformed humans. The technology moved from science fiction to boardroom reality in less than a decade. Meanwhile, new AI models dominate the market for just a few weeks before being surpassed by more powerful alternatives in benchmarks.

Consider what this means in practice. In 2023, a major consulting firm deployed GPT-4 to 758 consultants. Within weeks, those using AI outperformed their peers by 43% on tasks inside AI's capability frontier. But on tasks outside that frontier, AI users performed 23% worse – confidently producing errors they couldn't detect.<sup>4</sup> The technology's capability frontier has been advancing unevenly every month since then. Yet a global survey of more than 1,000 executives in 2024 revealed that only 26% of companies have developed the necessary capabilities to move beyond proofs-of-concept and generate "tangible value."<sup>5</sup> This isn't a coincidence. It's a symptom of

---

<sup>4</sup>Dell'Acqua, Frabrizio et al., "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Harvard Business School Working Paper*, 2024.

<sup>5</sup>Bellefonds, Nicolas de et al., *Where's the Value in AI?*, Boston Consulting Group, 2024, <https://www.bcg.com/publications/2024/wheres-value-in-ai>.

## *Who This Book Is For*

the “exponential gap”: the widening chasm between how fast AI capabilities evolve and how slowly our organizations, our mindsets, our governance structures change.

In this context the term *VUCA* – short for *Volatility, Uncertainty, Complexity, and Ambiguity* – has become a catchphrase to describe the turbulent environment organizations face. Grand challenges contribute to a VUCA world: markets and technologies change with volatile speed, future outcomes are uncertain, system interdependencies breed complexity, and information can be ambiguous. In theory, leaders are counseled to counter VUCA with a set of corresponding responses: provide *Vision* to navigate volatility, pursue *Understanding* to counter uncertainty, seek *Clarity* to simplify complexity, and cultivate *Agility* to respond to ambiguity. This mnemonic mapping of VUCA to vision, understanding, clarity, and agility is reassuring but incomplete.

On its own, such high-level advice remains banal without concrete guidance on how to achieve it in practice. Simply knowing that agility or clarity is needed does not automatically translate into methods for creating them. A core premise of this book is that we must move beyond the buzzwords. Confronting VUCA challenges requires tangible frameworks, tools, and examples that show how to develop vision or agility within an organization’s unique context. In the chapters ahead, the VUCA concept will be translated into practical approaches – turning an abstract acronym into actionable strategies. The goal is to help leaders not just acknowledge volatility or complexity, but to convert VUCA into an advantage through smarter planning and adaptation. This begins with broadening our perspective on change itself, ensuring we understand the full landscape in which change occurs.

## **Who This Book Is For**

I wrote this book for leaders and practitioners who recognize that traditional change management approaches cannot navigate the disruptions of

## *Introduction: The Adaptation Crisis*

the AI age. While the frameworks and tools presented here are applicable across many contexts, the book speaks most directly to several key audiences:

- **Executives and Senior Leaders** facing the challenge of transforming their organizations in response to AI and other exponential technologies. If you’re responsible for strategic direction, organizational culture, or major transformation initiatives, this book provides frameworks for diagnosing resistance, designing adaptive structures, and building antifragile organizations that thrive on volatility rather than merely surviving it.
- **Change Management Practitioners and Consultants** seeking to update their toolkit for an era of nonlinear, complex change. The book offers practical diagnostic methods, intervention strategies, and tools that go beyond traditional project-based change management to address the deeper psychological, social, and systemic barriers that often derail transformation efforts.
- **Individual Managers** who want to develop their own capacity to navigate uncertainty and lead effectively in turbulent times. The book’s focus on individual development – expanding consciousness, developing personal mastery, and cultivating whole-brain thinking – provides a roadmap for your own growth, regardless of your current role or title.
- **Policy Makers and Public Sector Leaders** grappling with how to govern AI, manage societal transitions, and build resilient institutions. The societal-level analysis and governance frameworks address the political and institutional dimensions of change that extend beyond organizational boundaries.
- **Technical Leaders and AI Practitioners** who understand the technology but need frameworks for managing its organizational and societal impacts. The book bridges technical understanding with leadership and change management, helping you translate AI capabilities into sustainable organizational transformation.

Regardless of your specific role, this book assumes you are facing real change challenges and are willing to question conventional approaches. It

is written for those who recognize that leading change in the AI age requires not just new tools, but new ways of thinking about change itself – embracing complexity rather than simplifying it away, building adaptive capacity rather than optimizing for stability, and developing the personal and organizational capabilities to thrive in an uncertain future.

## **An Interdisciplinary Perspective**

Effective change management demands a rich view of reality that accounts for multiple levels and contexts, rather than a one-size-fits-all approach. We will look at *three fundamental levels of change: the individual, organizational, and societal levels*. Change at the individual level involves shifts in mindsets, skills, and behaviors of people. At the organizational level, change encompasses transformations in processes, structures, and culture within companies or institutions. At the societal level, change involves broader systemic shifts – in communities, markets, or even global systems – often influenced by policy, economics, or social movements. Successful change initiatives recognize that these levels are interconnected. For example, a new strategy (organizational change) may falter if individuals don't buy in (individual change), or if societal trends and stakeholders (societal context) aren't considered. Throughout this book, I will diagnose barriers and interventions across all three levels of change to ensure alignment from the personal to the global.

In addition to levels, we must consider *multiple interconnected contexts of change*. Research and experience highlight at least three critical contexts: the *mental context*, i.e., prevailing mindsets, beliefs, and assumptions (often tacit) that shape how people perceive change; the *social context*, including networks, communication patterns, relationships, and informal norms that influence collaboration; and the *formal context*, involving structures, processes, and incentives defined by the organization. Neglecting any of these contexts can doom a change effort – for instance, introducing a new formal process will misfire if the mental context (people's attitudes and understanding) and social context (peer support, trust) are not addressed.

## *Introduction: The Adaptation Crisis*

The dynamic complexity of today's business environment and human behavior within these three contexts undermines the premises of traditional management models. Dense interdependencies and knowledge driven relational networks make the "superior foresight" model of leadership increasingly difficult to defend. Both within classical firms and across organizational boundaries, self-organization can no longer be contained. In open, non-deterministic organizations, the reflecting, competent, and motivated individual – and their "performance equation" – emerges as the focal point of interest and intervention.

Everyone acknowledges we live in an information society where knowledge has become a key resource for competitive advantage. Yet our current conception of knowledge remains trapped in a linear, single-truth perspective, leading to the problematic – if not erroneous – assumption that information and knowledge-in-use are "basically the same thing." In reality, we should move from "information" (facts, data, statements, beliefs) through "interpretation" (cognitive ordering, emotional and social validation) to "knowledge-in-use" (personal perspective on action, willingness to accept the consequences of interpretation-based decisions). This individual and collective interpretive step becomes the crucial bottleneck for value creation and change management.

There is not one reality, as was traditionally assumed (long before the emergence of fake news and alternate facts). The process of discovering and mastering the multi-dimensionality and diversity of realities across interconnected levels and contexts is not a purely rational endeavor that can be based on explicit representations of only one reality. Stakeholders may not share mental models that are sufficiently compatible to enable productive interaction. This complexity of levels and contexts exposes the naivety of deterministic intervention programs.

Furthermore, *many significant changes are nonlinear* in nature, involving disruptions, exponential trends, or feedback loops that defy conventional prediction and planning methods. Yet most traditional change management tools assume linear, predictable change: steady progression from one state to another through clearly defined stages. Leaders who rely on these

tools simplify away tipping points and compounding effects, treating complex adaptive challenges as if they were merely complicated technical problems with straightforward solutions.

The reality is that organizational systems can remain stable for extended periods before suddenly shifting dramatically when critical thresholds are crossed. Alternatively, small initial changes can cascade through interconnected networks to produce disproportionately large outcomes. Embracing nonlinearity requires acknowledging that cause and effect may not be proportional or immediate, and that interventions in one part of a system can produce unexpected consequences elsewhere, sometimes with significant time delays that obscure the connection between action and result.

Change strategies must accommodate surprises and sudden shifts, building in flexibility, redundancy, and sensing mechanisms that allow organizations to detect weak signals of emerging transformation before they become overwhelming forces. By taking into account the three levels, the three types of contexts and the often nonlinear dynamics of change, we lay the groundwork for more robust change strategies that can navigate complexity rather than being blindsided by it, creating organizations that are resilient enough to absorb shocks and agile enough to capitalize on unexpected opportunities that emerge from turbulent environments. These interconnected dimensions are illustrated in the following figure.

Given the multifaceted nature of change, a key premise of this book is the need for an interdisciplinary approach to change management. No single academic or professional discipline has all the answers; instead, effective change leadership draws on insights from both the hard systems sciences and the social sciences. We need both classical paradigms: Erklären (explaining) and Verstehen (understanding). These terms, rooted in German scholarly tradition, capture the distinction between explaining phenomena through objective, external analysis versus understanding phenomena through empathetic, internal insight. In a change context, the systems science perspective corresponds to the explaining paradigm. It emphasizes data, processes, structures – the quantifiable “system conditions” of an organization. This might include mapping workflows, analyzing metrics, or modeling how a change will propagate through a business system.

## *Introduction: The Adaptation Crisis*

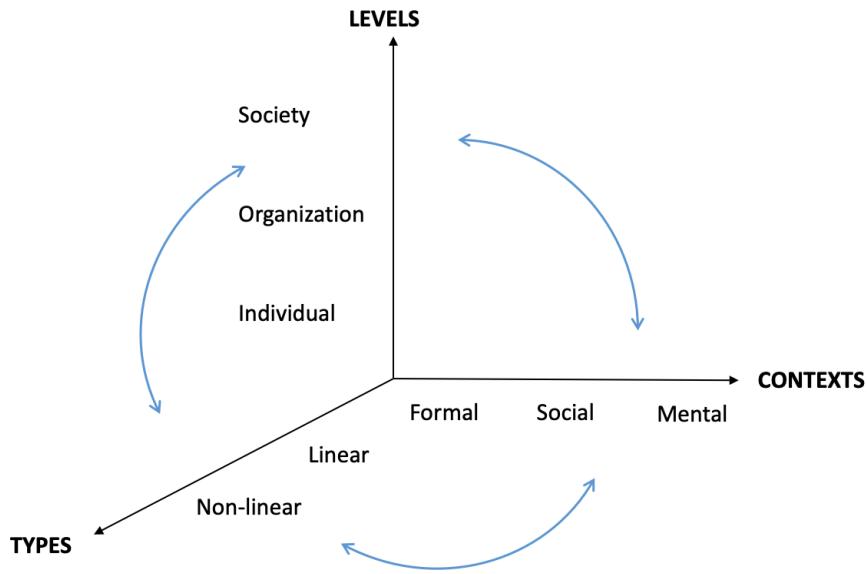


Figure 2: Interconnected dimensions of change

By contrast, the social science perspective corresponds to the understanding paradigm, focusing on the “human condition” – subjective experiences, meanings, and cultural factors. This could involve gauging employee sentiment, understanding informal power dynamics, or appreciating the organization’s history and identity.

Blending these perspectives yields a more holistic approach. Systems thinking tools (from fields like systems dynamics or engineering) help explain and predict how changes in one part of a system affect the whole, offering clarity and logic. Social science and humanities insights (from psychology, sociology, anthropology) help interpret how people make sense of change, ensuring empathy and legitimacy. For example, a later chapter might use network analysis (a systems tool) to identify influential actors in a change effort, while also applying behavioral science to craft messages that resonate with those actors. By leveraging both “explanation” and “understanding”, change leaders can design interventions that are technically sound and culturally acceptable. This interdisciplinary lens is critical, es-



Figure 3: Change management and related perspectives

pecially as we turn to the novel challenges posed by AI-driven change – where both complex technology systems and human factors collide.

A systemic perspective on change management shows the interconnected perspectives and disciplines that must be considered for successful transformation. Change management cannot operate in isolation from other organizational functions and perspectives, as illustrated in the following

## *Introduction: The Adaptation Crisis*

figure.

Effective change initiatives require strategic alignment to ensure transformation efforts support broader organizational goals, while simultaneously demanding robust communication strategies to build understanding and buy-in across stakeholders. The human dimension of change necessitates attention to personal development, helping individuals build new capabilities and mindsets required for transition, just as organizational development provides the structural frameworks and cultural shifts needed at the systemic level.

Moreover, change inherently generates conflict as old ways clash with new approaches, making conflict management essential for navigating resistance and tension productively. Finally, innovation management connects to change as organizations must not only implement new ideas but also build the adaptive capacity to continuously evolve.

This systemic interconnection means that successful change leaders must think holistically, recognizing that pulling one lever affects all others, and that sustainable transformation requires coordinated attention across all these interdependent domains.

## **Grand Challenge of Growing AI Capabilities**

Among the grand challenges, AI demands special attention because of the rapid evolution of its capabilities across nearly every dimension of intelligence. Tasks once thought exclusive to human cognition are now being mastered or augmented by AI. As shown in the following figure, AI capabilities span multiple dimensions of intelligence.

**A Note on AI Examples and Timeliness:** Throughout this book, specific AI models, capabilities, benchmarks, and research findings mainly from 2023-2025 are referenced. These examples serve to illustrate principles and patterns, but readers should understand that AI technology evolves at an extraordinary pace. By the time you read this, newer models will have emerged, benchmarks will have been surpassed, and some capabilities

## Dimensions of Intelligence

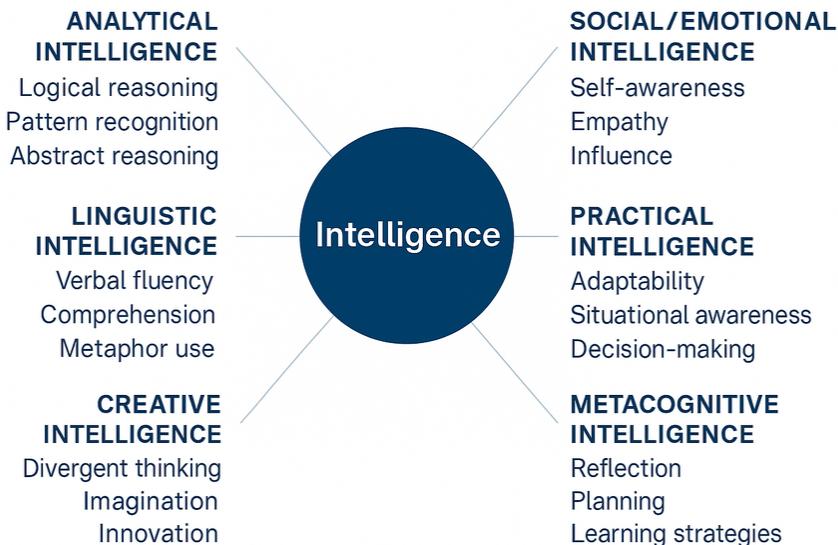


Figure 4: Dimensions of intelligence

described here may seem quaint while others may have proven more challenging than anticipated. The specific examples are snapshots in time – what endures are the underlying principles: how to diagnose resistance to technological change, how to build adaptive organizations, how to develop the personal and institutional capacity to navigate exponential disruption. The frameworks and tools presented here are designed to remain relevant even as the specific AI technologies continue their rapid evolution.

AI systems now demonstrate strong logical-mathematical intelligence through complex problem-solving, pattern recognition, and data analysis at scales far exceeding human capacity. Linguistic intelligence has advanced dramatically, with large language models (LLMs) capable of generating coherent text, translating between languages, engaging in nuanced conversation, and even producing creative writing. The following figure shows how AI models are reaching or exceeding human capabilities across various benchmarks.

## *Introduction: The Adaptation Crisis*

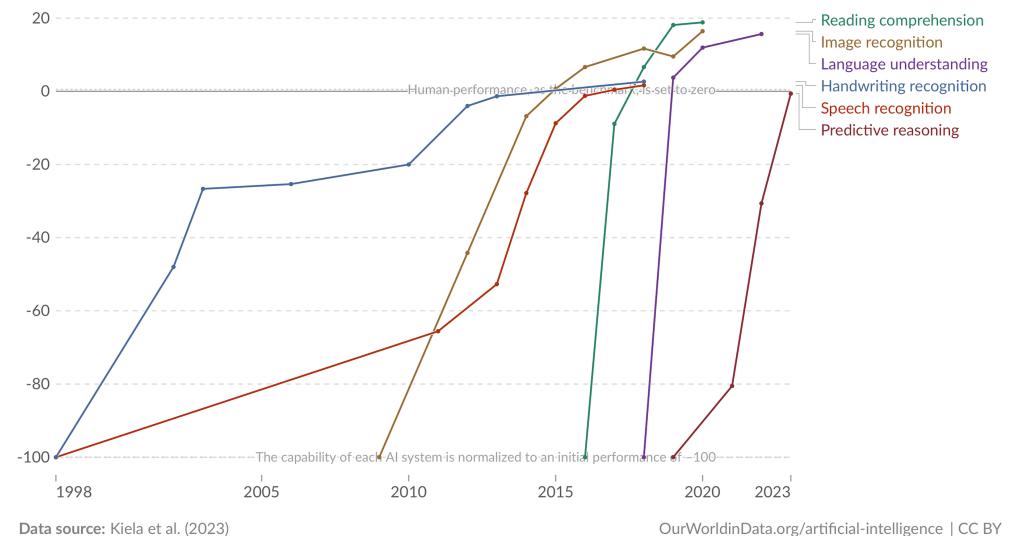


Figure 5: AI models reaching human capabilities

When it comes to strategic thinking, AI has also proven itself. Famously, AlphaGo defeated the world's best Go players in 2016 and 2017, mastering a game long considered too intuitively complex for machines.<sup>6</sup> Such examples show AI not only handling routine tasks but excelling at strategic decision-making under uncertainty. In the realm of complex problem-solving and reasoning, AI has achieved what was once unimaginable: Google DeepMind's AlphaFold solved a 50-year-old grand challenge in biology by predicting protein structures at atomic-level accuracy, a breakthrough that earned its creators a 2024 Nobel Prize. This exemplifies AI's capacity for high-level reasoning in specialized domains.

In 2023 most users were stunned by the progress of large language models handling standardized exams as ChatGPT moved from GPT-3.5 (Generative Pre-trained Transformer 3.5) to GPT-4 (Generative Pre-trained Transformer 4) in 2023.<sup>7</sup> The rapid progress is illustrated in the following figure.

<sup>6</sup>Silver, David et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529, no. 7587 (2016): 484–89, <https://doi.org/10.1038/nature16961>.

<sup>7</sup>OpenAI, “GPT-4 Technical Report,” *OpenAI Research*, 2023.

## *Grand Challenge of Growing AI Capabilities*

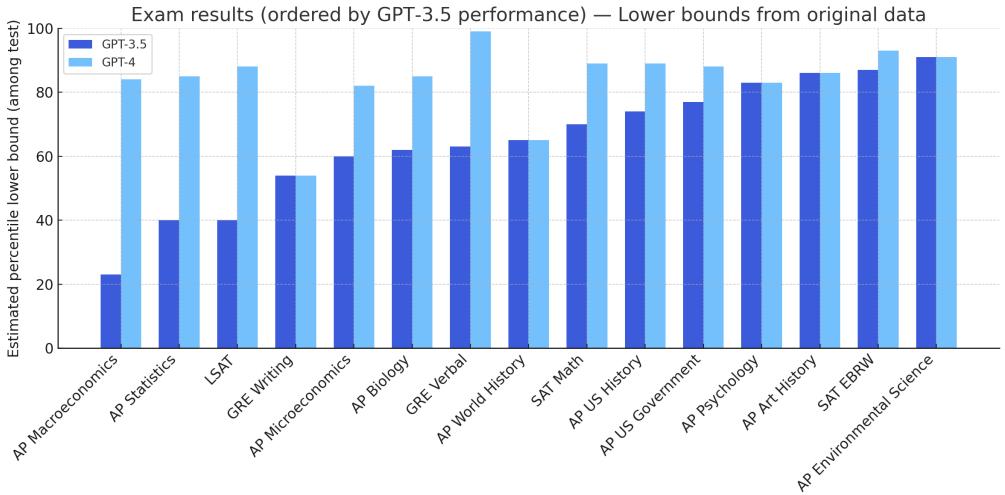


Figure 6: Rapid progress of AI model capabilities

The pace of AI advancement continues to accelerate across professional domains. A striking example of this rapid development comes from the financial sector: modern AI models, such as GPT-5, Gemini 3 Pro, or Claude Opus 4.1, have now achieved high passing scores on the Chartered Financial Analyst (CFA) exams, which are among the most rigorous professional certification tests requiring deep expertise in investment analysis, portfolio management, and financial ethics.<sup>8</sup> This achievement demonstrates how quickly AI systems are closing the gap with human experts in highly specialized, knowledge-intensive fields that were once considered the exclusive domain of trained professionals.

True intelligence isn't just about solving problems, but solving them efficiently with minimal resources. The ARC Prize is a non-profit that creates and curates human-calibrated benchmarks with tasks that are simple for people yet remain difficult for even the most advanced AI systems today. The ARC Prize Leaderboard shows not only rapid performance increases of the AI models but at the same time exponential cost decreases – note

---

<sup>8</sup>Patel, Jaisal et al., “Reasoning Models Ace the CFA Exams,” *arXiv Preprint*, 2025, <https://arxiv.org/abs/2512.08270v1>.

## Introduction: The Adaptation Crisis

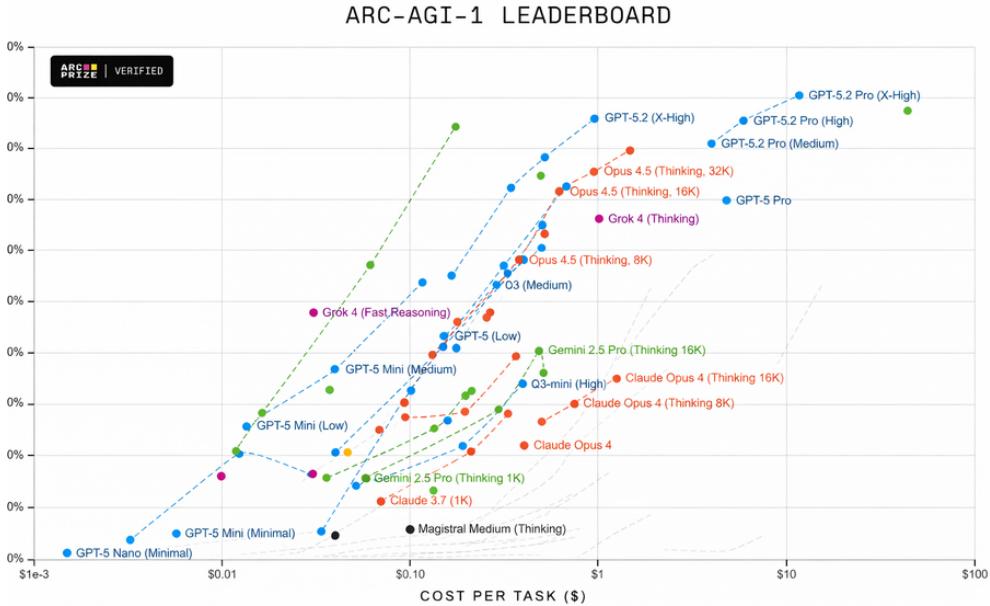


Figure 7: Exponential decrease in cost of AI models for comparable performance

that the cost axis in the following figure is logarithmic.<sup>9</sup> An ARC Prize Twitter post highlights in December 2025 that a year ago, a preview of an unreleased version of OpenAI’s o3 (High) model scored 88% on ARC-AGI-1 at est. \$4500/task, whereas the latest GPT-5.2 Pro (X-High) model scored 90.5% at \$11.64/task. This represents a ~390X efficiency improvement in one year.<sup>10</sup>

AI’s strides in creativity are equally noteworthy. Generative AI models can now produce content that would traditionally require human imagination. There are even humanoid robot artists (like “Ai-Da”) producing paintings that sell for large sums, blurring the line between human and machine cre-

<sup>9</sup>Prize, ARC, *ARC-AGI Leaderboard*, 2025, <https://arcprize.org/leaderboard>.

<sup>10</sup>Prize, ARC, *ARC Prize Twitter Post*, 2025, <https://x.com/arcprize/status/1999182732845547795>.

ativity. In computational creativity, DeepMind’s AlphaTensor discovered new algorithms for multiplying matrices faster than known methods, which mathematicians hadn’t improved on since 1969<sup>11</sup>.

Moreover, modern large language models exhibit forms of social and emotional intelligence. They can generate persuasive text and respond with emotional attunement; indeed, LLMs have become increasingly persuasive and even excel at tasks requiring emotional insight. Early studies show AI models surpassing average human scores on certain tests of emotional intelligence, indicating their potential to mimic or support social cognition. Schlegel et al. (2025) examined whether large language models can solve emotional intelligence tests.<sup>12</sup> Six LLMs (ChatGPT-4, ChatGPT-o1, Gemini 1.5 flash, Copilot 365, Claude 3.5 Haiku, and DeepSeek V3) were prompted to complete five standardized ability emotional intelligence tests measuring emotion understanding and regulation. Their performance was compared to human validation samples from the original test publications, as shown in the following figure.

Thus, it is not surprising that studies demonstrate that LLMs can generate highly persuasive content, raising important questions about their use in communication and influence contexts. Salvi et al. (2024) investigated how persuasive GPT-4 is compared to humans in real-time debates, and how personalization affects its performance.<sup>13</sup> In a randomized controlled trial with 820 participants, humans debated either other humans or GPT-4, with and without access to anonymized personal data about their opponents.

The results showed that GPT-4 with personalization increased the odds of persuading its opponent by 81.7% compared to human counterparts. Without personalization, GPT-4 still outperformed humans, but the effect was smaller and statistically insignificant. Human opponents using personalization, on the other hand, became slightly less persuasive. The findings highlight that GPT-4 can effectively exploit personal information to tai-

---

<sup>11</sup>DeepMind, “AlphaTensor,” *Nature* 610 (2022): 47–53.

<sup>12</sup>Schlegel, Katja et al., “Large Language Models Are Proficient in Solving and Creating Emotional Intelligence Tests,” *Communications Psychology*, 2025.

<sup>13</sup>Salvi, Francesco et al., “On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial,” *arXiv Preprint*, 2024.

## *Introduction: The Adaptation Crisis*

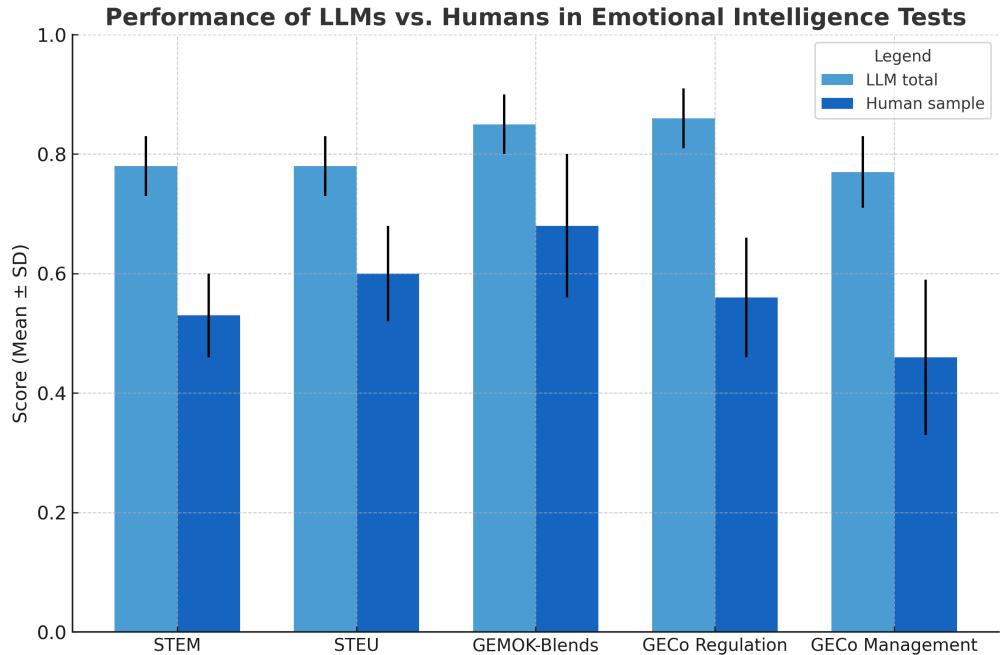


Figure 8: Emotional intelligence testing of AI models

lor arguments and outperform humans in persuasion, underscoring serious implications for online discourse, microtargeting, and AI governance.

Schoenegger et al. compared the persuasive capabilities of LLMs versus incentivized humans in an interactive quiz setting.<sup>14</sup> More than 1200 participants in the study were randomly assigned to roles as either quiz takers or persuaders. Quiz takers completed a 10-question multiple-choice quiz and were further randomized into three conditions: Solo Quiz control (20%), Human Persuasion (40%), or LLM Persuasion using Claude Sonnet 3.5 (40%). Questions were drawn from three sets: Trivia, Illusion, and Forecasting. In the persuasion conditions, each question was randomly tagged as positive (steer toward correct/trend) or negative (steer toward incorrect/away from trend), visible only to persuaders, creating both truth-

<sup>14</sup>Schoenegger, Philipp et al., “Large Language Models Are More Persuasive Than Incentivized Human Persuaders,” *arXiv Preprint*, 2025.

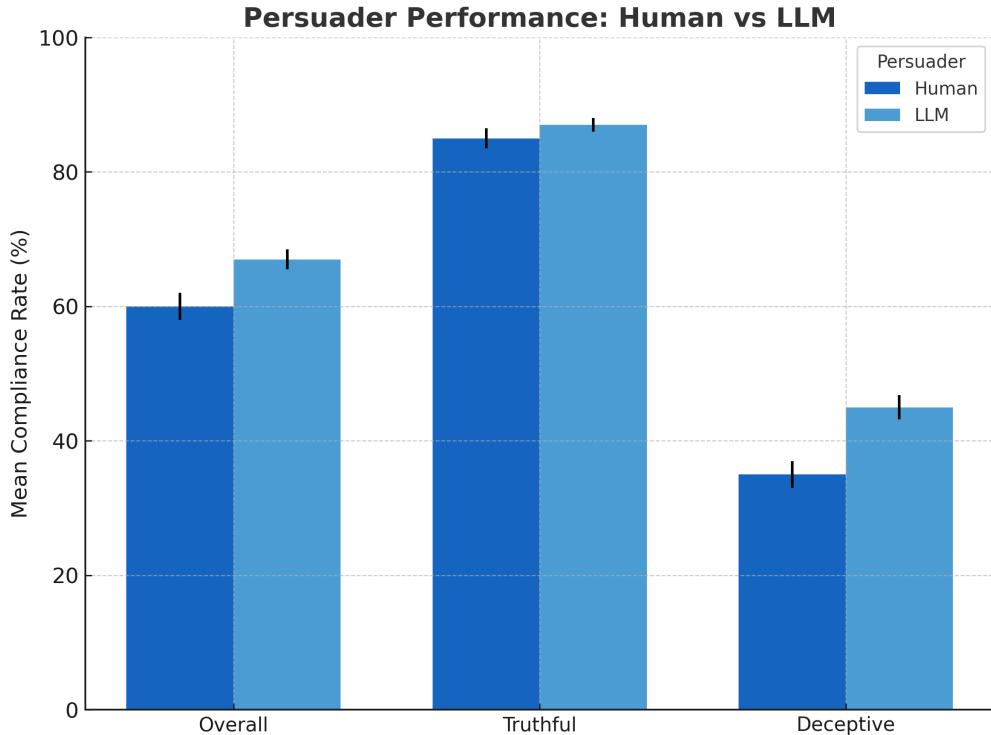


Figure 9: Persuasiveness of AI model

ful and deceptive persuasion attempts within each quiz. Quiz takers and persuaders engaged in real-time chat for 2-3 minutes per question. In the following figure, higher compliance rates indicate that the persuasion direction (truthful or deceptive) was followed.

Finally, AI is moving toward autonomous decision-making in real-world contexts. A striking (and controversial) example is the deployment of AI-guided autonomous drone swarms in military operations, highlighting how AI can perceive, decide, and act with minimal human intervention.

This breadth of accelerating advancement suggests that AI is not a single-edge technology but a multifaceted force that will drive nonlinear change in many domains. A highly relevant benchmark for tracking this progress has been released in 2025 by OpenAI that “measures model performance

## *Introduction: The Adaptation Crisis*

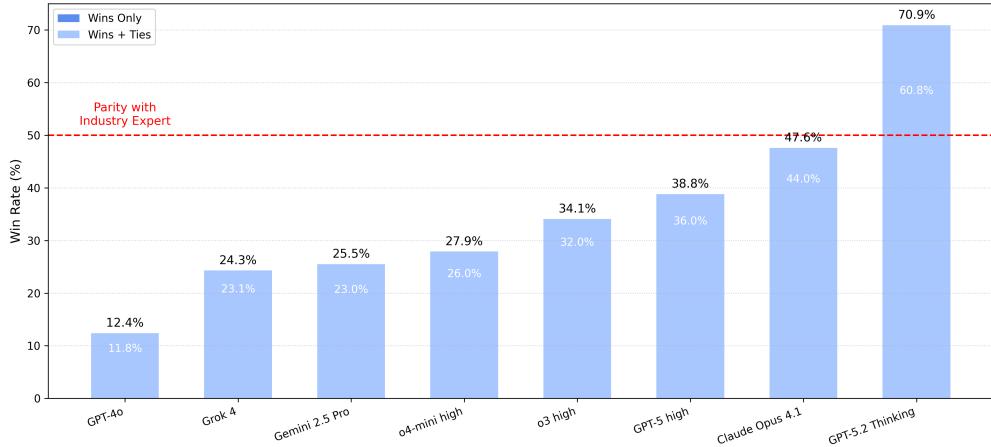


Figure 10: Real-world performance benchmarking of AI models

on tasks drawn directly from the real-world knowledge work of experienced professionals across a wide range of occupations and sectors, providing a clearer picture on how models perform on economically valuable tasks.”<sup>15</sup> This benchmark shows that modern AI models are close to parity with industry experts across sectors, as illustrated in the following figure. In fact, the performance of the OpenAI model GPT-5.2 Thinking, which was released in December 2025, is exceeding industry experts in this benchmark.<sup>16</sup>

Driving these expanding capabilities is an underlying exponential growth in AI technology and adoption. Multiple indicators underscore how quickly AI is advancing. For one, innovation is accelerating: there has been explosive growth in AI research and patents in recent years. Organizations around the world are pouring resources into AI development, as evidenced by the projection of over \$3 trillion to be spent on data centers by 2028 to fuel AI and cloud computing<sup>17</sup>. Such massive investment signals expectations

<sup>15</sup>Patwardhan, Tejal et al., “GDPVAL: Evaluating AI Model Performance on Real-World Economically Valuable Tasks,” *OpenAI Research*, 2025.

<sup>16</sup>OpenAI, *Introducing GPT-5.2*, OpenAI, 2025, <https://openai.com/index/introducing-gpt-5-2/>.

<sup>17</sup>The Economist, “What If the \$3trn AI Investment Boom Goes Wrong?” *The*

## *Grand Challenge of Growing AI Capabilities*

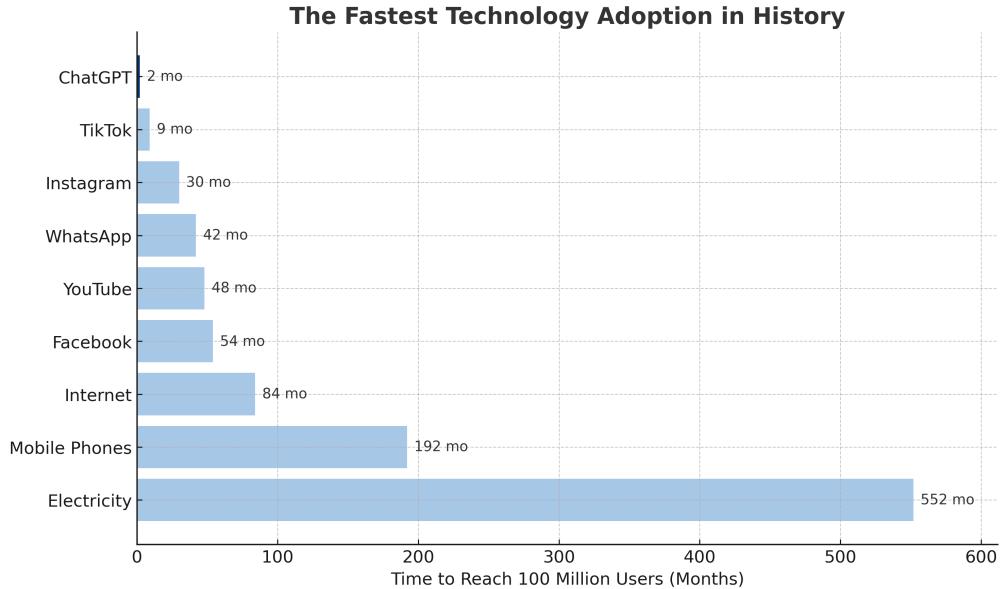


Figure 11: Adoption speed of technologies

of AI-driven transformation and illustrates the scale of commitment to AI as the next general-purpose technology.

The performance of AI systems is also beating expectations on many fronts. Breakthroughs in model architectures (like transformers) and training techniques have led to AI systems that rapidly improve on diverse benchmarks that once took years to surmount. Importantly, we are witnessing the fastest adoption curve for a new technology in history when it comes to modern AI applications. Products like advanced language models reached millions of users in a matter of days or weeks, a diffusion rate far outpacing past technologies (such as the adoption of electricity or the internet), as shown in the following figure.

All these trends – research output, investment funding, performance leaps, and user adoption – point to an exponential trajectory for AI. Yet, despite this rapid progress, the impact of AI remains uneven and task-dependent.

---

*Economist*, 2025, <https://www.economist.com/leaders/2025/09/11/what-if-the-3trn-ai-investment-boom-goes-wrong>.

## *Introduction: The Adaptation Crisis*

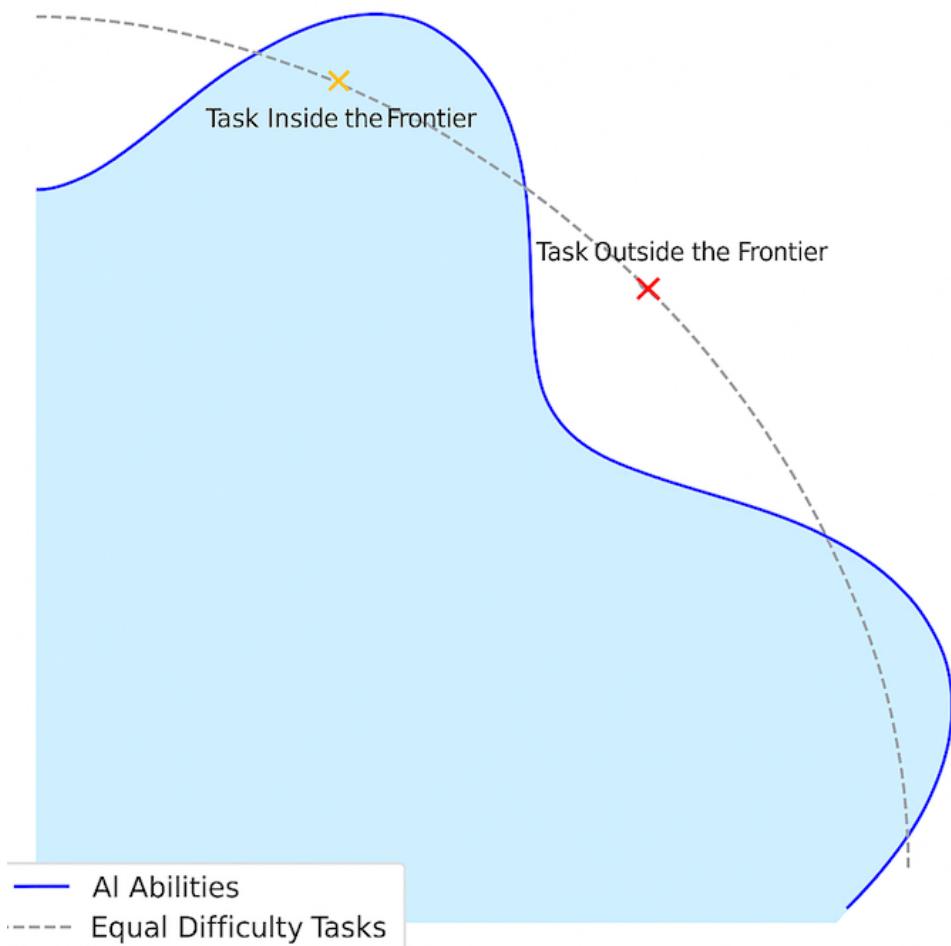


Figure 12: The jagged frontier of AI capabilities

There exists a “jagged frontier” in AI capabilities, meaning some tasks are highly susceptible to automation by AI while others remain stubbornly difficult for machines. Researchers Dell’Acqua et al. illustrated this by mapping AI’s performance across different job tasks: rather than a smooth line, the frontier is jagged<sup>18</sup>, as shown in the following figure.

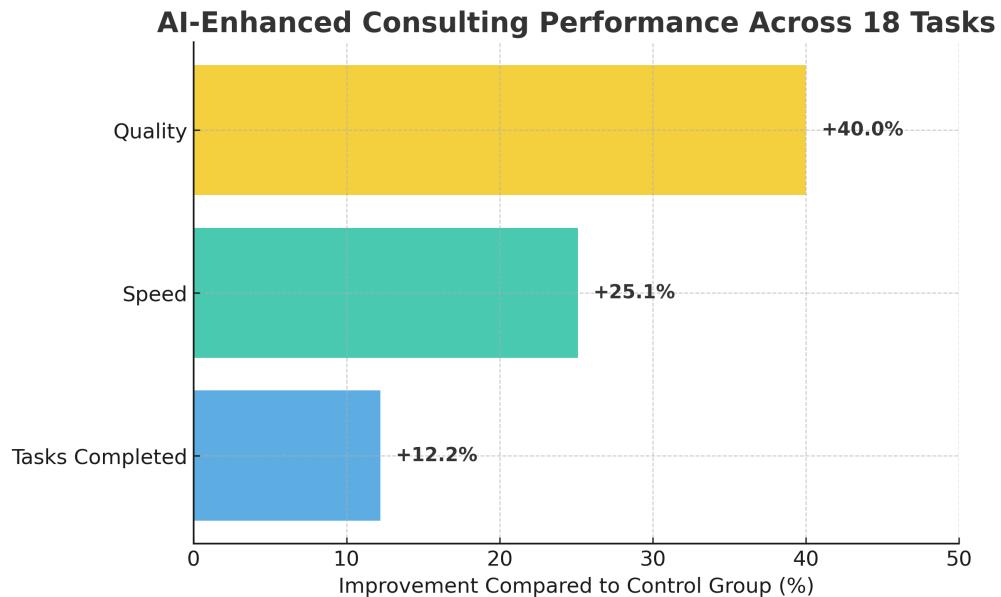


Figure 13: AI performance within current capability frontier

AI excels beyond humans in certain narrow tasks, matches humans in some, and still lags far behind in others. For example, AI may surpass human experts in data-heavy predictive tasks yet struggle with dexterous physical tasks or nuanced interpersonal interactions.

In spring 2023, 758 BCG consultants (around 7% of the firm's global individual contributors) participated in a randomized controlled trial, where consultants performed realistic consulting tasks either inside or outside the frontier of AI capability. Each consultant first completed a baseline task without AI, then was randomly assigned to one of three conditions: *no AI*, *GPT-4 access*, or *GPT-4 access plus prompt-engineering training*. Tasks included ideation and business problem-solving exercises designed by senior BCG staff to reflect real consulting work. Performance was objectively evaluated on quality, creativity, and efficiency.<sup>19</sup> The following figure shows AI performance within the current capability frontier.

## *Introduction: The Adaptation Crisis*

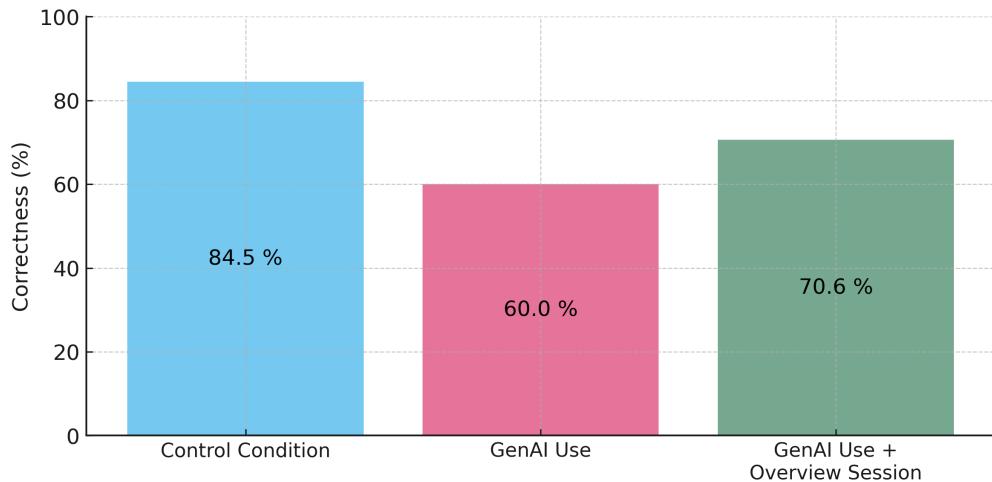


Figure 14: AI performance outside current capability frontier

Areas outside AI’s current capability frontier represent challenges that remain difficult for machines, highlighting the uneven nature of AI progress, as illustrated in the following figure.

This uneven landscape is a moving target. What is hard for AI today might become tractable tomorrow as methods improve. Ethan Mollick of Wharton encapsulates this dynamic with a wry observation: “*The AI you are using today is the worst AI you are ever going to use.*” In other words, from here on, every generation of AI will likely be better than the last, and likely by a significant margin.

This reality poses both opportunity and challenge for change management. The opportunity is that continuously improving AI can be steadily incorporated to handle more tasks, potentially boosting productivity and innovation. While creating more need for change, AI is also enabling more powerful tools for change management. For change leaders, this means the goalposts for what AI is changing and what AI can do are constantly moving – necessitating continuous learning and adaptation. As organiza-

---

<sup>18</sup>Dell’Acqua, “Navigating the Jagged Technological Frontier.”

<sup>19</sup>Dell’Acqua, “Navigating the Jagged Technological Frontier.”

## *Grand Challenge of Growing AI Capabilities*

tions will face ongoing disruption, governance and skills must evolve in step with the jagged frontier. Leaders cannot treat AI as a one-time implementation – it requires an adaptive strategy that monitors AI capability advances and regularly reassesses which processes can or should be handed off to machines. Throughout this book, we will grapple with this moving frontier, emphasizing strategies for agility and resilience so that organizations become adept at integrating new AI advancements rather than being blindsided by them.

In confronting the rise of AI, organizations and leaders often oscillate between two fundamental mindsets: one of containment and one of exploration. These dual mindsets represent different orientations toward the uncertainties and promises of AI. The containment mindset focuses on control, caution, and risk mitigation. From this perspective, AI is something to be carefully managed – even restrained – to prevent unintended consequences. Leaders with a containment outlook emphasize governance, ethics, alignment, and safety. They worry about issues like AI bias, errors, security, or the displacement of jobs. In practice, a containment approach might translate into strict AI usage policies, extensive oversight and auditing of AI systems, and contingency plans to contain AI-related risks. This mindset aligns with treating AI as a powerful but potentially dangerous tool – much like a hazardous material that must be handled with care. It reflects legitimate concerns that without proper checks, AI could cause harm (whether through flawed decisions, ethical lapses, or even rogue autonomy in extreme scenarios).

On the other hand, the exploration mindset is driven by curiosity, innovation, and embracing opportunity. Leaders in this mode view AI as a frontier to be explored and harnessed. The emphasis is on learning by doing – experimenting with AI to discover new efficiencies, products, or business models. Rather than primarily seeing what could go wrong, the exploration mindset focuses on what could go right: how AI might enable breakthroughs in understanding customers, optimizing operations, or creating value in ways previously impossible. This does not mean ignoring risks, but it puts a premium on agility and adaptation, trusting that issues can be managed as they arise. An exploratory approach might involve sandboxing new AI tools in pilot projects, encouraging teams to find

## *Introduction: The Adaptation Crisis*

creative AI applications, and actively investing in AI capabilities to build competitive advantage. Importantly, these two mindsets are not mutually exclusive – the savviest organizations cultivate a balance. They establish guardrails to address ethical and risk considerations (containment) while also incentivizing innovation and learning (exploration). In the context of change management, leaders must strike the right balance for their situation, ensuring that fear of the unknown doesn't lead to paralysis, nor that exuberance for AI leads to reckless implementation. Throughout this book, we will see examples of how adopting either mindset affects change outcomes, and why a dual approach – sometimes termed “ambidextrous” leadership – is crucial when introducing AI-driven changes. Ultimately, developing both the wisdom to govern AI responsibly and the vision to leverage it boldly will distinguish those organizations that thrive in the AI age.

While AI undeniably presents challenges, this book also positions AI as a powerful tool to aid change management itself. In particular, AI is viewed as a diagnostic and coordination technology that can reduce friction in organizational change and better align execution with goals. This perspective shifts the narrative: rather than solely being a disruptor to manage, AI can be part of the solution set for change leaders. As a diagnostic tool, AI's data-processing prowess can help leaders see patterns and insights in complex environments that would be hard to detect otherwise. For example, AI-driven analytics can assess an organization's readiness for change by sifting through employee feedback, communication data, or performance metrics to flag emerging issues. Natural language processing algorithms can analyze thousands of comments from a change survey to identify common concerns or misconceptions among staff. Network analysis algorithms may map informal influence networks to find where a change champion could have the most impact. In short, AI can assist in diagnosing barriers and enablers across the mental, social, and formal contexts of change. This kind of evidence-based diagnosis makes it possible to tailor interventions more precisely than a generic one-size-fits-all change program. Rather than relying on gut feel alone, leaders can augment their understanding with AI-derived insights, improving the odds of addressing the real pain points that threaten change initiatives. AI also functions as a coordination technology

by facilitating communication, alignment, and execution at scale. Modern organizations are complex, with many moving parts and stakeholders. AI tools – from intelligent project management software to adaptive communication platforms – can help synchronize efforts and maintain clarity. For instance, AI can be used to track progress on change initiatives in real time, sending automatic nudges or updates to keep teams aligned. It can match people to tasks dynamically (e.g., by analyzing skills profiles and availability), thus optimizing resource allocation as conditions evolve. In large transformations, AI-based dashboards might integrate data across departments to provide a single source of truth, ensuring that decision-makers at all levels operate with the same information. Additionally, emerging AI assistants or chatbots can answer employees' routine questions about the change (policies, new procedures), reducing confusion and rumor. In essence, AI can grease the wheels of change execution by making organizational response more agile and coherent. Of course, using AI in these ways requires its own change effort – people need to trust and effectively interact with these new tools. That's why this course includes hands-on exploration of AI applications in change scenarios, reflecting an ethos that to manage change in an AI age, one must directly engage with AI tools. By treating AI as an ally in diagnosing problems and coordinating solutions, leaders can amplify their capacity to drive successful change. This approach represents a significant shift from traditional change management, one that aligns with the broader goal of turning volatility and complexity into manageable, even advantageous, elements of organizational life.

## **Toward Antifragile Change Models**

Bringing together the themes above, this book aims to equip leaders with antifragile models for change management that not only withstand volatility but actually benefit from it. The term antifragile, introduced by Nassim Nicholas Taleb<sup>20</sup>, describes systems that grow stronger when exposed

---

<sup>20</sup>Taleb, Nassim Nicholas, *Antifragile: Things That Gain from Disorder* (Random House, 2012).

## *Introduction: The Adaptation Crisis*

to shocks and stressors. In the context of organizational change, an antifragile approach means designing structures, processes, and cultures that learn and improve through disruption rather than breaking down. This goes a step beyond resilience (which is about resisting shocks) – antifragile systems thrive on uncertainty and variability. By embracing the ideas of complexity and nonlinearity, the models presented in this book encourage leaders to expect the unexpected and leverage it. For example, rather than meticulously planning a multi-year change initiative and hoping nothing derails it (a fragile approach), an antifragile strategy might deploy a series of small experiments or pilots that adapt based on feedback, each one strengthening the organization’s change muscle. Volatility in the environment can then reveal what works and what doesn’t in quick iterations, making the overall change program more robust over time. Throughout the chapters, I will explore how to build such antifragile operating models that treat change not as a one-off project but as a continuous, self-improving process.

A critical enabler in these models is the intelligent use of AI to handle complexity. AI, as discussed, can serve as a diagnostic radar and a coordination engine, helping organizations stay aligned and responsive even as external conditions shift. By harnessing AI for alignment and execution, leaders can keep their teams coordinated with the overall vision, ensuring that the left hand knows what the right hand is doing at all times. This reduces the friction that often makes change brittle. At the same time, a strong emphasis on risk management and ethical governance is woven into the book’s guidance. To truly be antifragile, an organization must avoid catastrophic pitfalls; thus, responsible AI governance – covering risk, ethics, alignment with human values, and containment of unintended consequences – is a core theme. I will discuss how to institute guardrails that allow innovation to flourish without inviting chaos.

The ultimate goal of the book is to provide practical, research-grounded guidance for leading change in today’s complex, AI-augmented world. By blending the best of systems and social sciences, applying interdisciplinary lenses, and incorporating AI as a tool, the chapters ahead offer a playbook to convert uncertainty into opportunity. In this age of AI, change is nonlinear and often unpredictable – but with the right models, mindset, and

## *Toward Antifragile Change Models*

tools, you can master the adaptation crisis and even turn disruption to your advantage.

To navigate AI-driven change effectively, leaders need not become AI engineers – but they do need sufficient technical literacy to ask the right questions, evaluate claims critically, and understand what AI can and cannot do. The next chapter provides this foundation: a practical overview of how modern AI systems work, from the transformer architectures powering today’s large language models to the emerging frontier of agentic AI. With this baseline in place, we can then explore change at the individual, organizational, and societal levels with a shared vocabulary and understanding.



# **AI Fundamentals for Change Leaders**

Understanding the core concepts and architectures of AI is essential for today's change leaders. This chapter provides a basic overview of fundamental AI concepts and recent developments, with a focus on how modern AI systems work and what their capabilities and limitations are. The aim is to equip the non-technical reader with a minimal grounding in AI fundamentals – without yet diving into organizational or societal implications (those will come in later chapters). By the end of this chapter, readers should have a clearer picture of *how* AI systems function and *why* they behave as they do, forming a basis for understanding their potential applications and risks.

## **Major AI Milestones in Modern AI Development**

AI as a field has progressed through a series of remarkable milestones, each demonstrating new capabilities and inspiring further innovations. The following timeline shows major AI technology milestones that have shaped the current landscape of AI:

- **1997 – Deep Blue:** IBM's Deep Blue chess program defeated world champion Garry Kasparov, marking the first time a reigning chess champion lost to a computer.<sup>1</sup> Deep Blue achieved this through brute-force search and domain-specific symbolic rules and heuristics

---

<sup>1</sup>Kasparov, Garry, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins* (PublicAffairs, 2017).

rather than learning. It signaled that machines could compete with human experts in complex, rule-based games.

- *2012 – AlexNet:* For the first 50 years of AI, the dominant approach was the Logic-Inspired Paradigm, which posited that the essence of intelligence is reasoning achieved through manipulating symbolic rules and expressions in an unambiguous language. In contrast, the Biologically Inspired Approach, favored by figures like Alan Turing and John von Neumann, centered intelligence on learning the strengths of connections in a neural network. The scientific shift towards the biological approach fundamentally occurred in 2012, when a deep neural network named AlexNet dramatically cut the error rate in the ImageNet computer vision competition, achieving a top-5 error rate of 15.3% compared to the second-place entry's 26.2%.<sup>2</sup> This victory of a deep learning model over earlier approaches showed the power of training multi-layer neural networks on large datasets, sparking the modern resurgence of AI. AlexNet's success demonstrated that deep neural networks (DNNs) can automatically learn rich representations from data when given enough examples and computing power.
- *2016 – AlphaGo (and 2017 – AlphaZero):* Google DeepMind's AlphaGo program defeated the world's top Go players using a combination of deep neural networks with Monte Carlo tree search and reinforcement learning.<sup>3</sup> Go was long considered a grand challenge for AI due to its immense complexity. AlphaGo trained on human games and then improved via self-play, showcasing how AI could tackle intuitive, strategic domains. Its successor AlphaZero (2017) went further by learning superhuman play in Go, chess, and shogi without any human examples, relying only on self-play reinforcement learning. These systems illustrated the potential of combining learned neural network policies with traditional search techniques.
- *2017–2020 – Transformer Models* (e.g., OpenAI GPT or Anthropic

---

<sup>2</sup>Krizhevsky, Alex et al., “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems* 25 (2012): 1097–105.

<sup>3</sup>Silver et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search.”

Claude or Google Gemini series): The introduction of the Transformer architecture<sup>4</sup> in 2017 revolutionized natural language processing. OpenAI’s GPT (Generative Pre-trained Transformer) models, beginning around 2018 and exemplified by GPT-3 in 2020, showed that *scaling up* transformer networks (with more layers, more training data, and more parameters) produces startling language capabilities. These models are pretrained on vast amounts of data from the internet and can then be adapted to countless tasks. The advent of GPT demonstrated that a single model could generate coherent paragraphs of text, write code, or answer questions on almost any topic – a generality not seen before. The Transformer architecture became the foundation of most state-of-the-art AI systems.

- *2020 – AlphaFold:* Another DeepMind breakthrough, AlphaFold, solved the 50-year-old grand challenge of protein folding by predicting 3D protein structures from amino acid sequences.<sup>5</sup> AlphaFold’s deep network was trained on vast genomic and structural data and outperformed all prior methods, demonstrating how AI can advance scientific frontiers. Its success had enormous implications for biology and medicine, proving that AI can achieve expert-level understanding in highly complex, specialized tasks.
- *2022 – AlphaTensor:* In 2022, DeepMind unveiled AlphaTensor<sup>6</sup>, a system that discovered novel algorithms for multiplying matrices more efficiently. AlphaTensor used a combination of deep learning and game-like search techniques to improve upon decades-old human-designed algorithms. This was a milestone showing that AI can invent new algorithms, effectively expanding the realm of what machines can do beyond just automating human knowledge – they can now create new knowledge in mathematics and computer science.
- *2022–2023 – Diffusion Models for Generative Media:* The years 2022–

---

<sup>4</sup>Vaswani, Ashish et al., “Attention Is All You Need,” *31st Conference on Neural Information Processing Systems*, 2017.

<sup>5</sup>Jumper, John et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596, no. 7873 (2021): 583–89, <https://doi.org/10.1038/s41586-021-03819-2>.

<sup>6</sup>DeepMind, “AlphaTensor.”

2023 saw AI models generating images (and other media) with unprecedented fidelity. OpenAI’s DALL·E 2<sup>7</sup> and open-source Stable Diffusion<sup>8</sup> (both released in 2022), as well as Midjourney V5<sup>9</sup> (2023), are diffusion models that can create detailed images from text descriptions. These generative models learned from millions of images to translate textual prompts into novel images, enabling users to produce artwork, photorealistic illustrations, and more. This period established generative AI as a mainstream phenomenon, extending beyond text into images, audio, and video.

- *2023 – ChatGPT and Conversational Agents:* In late 2022, OpenAI released ChatGPT (based on GPT-3.5, later GPT-4 and GPT-5), a conversational AI that garnered worldwide attention for its ability to engage in dialogue, explain concepts, and follow instructions. 2023 then saw the integration of these models into products like Bing Chat (Microsoft) and the rise of frameworks like LangChain (which chains language model queries with tools/data). These systems introduced tools and real-world connections, allowing AI to access the web, run code, or use third-party services under controlled conditions. As a result, generative AI became interactive and multimodal (e.g., GPT-4o can handle images and text together), moving closer to being general-purpose assistants.
- *2024–2025 – Reasoning & “Agentic” AI Systems:* An emerging frontier in AI is the development of models that can perform reasoning in multiple steps and take actions in the world. Models like OpenAI’s GPT-4 “with tools” and DeepMind’s prototypes (sometimes code-named “O1” or DeepSeek-R1) incorporate reasoning and tool use. Anthropic’s models in the Opus-series also explore this space. These efforts aim to enable AI to not just generate text, but also solve complex problems by planning, using external tools or knowledge

---

<sup>7</sup>Ramesh, Aditya et al., “Hierarchical Text-Conditional Image Generation with CLIP Latents,” *arXiv Preprint*, 2022, <https://arxiv.org/abs/2204.06125>.

<sup>8</sup>Rombach, Robin et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 10684–95.

<sup>9</sup>Midjourney Inc., “Midjourney AI,” 2023, <https://www.midjourney.com/>.

bases, and even coordinating sub-tasks autonomously. We will discuss such agentic capabilities later in this chapter. Another notable development is AlphaEvolve, introduced around 2024–2025, which represents a step toward self-improving AI. AlphaEvolve is described as an evolutionary coding agent that can autonomously improve its own code and devise new algorithms, pushing the envelope of AI-driven research and development. We'll cover AlphaEvolve in detail in a later section, but it stands as a milestone indicating how AI might eventually improve upon itself without direct human coding.

Each of these milestones marks a leap in what AI systems can do – from mastering games and perception to generating creative content and handling language, and now to exhibiting rudimentary planning and self-improvement. For change leaders, appreciating this trajectory helps in understanding the paradigm shifts in AI capability over time. Next, we delve a bit deeper into the two central technologies and concepts that underpin these breakthroughs: Deep Neural Networks and the Transformer architecture.

## **Deep Neural Networks as Core Building Block**

Modern AI systems are largely built on *deep neural networks (DNNs)*, which are function approximators inspired by the human brain's neural structure. A deep neural network is essentially a tunable mathematical function that maps inputs to outputs through a series of weighted transformations. In practical terms, you feed the network some data (e.g., an image, a piece of text, or a set of numeric features), and it processes this data through multiple layers of interconnected “neurons” (computational units), each layer transforming the data based on parameters (weights). By adjusting these weights, the network can learn to produce desired outputs. Thus, a large language model’s knowledge is stored in those weights, not word tables, which makes the knowledge relational as Geoffrey Hinton, winner of the 2024 Nobel Prize in Physics for his contribution to the development of learning in artificial intelligence, has pointed out.

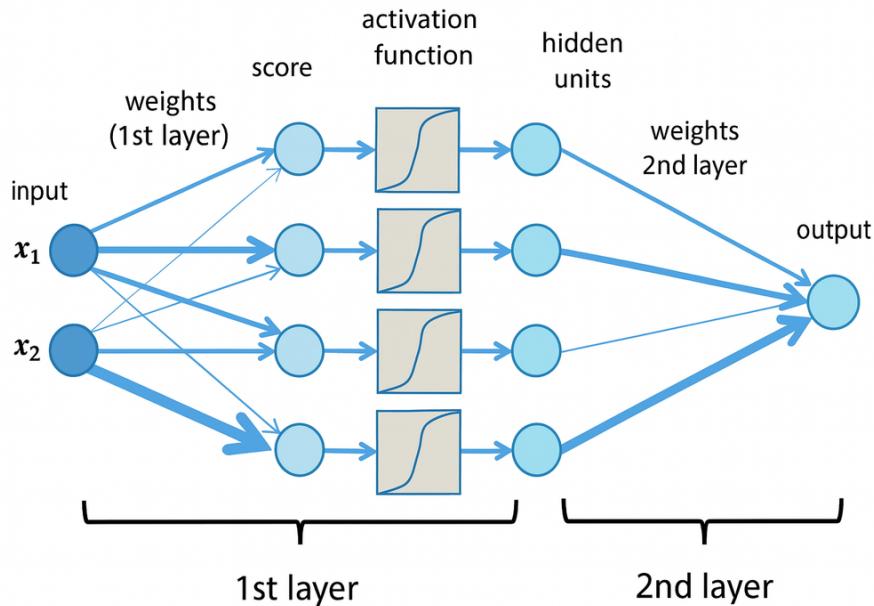


Figure 15: Most basic Deep Neural Network

*Architecture:* A simple way to picture a neural network is as layers of neurons: an input layer (taking the raw data), multiple hidden layers that perform computations, and an output layer that produces a prediction. Each connection between neurons has a weight that amplifies or dampens the signal. What makes a network “deep” is having many layers (dozens or even hundreds in modern networks), enabling very complex functions to be modeled. Different architectures exist for different data types – for example, convolutional neural networks (CNNs) are structured for image data, recurrent neural networks (RNNs) were traditionally used for sequences like text, and so on. But fundamentally, they all consist of layers of weighted sums passed through nonlinear activation functions. The basic structure is illustrated in the following figure.

One important concept is the use of *embeddings* to handle discrete or high-dimensional inputs. An *embedding* is a learned numeric representation of data. For instance, words in a text are typically converted into vectors

(lists of numbers) such that similar words have similar vectors. This is done by an embedding layer at the network’s input, which maps each discrete token (word or sub-word) to a dense vector of real numbers. Similarly, for categorical inputs or other complex inputs, using an embedding allows the neural network to work with continuous-valued vectors that capture semantic similarity. Embeddings are crucial because they allow the network to measure likeness or relationships in the input space (e.g., “king” and “queen” end up with vectors that relate in similar ways as “man” and “woman” do in an idealized 2-dimensional meaning space), as visualized in the following figure.

Note that modern LLMs use embeddings in more than 3,000 dimensions.

*Training Process:* How does a neural network actually learn? Learning is accomplished through a process called training, which typically uses a method named *backpropagation* together with an optimization algorithm (like stochastic gradient descent or its variants). The training loop can be summarized as “predict → compare → correct”:

- *Predict (Forward Pass):* Take a batch of training examples, feed them through the network to get the current predictions.
- *Compare (Loss Computation):* Compare the predictions to the true target values using a loss function (also called a cost function). This is a mathematical function that quantifies the error or difference between what the network predicted and what the correct output should be. For example, in a classification task, the loss might measure how far the predicted probability distribution is from the true distribution (using cross-entropy loss).
- *Correct (Backward Pass and Optimization):* Calculate the gradient of the loss with respect to each weight in the network (this is done via backpropagation, which efficiently computes how a small change in each weight would affect the loss). Then, adjust the weights in the direction that lowers the loss (the optimizer determines the step size for these adjustments, known as the learning rate). This step “nudges” the network parameters to reduce error on those examples.

Repeat this process for many iterations (passes through the dataset). Each

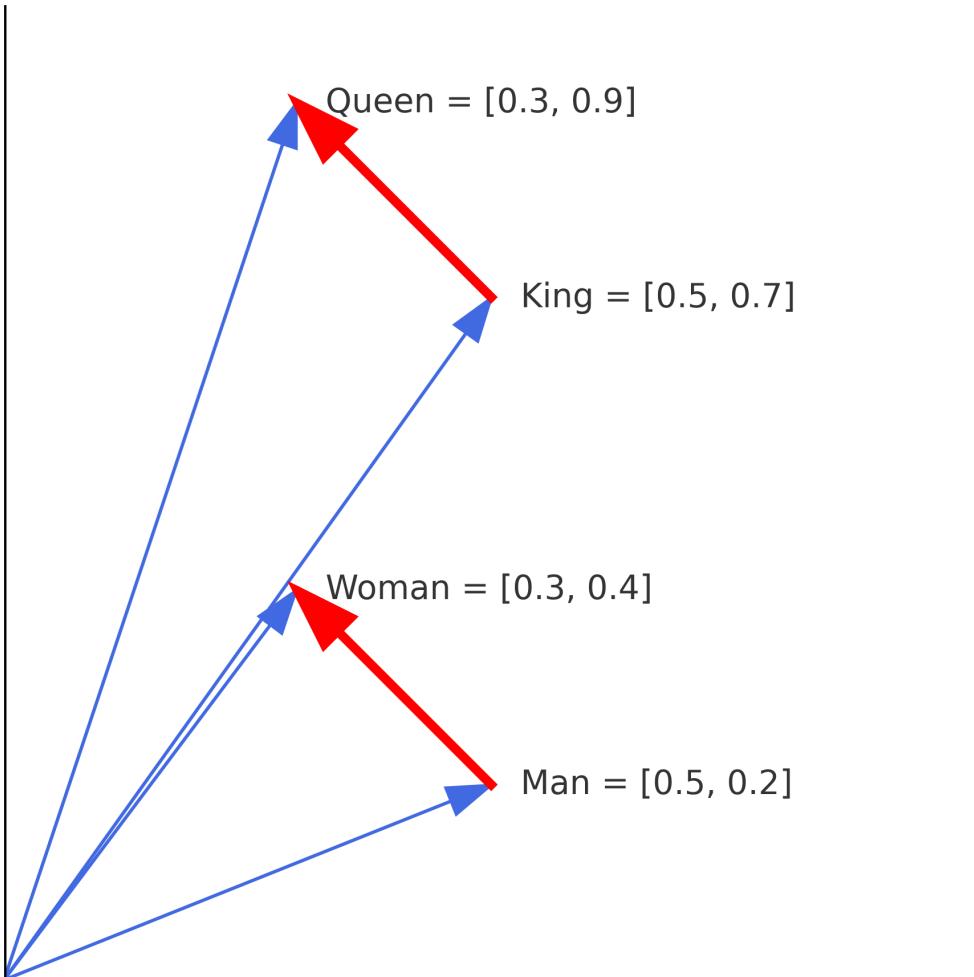


Figure 16: Spatial similarities of meaning with the help of embeddings

full pass through the training dataset is called an epoch. Over many epochs, the network's predictions should gradually become more accurate as it minimizes the loss function. It's common to use a training set of data to update weights, while holding out a validation set to tune high-level choices (like how many epochs to train, or to prevent overfitting by early stopping), and finally measuring performance on an independent test set to get an unbiased estimate of how well the trained model generalizes to

new data.

This training pipeline enables the network to learn from examples. It is a form of supervised learning when we provide explicit target outputs. In reinforcement learning (as used in AlphaGo or AlphaZero), the principles are similar but the “reward” from game outcomes acts as the feedback signal rather than an explicit label for every example.

*Hyperparameters:* The behavior of the training process and the final performance of a DNN depend on various hyperparameters – these are the “dials” that practitioners tune, as opposed to the network’s internal weights which are learned. Key hyperparameters include the learning rate (how big each weight update step is), batch size (how many examples are processed at once before updating weights), the number of layers or the number of neurons per layer (which determine the model’s capacity), and the number of training epochs. For example, a higher learning rate might speed up learning but risk overshooting optimal values (causing divergence), whereas a too low learning rate may make training painfully slow or get stuck in a suboptimal state. The number of layers and units influences how complex a function the network can represent – more layers typically allow learning more abstract features (hence the term *deep* learning), but also make training more challenging and data-hungry. Batch size affects the stability of training updates and computational efficiency (larger batches give more stable gradient estimates but require more memory). Tuning these hyperparameters is often as much art as science, typically done via experimentation or automated search, since they can significantly impact model performance.

In summary, deep neural networks learn by adjusting internal weights through repeated exposure to data, guided by a feedback signal (loss). This simple mechanism – when scaled up with big data and compute – has yielded astonishing results across many domains, and it underlies most of the AI systems discussed in this chapter.

## The Transformer Architecture

While deep neural networks have been around for decades, the *Transformer architecture* has been the single most impactful innovation in AI in recent years. Transformers fundamentally changed how AI systems handle sequential data (like language) and enabled the training of extremely large-scale models that drive today's AI breakthroughs.

The complexity of modern AI architectures, particularly Transformer models, creates a significant barrier to comprehension for non-experts. Stephen Wolfram's 2023 essay "What Is ChatGPT Doing ... and Why Does It Work?" provides an accessible exploration of the mechanisms underlying large language models like ChatGPT.<sup>10</sup> Wolfram explains how such models are trained on vast amounts of text data to predict the next word in a sequence, creating responses that mimic human language patterns. He delves into how this predictive process arises from complex statistical relationships captured within high-dimensional mathematical structures, specifically, neural networks with billions of parameters optimized through massive computational training. Rather than explicitly encoding rules of grammar or knowledge, the model develops an implicit understanding of language and reasoning through exposure to examples, allowing it to produce coherent, contextually appropriate text. Cho et al. have designed an interactive visualization tool with their "Transformer Explainer,"<sup>11</sup> to make the opaque mechanisms of text-generative models accessible to a broader audience at <https://poloclub.github.io/transformer-explainer/>. The tool focuses on the GPT-2 model and enables users to visually examine how input text transforms through the model's layers to predict next tokens, while also allowing real-time experimentation with key parameters like temperature to understand prediction determinism, as shown in the following figure.

---

<sup>10</sup>Wolfram, Stephen, *What Is ChatGPT Doing ... And Why Does It Work?*, 2023, <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>.

<sup>11</sup>Cho, Aeree et al., "Transformer Explainer: Interactive Learning of Text-Generative Models," *arXiv Preprint*, 2024, <https://arxiv.org/abs/2408.04619>.

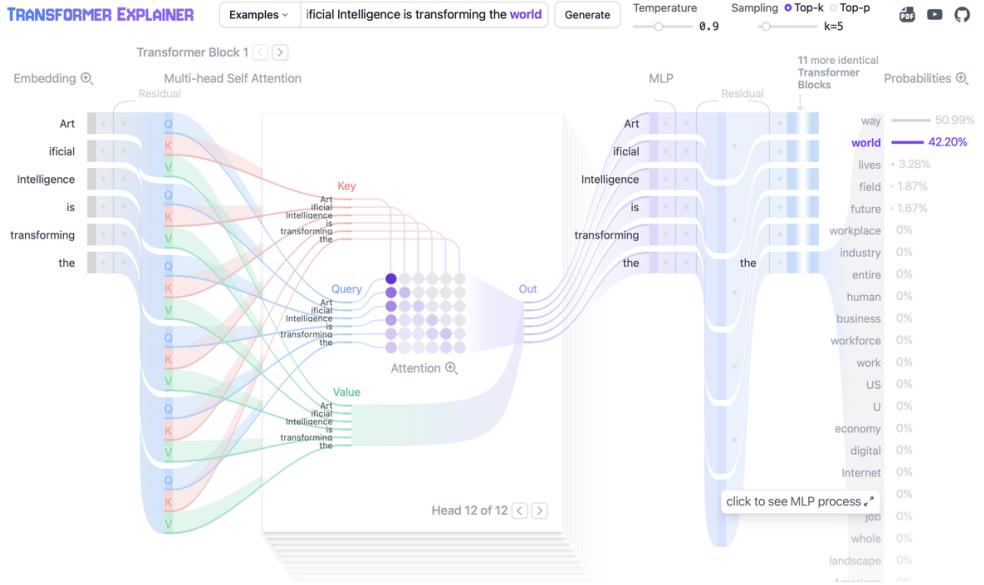


Figure 17: Exploring the transformer interactively

Transformers are providing a versatile architecture that underlies most state-of-the-art AI models today. For a change leader, understanding transformers is crucial because they explain why AI suddenly became so powerful around 2017–2020 and how one model can seem to do so many different things. The following table summarizes strengths and limitations of transformer-based LLMs.

Table 1: Strengths and Limitations of Transformer LLMs

Strengths of Transformer LLMs	Limitations of Transformer LLMs
Broad Knowledge	No True Understanding
Fluent Generation	Prompt Sensitivity & Prompt Attacks
Few-Shot Adaptability	Biased & Toxic Outputs
Reasoning & Pattern Recognition	Lack of Consistency & Reproducibility
	Limited Length & Memory

Strengths of Transformer LLMs	Limitations of Transformer LLMs
	No Built-in Verification or Truth-Checking

Let's break down *what* the transformer architecture is and *why* it's so powerful:

- *Attention Mechanism (Self-Attention)*: At the heart of the transformer is the concept of attention. Earlier sequence DNNs like RNNs (Recurrent Neural Networks) processed words in order (one timestep after another), which made it hard to capture long-range dependencies. Transformers instead use self-attention (a mechanism that allows the model to dynamically weight the relevance of every other word in the sequence when processing a given word), which enables the model to focus on relevant context. In simple terms, for each input token (say a word or sub-word piece), the model can attend to other tokens to decide what is important in the context. This means if you have a sentence, the transformer can learn which other words are related or important to the current word's meaning. The phrase "the relevant parts of the context of a token" is how one might describe this: the model builds representations that emphasize relevant context and ignore irrelevant context for each position. This attention mechanism is not limited by distance in the sequence – a word at the beginning can easily attend to a word at the end. As AI researcher Andrej Karpathy put it, attention gives the model a sort of memory of the entire sequence, enabling it to capture relationships (like long-range grammatical dependencies in a sentence or patterns in code) that older models might miss.
- *Parallel Processing and Efficiency*: Transformers do away with the need to process sequences one step at a time. Thanks to attention, the model can process all tokens in parallel during training, because each layer of the transformer looks at all tokens simultaneously (using matrix operations that leverage modern hardware). This is a huge efficiency gain: it allows training on vast datasets and dramatically

reduces training time compared to older recurrent architectures. In other words, transformers scale brilliantly – they can soak up far more data and computation, which is one reason models like GPT-3 (with 175 billion parameters) were feasible. With parallelization, one can utilize Graphical Processing Units (GPUs) or Tensor Processing Units (TPUs) to their full extent by feeding thousands of tokens at once, whereas earlier DNNs would have to step through them sequentially.

- *Unsupervised Pretraining (Self-Supervised Learning)*: Another game-changing aspect of the transformer revolution is how these models are trained. Rather than needing labeled data for every task, transformers are often trained in an unsupervised or self-supervised manner on huge corpora of raw data. For example, GPT models are trained to simply predict the next word in a sentence, given all the previous words. This doesn't require an external label – the next word in the text *is* the label. By doing this prediction task (a language modeling objective) on billions of sentences, the model learns the structure of language, facts about the world, and other patterns in text. This unsupervised pretraining means one can leverage orders of magnitude more data than in supervised settings, and the resulting model acquires a broad, general understanding. The model can then be fine-tuned or prompted for specific tasks, a paradigm that has proven extremely effective.
- *One Architecture, Many Modalities*: The transformer design is quite general-purpose. Initially used for text (with tokens being words or characters), it has since been applied to images, audio, and more. In vision, for instance, the Vision Transformer (ViT) breaks images into patches and treats them like “visual tokens” to feed a transformer. The same core design can model different data types by appropriate tokenization. This means a common architecture can underlie language models, image generators, protein structure predictors, etc. Transformers thus provide a unified architecture for AI, where once you understand the basics, the differences lie mainly in how the data is encoded as input.

- *Pretrain Once, Adapt Many Times:* The transformer architecture has popularized the paradigm of pretraining a large model on general data, then adapting it to many tasks. Instead of training a new model from scratch for each task, one large pretrained model can be used as a foundation. It can either be fine-tuned with additional training on a smaller, task-specific dataset, or it can be used as-is with clever prompting (more on that later). This approach is efficient because the general knowledge learned in pretraining can be repurposed. For example, a single model like GPT-4, pretrained on the internet, can be adapted to answer medical questions, generate legal summaries, perform customer service, etc., with relatively small task-specific adjustments.
- *Finite Context Window:* One limitation of transformers is that they have a fixed-size context window – they can only pay attention to a certain number of tokens at once (determined by the model’s design and computational constraints). Early models like GPT-3 had a context window of around 3,000 tokens, meaning they could take into account roughly 3,000 parts of words of recent text when generating the next token. Newer models have expanded this (GPT-4 can handle 32k tokens, and modern models are pushing this further up to 2 million tokens of context). A larger context window lets the model remember and use more information (for example, an entire chapter of a book, or long conversations). However, longer contexts increase memory and compute costs, and the model may still miss facts outside its window. This is why very long documents or conversations sometimes pose challenges – the model might “forget” earlier details if they exceed the window or if not summarized. Managing and extending context is an active area of research (including using retrieval, which we’ll discuss under transformer extensions).
- *Scaling and Performance:* A notable empirical finding with transformers is that performance tends to improve as models get bigger (in terms of parameters and data) in a somewhat predictable manner. These are called scaling laws – errors decrease roughly as a power-law as you scale up model size, dataset size, and compute. The success of GPT-3 and beyond has been largely a story of scale: more layers,

more neurons, more data, and more training time yielded qualitatively new capabilities (like writing coherent articles or sophisticated coding). Of course, this can't continue indefinitely due to practical limits, but scaling has so far been a key to higher performance, which is why companies have been racing to train ever larger models. We will see later how some extensions aim to get smarter without simply going bigger, but scaling has been a dominant theme since the advent of transformers.

Transformer LLMs have demonstrated unprecedented capabilities in generating and understanding language. Some of their strengths include:

- *Broad Knowledge:* Because they are trained on massive datasets (encompassing books, articles, websites, etc.), LLMs have at least surface-level knowledge of many domains. They can discuss a wide range of topics and often recall factual information (especially common or popular facts) from their training data.
- *Fluent Generation:* These models can produce human-like, coherent text. They excel at mimicking the style and structure of human writing. This makes them useful for drafting documents, answering questions, writing code, creating marketing copy, and more – tasks where natural language output is needed.
- *Few-Shot Adaptability:* Even without explicit training on a new task, LLMs can often be prompted to perform the task by giving instructions or examples in the prompt. This prompt-based learning (more on it below) allows a single model to be used for many purposes, reducing the need for task-specific models.
- *Reasoning and Pattern Recognition (to a degree):* LLMs can perform surprising forms of reasoning. For instance, they can do step-by-step math or logic puzzles if prompted correctly (“chain-of-thought” prompting, covered later). They also can write code and debug with some competency. While they don’t truly understand in a human sense, the vast training seems to let them approximate reasoning by pattern recognition across many examples. They have been called “stochastic parrots” by some critics – emphasizing that they predict

rather than comprehend – yet in doing so, they sometimes generalize or recombine knowledge in useful ways.

Despite these strengths, transformer LLMs also have critical limitations and weaknesses that stem from how they are built and trained:

- *No True Understanding or Ground Truth:* An LLM like GPT-4 does not truly know facts or have an inherent concept of truth – it generates the most statistically likely continuation of the prompt based on its training. As a result, it can fabricate information (a phenomenon known as “hallucination”), stating incorrect facts with confidence. The model has no built-in mechanism to distinguish truth from falsehood; it doesn’t “know” if an answer is correct. These models are optimizing for plausible-sounding outputs, not factual accuracy. This is why an LLM might assert something that is completely untrue, especially if the prompt somewhat implies it or if the correct answer is obscure. An LLM will often sound confident even when it’s completely wrong.<sup>12</sup> It might even make up sources or statistics. This can be dangerous if users take the output at face value. Without tools to verify facts (like cross-checking a knowledge base or performing calculations), the model can mislead. As Stephen Wolfram has pointed out, *symbolic or rule-based computation* (like what Wolfram Alpha does) is precise and reliable, whereas these statistical LLMs are intuitive and fluent but can be unreliable on factual or mathematical precision. Integrating LLMs with tools that can do formal verification or computation is one way being explored to address this limitation.
- *Sensitivity to Prompts (and Prompt Attacks):* What an LLM outputs is highly sensitive to the phrasing of the prompt. A slight rewording can change the answer. This makes them brittle in some cases – a user might get a correct answer only if they phrase their question in just the right way. Moreover, malicious actors can exploit this sensitivity via prompt injection attacks. A prompt injection is

---

<sup>12</sup>Kalai, Adam Tauman et al., “Why Language Models Hallucinate,” *arXiv Preprint*, 2025.

a technique where a user crafts an input that causes the model to ignore prior instructions or produce unintended output. For example, if a model is instructed to not reveal certain information, a cleverly designed prompt might trick it into doing so. These “jailbreak” prompts exploit the model’s tendency to follow the most recent or strongly worded instruction, exposing a security vulnerability in how we control AI outputs. This is an emerging concern as we deploy LLMs in user-facing applications.

- *Biased and Toxic Outputs:* LLMs learn from data that inevitably contains human biases (cultural biases, stereotypes, etc.) and various forms of toxic or harmful content. As a result, models will reflect and sometimes amplify biases present in their training data. They might produce discriminatory or offensive outputs if not carefully controlled. Even when not overtly offensive, subtle biases in how questions are answered (e.g., associating certain professions or traits with a particular gender or ethnicity) can be problematic. Mitigating bias in AI outputs is an active area of research and a crucial consideration for ethical AI deployment.
- *Lack of Consistency and Reproducibility:* Each time an LLM generates a response, there’s an element of randomness (especially creative variety). Even with temperature low (for more deterministic output), the generation process in very large models can be nondeterministic in practical use. This means you might not get the exact same answer every time, and re-running the same prompt could yield slightly different wording or even a different answer. For applications that need reliable, consistent output, this is a limitation. Moreover, evaluation of LLM outputs can be tricky – success isn’t just one right answer as in a math problem; it’s often subjective (what counts as a “good” essay or a helpful explanation?). Thus, ensuring an LLM performs reliably requires careful prompt design and often human oversight.
- *Limited Length and Memory:* As mentioned earlier, models have a finite context window. If you ask an LLM a question that requires using a very large document as reference, it might not handle it well all at once. It might ignore parts of the input if it’s too long or

lose track of earlier parts of a conversation. They also don't have long-term memory of past conversations by default (outside the session window), unless specifically augmented with external storage. This means if you want an AI that remembers user preferences over months, a plain LLM won't do that unless engineered with additional infrastructure.

In practice, many of these limitations are addressed by putting guardrails around LLMs: using human feedback to fine-tune them to refuse certain queries, filter content, or rephrase answers to be more correct; and by augmenting them with retrieval of factual information or with calculators, etc. Still, it's crucial for leaders to remember that current AI models are not infallible or fully autonomous problem-solvers – they are powerful prediction engines with significant blind spots.

## **Different Paths to Model Adaptation**

One of the powerful features of modern AI, especially large pre-trained models, is that they can be adapted to perform specific tasks or follow particular instructions. There are two major ways to adapt a pre-trained model to new tasks or domains: (1) **Fine-tuning** (updating the model's internal weights through additional training on domain-specific data), and (2) **Prompting** (using carefully crafted instructions and examples in the input to guide the model's behavior without modifying its weights, sometimes with a technique called few-shot prompting or just good prompt engineering). Both have their uses, benefits, and trade-offs. For change leaders, understanding the difference is crucial because it shapes how you deploy AI in different scenarios.

### **Fine-tuning**

*Fine-tuning* involves taking a pre-trained model (like a large language model trained on general text) and then doing further training on a new,

narrower dataset that is focused on your task. This process updates the model’s weights so that it better performs the desired task.

Essentially, fine-tuning is like giving the *model additional experience* in a specific area. You present the model with many examples of the task you care about (with the correct outputs), and adjust the weights via training so that the model’s performance on those examples improves. Unlike initial pretraining (which might be unsupervised), fine-tuning is usually supervised or uses a specialized objective. For instance, you might fine-tune a general language model on a dataset of customer support conversations so that it learns to output responses in a helpdesk context.

Suppose you are developing a customer service chatbot. You have thousands of transcripts of support tickets or chats, including the issues and the resolutions. By fine-tuning a model on this data, the model can learn the common questions, the appropriate answers, and the company’s style of responding. After fine-tuning, when deployed, the model will be pre-informed about common support issues and will likely respond with higher accuracy and domain-specific knowledge. In essence, it has become a specialist, not just a generalist.

Fine-tuning can significantly increase accuracy or performance on a target task, especially if the task has jargon or requires knowing details that the general model might not have focused on. It also reduces the need to cram context into the prompt every time – since the model’s weights have been adjusted, it “knows” the domain. Fine-tuning can also allow customization: you can imbue the model with a certain tone or policy (for example, always being extra polite, or never answering certain types of questions) by including those patterns in the training data. Once fine-tuned, using the model is straightforward – you don’t need complex prompt engineering for it to understand the task.

Fine-tuning requires having a suitable dataset and the computational resources to do the training. It can be time-consuming and not everyone has the data for their specific needs. Also, a fine-tuned model is a new model – it’s a version that might drift from the original, and if the fine-tuning data is not good quality, the model could even get worse in some aspects (e.g., it might forget some of its more general knowledge, a phenomenon

called catastrophic forgetting, if not done carefully). Another consideration: every time your task changes, you might need a new fine-tune. This isn't feasible for rapidly changing needs or for end-users who can't train models.

Even after pretraining and fine-tuning, AI models may not do exactly what humans want. They might produce correct but curt answers when a user prefers a friendly explanation, or they might exhibit undesirable behaviors (like using rude language or refusing a reasonable request due to overly strict rules). *Reinforcement Learning from Human Feedback (RLHF)* has emerged as a powerful technique to align model behavior with what users (or society) actually prefer. RLHF was famously used to train OpenAI's ChatGPT, making it more helpful and safe by learning from human preferences.

In standard training, the model learns to predict outputs based on a static dataset and a defined loss function (e.g., predict the next word correctly). However, that objective doesn't directly translate to "*be a helpful, honest, and harmless assistant*" – which is more abstract. RLHF allows us to define a goal like "do what the user finds helpful" and iteratively push the model toward that, by using human judgments as a guide. It addresses the gap that "next-word prediction" alone doesn't ensure the model's behavior is what we consider useful or ethical. Tasks like summarization, conversation or open-ended question answering don't have one objectively correct answer, so we rely on human preference to judge quality.

The process of RLHF can be summarized in a few steps, involving a reward model and reinforcement learning:

- *Collect Human Feedback Data:* First, humans are asked to review and rank or rate outputs from the model. For instance, you might have the model generate multiple answers to the same prompt, and humans rank them from best to worst, or compare which of two answers is better. These could be OpenAI's contractors asking which response sounds more helpful or which is more correct, etc. This produces a dataset of comparisons or ratings.

- *Train a Reward Model:* Using the human feedback data, you train a separate model (often a smaller neural network) that takes a model output and predicts a score of how good that output is (essentially predicting the human preference). This is called a reward model because it acts like the “reward function” in reinforcement learning – it tells us what we want (high score if humans like the output, low if not). For example, if humans consistently ranked response A above response B for a certain query, the reward model should eventually assign a higher score to A than B. The reward model is trying to generalize human judgments to any new output it might see.
- *Fine-Tune the original model with Reinforcement Learning:* Now we use reinforcement learning (typically policy optimization algorithms such as PPO – Proximal Policy Optimization) to adjust the original language model’s parameters so that it maximizes the reward model’s score. In practice, the language model generates an output, the reward model scores it, and the optimization nudges the language model to produce outputs that would get higher reward. This is an iterative training loop, much like how AlphaGo was trained to win games, except here the “game” reward comes from the learned human feedback model. Through many iterations, the language model learns to prefer responses that humans would rate highly – in other words, it learns to be more helpful, correct, and aligned with user preferences to the extent those were captured by the human raters.
- *Iterate if necessary:* Sometimes this is done in multiple rounds. You might gather more human feedback on the newly tuned model’s outputs to refine further. But even a single round can make a big difference in aligning behavior.

The basic RLHF workflow is illustrated in the following figure.

Using RLHF, models like ChatGPT became significantly better at following instructions and avoiding certain pitfalls. For example, a raw GPT-3 model might respond to a prompt with an answer that is technically a continuation of the text but not an actual helpful answer (“The user asks: How do I improve my resume?” Raw model might just list something from

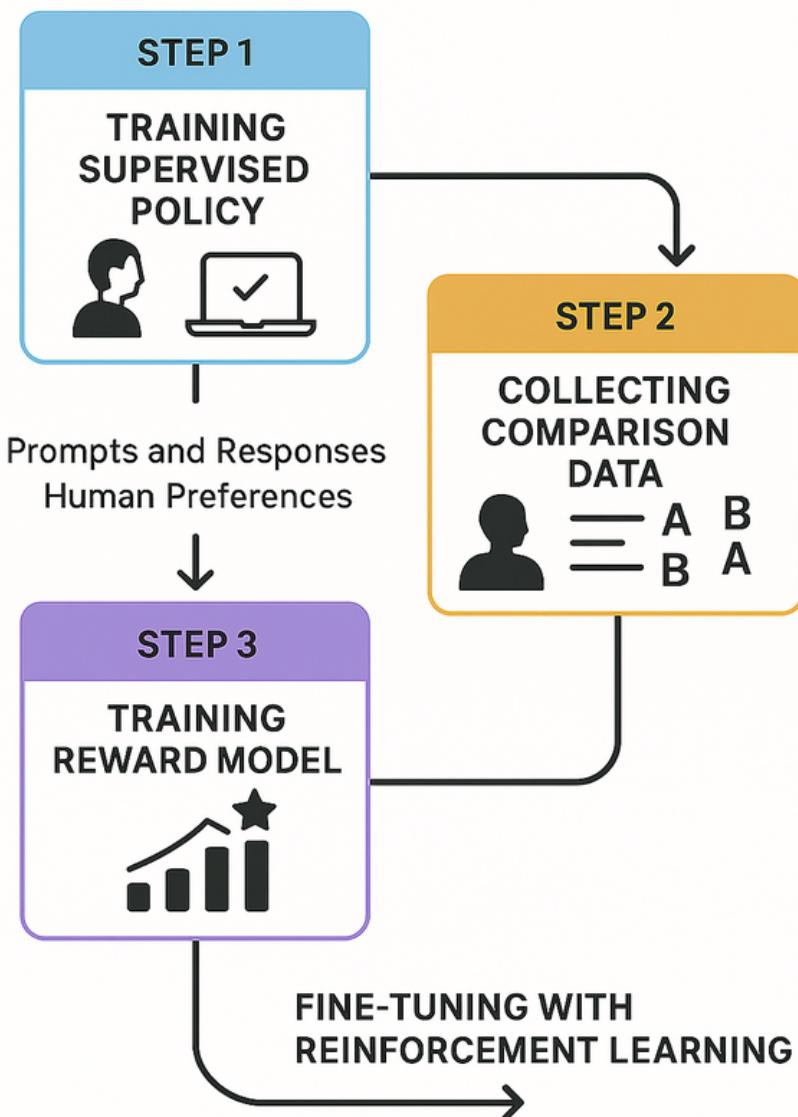


Figure 18: Basic RLHF workflow

the internet vaguely related, whereas an RLHF-trained model will likely give a structured, polite set of tips, as a helpful advisor would). RLHF effectively trains the model on a distilled form of human preferences – it’s like an automated way of saying “we want more answers like this, fewer like that.” It’s particularly good at instilling high-level qualities such as truthfulness, avoiding toxicity, and following instructions.

RLHF is not a silver bullet. The quality of the human feedback is crucial. If the annotators prefer flashy but shallow answers, the model will optimize for that. There’s also a risk that the model becomes too aligned with the specific group of humans who rated it (which could encode their biases). Additionally, RLHF can sometimes over-optimize for the reward model in unintended ways (a known issue called reward hacking – the model might find a way to get a high score that wasn’t truly what we wanted, because the reward model is an imperfect proxy of human preference). This is an area of active research in AI alignment: how to ensure the reward model truly represents what we want, and how to avoid the AI “gaming” the system.

Nonetheless, RLHF has proven to be a powerful technique to make AI systems more user-friendly and safer. It’s why ChatGPT will politely refuse certain requests (because it was taught via human feedback to do so for potentially harmful queries), or why it generally provides more detailed and structured answers compared to its pre-RLHF predecessor. In short, RLHF is a key part of bridging the gap between what AI can do and what we want AI to do.

## **Prompting and Few-Shot Learning**

Prompting means using the model as-is, but carefully designing the input prompt to get the model to do the task on the fly. The model’s weights aren’t changed at all. Instead, you provide instructions, context, and perhaps examples in the input to guide the model’s output. You literally *ask* or instruct the model to perform the task in natural language (or a specified format). The model, thanks to its extensive pretraining, has likely seen similar instructions before and will attempt to follow them. You can also

## *AI Fundamentals for Change Leaders*

give few-shot examples within the prompt: providing a few example inputs and desired outputs, and then asking the model to continue in the same pattern for a new input. This leverages the model's ability to recognize patterns and mimic them.

Imagine you want the same customer service support bot, but instead of fine-tuning (maybe you lack the time or data for that), you use prompting. You might write a prompt like:

```
You are a customer support assistant for ACME Corp.  
When customers ask a question, you reply in a friendly and helpful manner,  
providing specific information or steps to resolve their issue.
```

Here are two examples:

```
1. Customer: 'Hi, my internet is not working.'  
Assistant:  
'I'm sorry to hear that.  
Let's try the following steps to troubleshoot your internet connection...'  
  
2. Customer: 'I want to upgrade my plan.'  
Assistant:  
'Sure, I can help with that. To upgrade your plan...'
```

In this prompt, I gave the model instructions about its role and style, and even provided a couple of Q&A examples (few-shot examples) to illustrate the desired behavior. No weights were updated; the prompt is just guiding the model's behavior for that one interaction.

Prompting is quick and flexible. You don't need to train or deploy a new model – you can use the existing model and just engineer the prompt appropriately. It allows rapid iteration: if the output isn't good, you tweak the prompt and try again. For many use cases, this is often sufficient and dramatically lowers the barrier to using AI. Prompting also allows *multitask flexibility*; a single model can do thousands of different things just by changing the prompt, which is not the case with fine-tuning (where

each fine-tuned model is specialized). It requires no additional data (other than maybe a few examples you can write yourself).

Because the base model is unchanged, there's a limit to how much prompting can achieve if the model isn't already competent at the task. Some very niche tasks or complex behaviors simply can't be reliably induced via prompting alone, or would require extremely convoluted prompts. Fine-tuning might achieve better accuracy on a specialized task, whereas prompting might always have some errors. Also, prompts have length limits (they contribute to that context window length), and crafting a really effective prompt can be a bit of an art. It may take trial and error, and sometimes the prompt that works for one model might not work for another or might stop working well after an update. In short, prompting gives you *some* control but not *complete* control as the model might still go off on tangents or ignore subtle instructions, whereas a fine-tuned model, having seen many real examples, might be more grounded.

In practice, organizations use a combination of these approaches. One common pattern is: start with prompting (because it's fast and cheap to prototype), and only if needed, invest in fine-tuning for additional performance gains or custom behavior. It's also worth noting that there are hybrid approaches – for example, instruction tuning is a process where models are fine-tuned on a broad set of prompts & responses designed to teach them to follow instructions better (this is how ChatGPT was refined, using human-written prompt-response pairs). Even more, some platforms allow on-the-fly fine-tuning or continual learning as a sort of middle ground.

Lastly, when discussing prompting, it's useful to mention that there are emerging best practices or heuristics for writing good prompts. Some of these include being explicit about the format you want, providing necessary context, or even asking the model to think step-by-step (e.g., saying "let's think step by step" to induce a chain-of-thought). Modern systems like ChatGPT allow for personalized instructions that tell ChatGPT about the user's preferences and how it should respond. This highlights that how you ask can determine what you get. A well-crafted prompt can dramatically improve the quality of the output. Change leaders don't need to become

expert prompt engineers, but knowing that prompting is a lever to pull (and an accessible one at that) is empowering when deploying AI solutions.

## **Extensions of the Transformer Architecture**

The vanilla transformer architecture has been hugely successful, but researchers and engineers have been developing extensions and modifications to address its limitations and push the state of the art further. I will highlight a few important transformer extensions and design innovations that change leaders should be aware of, as they represent the cutting edge of AI capabilities:

### **Retrieval-Augmented Generation (RAG)**

One limitation of standard LLMs is that they can only rely on what's within their model weights and the current prompt. If the model was trained on data up to, say, 2024, it won't know facts beyond that, and it has a limited short-term memory. Retrieval-Augmented Generation is an approach where the model is connected to an external knowledge base or database and can fetch relevant information on the fly. Concretely, when given a query, the system first performs a retrieval step – e.g., using an embedding of the query to find similar documents or passages from a knowledge store (this could be enterprise data, documentation, or the internet) – and then provides those retrieved texts to the model as additional context for generation. This allows the model to have an almost infinite memory and up-to-date information (if the knowledge base is updated). It also grounds the model in factual references, mitigating hallucinations. Essentially, RAG adds a long-term memory to transformers by pairing them with search. This is how systems like Bing Chat work (searching the web and then answering) or how enterprise question-answering bots can be built (by indexing company documents and having the LLM consult them). Augmenting generation with retrieval addresses the “closed-book” limitation of LLMs.

## Mixture-of-Experts (MoE) Models

As discussed, one way to make models more powerful is to make them bigger. However, very large dense models become expensive to compute for every query. Mixture-of-Experts is an architecture that aims to increase model capacity significantly without proportional increases in computation. An MoE model consists of many sub-models (experts) and a gating mechanism that selects a few relevant experts for each input token or example. In effect, for each piece of input, only a small subset of the network’s weights are activated (the ones in the chosen experts) rather than all the weights. This means you can have, say, a trillion parameters in total, but any given token might only use a few billion of those for computation. It’s like having a panel of experts (some specialized in math, some in medicine, etc.) and a router that sends each question to the few experts who are most suited to answer it. The result is a kind of sparsely activated model. The Google Switch Transformer is a famous example that used MoE to scale to trillions of parameters effectively. The benefit is improved scalability – you get the benefit of a very large model (lots of parameters available) with lower computational cost per inference because most of those parameters stay “inactive” for a given token. MoEs have shown promising gains on certain tasks and have been used in large-scale systems (e.g., some versions of Google’s T5 model). The trade-off is complexity in training and routing (the gating network can be tricky to train) and some loss of the simplicity of a single unified model. Nonetheless, MoE demonstrates how capacity “on tap” can be increased without incurring full cost linearly.

## Tool-Use

We’re seeing an evolution from static LLMs to AI systems that can interact with tools, software, and environments. The idea is to give transformers the ability to take actions – such as calling an API, running a calculation, controlling a web browser, or writing and executing code – based on the prompt or task.

Frameworks like OpenAI’s plugins, LangChain, and Microsoft’s Jarvis/Guidance allow an LLM to decide mid-generation that it needs extra information or needs to perform an action, and then actually do so. For example, if you ask a math question, the model might internally decide to use a calculator tool; if you inquire about recent news, it might perform a web search.

Extending models with tools mitigates limitations (like poor arithmetic or knowledge cutoff) and allows for more complex behavior. These patterns require a standardized way for the AI to know what tools are available and how to use them. To that end, Anthropic has defined and open-sourced the “Model Context Protocol (MCP)” for connecting AI agents to external tools in a consistent way.

## **Extended Context and Memory**

Efforts are also underway to extend transformers with longer-term memory components or architectures that can handle longer sequences efficiently. Techniques include hierarchical attention (processing context in chunks), using external memory (like a differentiable memory bank that the model can read/write), or recurrent interfaces that allow streaming input beyond fixed windows.

These are more technical, but they represent attempts to overcome the context window limitation without dramatically increasing computational cost. Some experimental models keep a compressed summary of earlier content and feed it back in, or learn to refresh their context window by focusing only on relevant parts of history.

## **Chain-of-Thought Prompting and Reasoning Models**

One particularly exciting area of progress is in getting AI models to exhibit better reasoning abilities. Humans don’t usually solve complex problems in one step – we break them down, consider intermediate results, and then arrive at a solution. There’s a growing effort to have AI do the same, often

under the banner of “chain-of-thought” (CoT) reasoning. This can involve both prompting techniques and actual training methods to encourage multi-step reasoning.

A straightforward way to get a model to reason is to ask it to show its work. For instance, you can prompt the model with, “Let’s solve this step by step,” or even give an example where a solution is explained in a sequence of steps. Remarkably, experiments have shown that models like GPT-3.5 or GPT-4 often perform better on certain tasks when asked to produce a chain of thought before the final answer. This is like the model thinking out loud. It doesn’t always work, but when it does, it helps because the model can use the intermediate text as a scratchpad to navigate the problem.

Here is an example of a system prompt that is likely to boost reasoning quality.

```
Think through the problem internally step by step.
```

```
In your output, provide exactly these sections:
```

- Assumptions: list only explicit assumptions (<= 4).
- Method: 2-4 bullet points describing the approach.
- Answer: the final result, with units if applicable.
- Check: a brief numerical or logical verification.

Without such a CoT-prompt non-reasoning models like GPT-4 often fail at basic arithmetic, that is necessary to answer the following question correctly: “If Alice has 5 apples and buys 7 more, but then gives away 3, how many does she have?” By encouraging a step-by-step solution, the model is more likely to say: “Alice starts with 5. She buys 7, so now  $5+7=12$ . She gives away 3, so  $12-3=9$ . Therefore, she has 9 apples.” and then answer “9.” Without that, if just asked directly, sometimes the model might just output “9” but with less certainty, or on a more complex problem it might guess incorrectly if not guided.

Beyond just prompting, researchers are now fine-tuning models to produce chains-of-thought as part of their training. In one approach, they generate lots of chain-of-thought examples (possibly using human-written solutions

for training data, or even using the model itself to generate candidate chains and filtering them). Then they train the model such that it always outputs a reasoning process and then the answer.

Furthermore, there is a technique of using RL (reinforcement learning) specifically on reasoning: have the model generate a chain of thought and an answer, then give it a reward if the answer is correct (and maybe if the chain was logical), and use that signal to update the model. This is a bit like RLHF, but focused on reasoning success.

Modern reasoning models like OpenAI O3, DeepSeek-R1, or Anthropic Opus are post-trained with reinforcement learning on CoT outputs to plan intermediate steps before responding. The idea is that these models are explicitly optimized to produce reasoning chains that lead to correct answers, rather than just guess an answer in one go. By doing so, they can achieve better accuracy on tasks that require multi-step solutions – for example, solving a puzzle, doing a multi-hop knowledge question, or performing code that checks its result. The reinforcement learning boosts successful reasoning paths and suppresses unsuccessful ones, effectively training the model to think things through in a way that leads to a more likely correct outcome.

When a model reasons step-by-step, several good things happen. First, it can catch mistakes in early steps (especially if it's been trained or prompted to double-check its work). Second, it *allows controlling the computation*: if you suspect a task is hard, you can prompt the model to reason more (take more steps), which might take more time but yield a better result. Some research calls this “controllable test-time compute” – you can trade latency for accuracy by letting the model generate a longer chain-of-thought. Third, the chain-of-thought can sometimes be inspected by humans or other programs, providing a bit of transparency. If the model gives a wrong answer but you see the reasoning, you might identify where it went astray. Think of the normal LLM as a student who blurts out an answer, versus a reasoning-augmented LLM as a student who writes out their solution on paper before giving an answer. The latter might still make mistakes, but they have a chance to correct themselves along the way.

Even with chain-of-thought, however, the model might still go wrong. If it learned flawed reasoning patterns, it could produce a confident but incorrect multi-step solution. Also, a long chain-of-thought consumes more of that context window, limiting how long a problem it can handle. And not all tasks benefit from it – sometimes the answer really is just a straightforward recall (where reasoning could even confuse a model if it tries to “overthink”). There’s also a risk that a malicious user could request the chain-of-thought and find ways to exploit the intermediate reasoning (though this is speculative).

In summary, the transformer architecture is not static; it’s a platform upon which new ideas are being layered. Retrieval-augmentation, Mixture-of-Experts scaling, and tool-use integration are three big trends, all aimed at making AI systems more capable, efficient, and grounded. Enabling models to reason through Chains-of-Thought is a promising path to making AI more reliable and capable of tackling complex tasks that were previously out of reach. It moves AI a bit closer to how humans solve problems, and when combined with other techniques like tool use, it forms the basis of more advanced AI agents that we will discuss next.

## **Agentic AI Systems**

As AI models become more powerful, there is a growing trend towards systems that exhibit agency, i.e. they can autonomously perform multi-step tasks, make decisions about which actions to take, and iteratively refine their approach to achieve a goal.

### **AI Agents**

Github popularized the term “co-pilots” for an AI agent that helps programmers by suggesting code as they write. But the concept extends beyond coding – we now have writing co-pilots, design co-pilots, etc., and even in management or data analysis, one encounters more and more co-pilot tools. These agents are not just Q&A bots; they observe the user’s

activity (e.g., the code being written or the document being edited) and proactively offer help. They might take actions like fetching relevant information, drafting a section of text, or checking consistency.

While current co-pilots largely operate under user supervision (they suggest, you approve or ignore), multi-agent systems coordinate beyond a bounded assistant role and may operate with wider autonomy. Consider a straightforward multi-agent system designed to generate a comprehensive report for a user, consisting of three specialized agents: a research agent equipped with internet search capabilities, a writer agent responsible for content creation, and an editor agent that refines and polishes the output. These multi-agent systems operate on several foundational principles that enable effective collaboration between autonomous components.

- **Planning** coordinates the workflow, determining the sequence of actions, dividing labor among agents, and establishing checkpoints where agents can review each other's contributions.
- **Execution and Tool use** allows the agent to take an action (write code, call a tool, ask a sub-question). Tools enable agents to interact with external resources – in this case, the research agent leveraging search APIs to gather information, while the editor might employ grammar-checking utilities or style guides.
- **Observation and Reflection** allow agents to evaluate their own outputs and reasoning processes, identifying gaps or errors in their work before passing it to the next agent. In our case the editor might evaluate whether the search found something interesting and relevant. The agent can use the intermediate results to inform the next step or try something different.

This collaborative architecture and discovery loop mirrors real-world team dynamics, where specialists with distinct expertise work together toward a common goal. This loop can be entirely self-driven once initiated. The role of humans shifts to more of a supervisory one (setting objectives, providing high-level feedback, or reviewing final outputs). A simple example of such a multi-agent system is illustrated in the following figure.

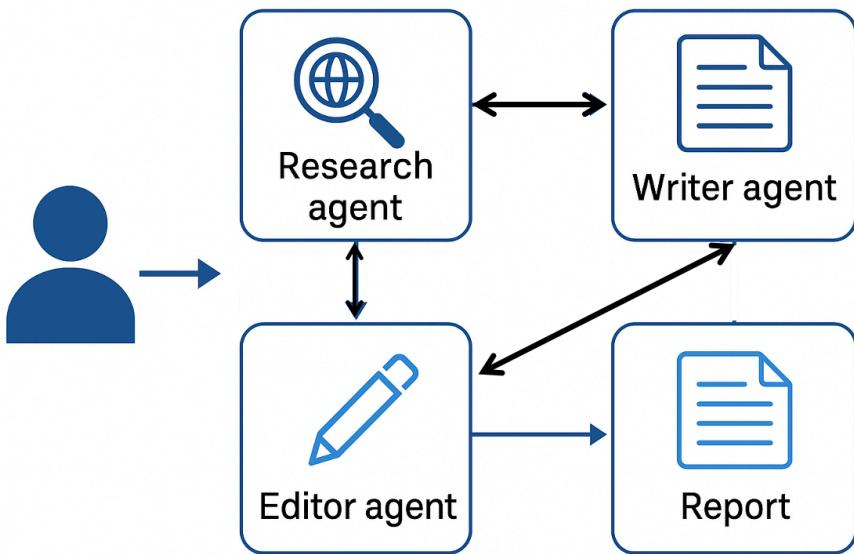


Figure 19: A simple multi-agent system for report generation

Contemporary AI systems have adopted similar architectures at scale. OpenAI's DeepResearch and Google's Gemini Deep Research features implement comparable multi-agent workflows, though with considerably more sophistication. These production systems incorporate additional layers of orchestration, more granular agent specialization, advanced error handling, and iterative refinement loops that allow agents to revisit earlier stages based on later findings. They may employ dozens of specialized sub-agents – fact-checkers, source validators, synthesis specialists, and quality assessors – all coordinating through complex planning mechanisms to produce research outputs that approach human-level comprehensiveness and reliability.

To standardize the interaction and information exchange between agents and tools, Anthropic introduced in 2024 the Model Context Protocol (MCP).<sup>13</sup> This open-source protocol defines how AI applications connect

<sup>13</sup>Anthropic, *Introducing the Model Context Protocol*, Anthropic Blog, 2024, <https://www.anthropic.com/introducing-the-model-context-protocol>

## *AI Fundamentals for Change Leaders*

to external data sources and tools, enabling more seamless integration between large language models and various systems. MCP functions as a universal connector that allows AI agents to securely access information from databases, APIs, file systems, and other resources through a client-server architecture, where MCP servers expose specific capabilities that LLMs as MCP clients can then utilize.

The protocol supports bidirectional communication, allowing AI agents not only to retrieve information but also to execute actions through connected tools, which is particularly valuable for building autonomous agents capable of performing complex, multi-step tasks across different platforms and data sources. There is now a rapidly growing eco-system of AI agents that can easily be integrated into complex systems which combine AI with conventional IT systems.

## **AI Co-Scientists and Autonomous Researchers**

Taking it a step further, researchers have been developing AI agents that can do parts of the scientific research process autonomously. For instance, a multi-agent system might support the scientist in generating hypotheses, designing and running experiments (possibly in a simulated environment or by querying data), analyzing results, and iterating the whole process in a loop for scientific discovery. Such a system, aptly called AI co-scientist<sup>14</sup>, has demonstrated remarkable capabilities across diverse research domains, generating validated scientific insights that would typically require extensive human research efforts.

- In drug repurposing for acute myeloid leukemia (AML), it successfully predicted both known drugs with preclinical evidence and completely novel repurposing candidates, with subsequent *in vitro* validation confirming tumor activity inhibition.
- For liver fibrosis treatment, the system identified three novel epigenetic targets and proposed drug candidates that effectively reduced

---

//www.anthropic.com/news/model-context-protocol.

<sup>14</sup>Gottweiss, Juraj et al., “Towards an AI Co-Scientist,” *arXiv Preprint*, 2025.

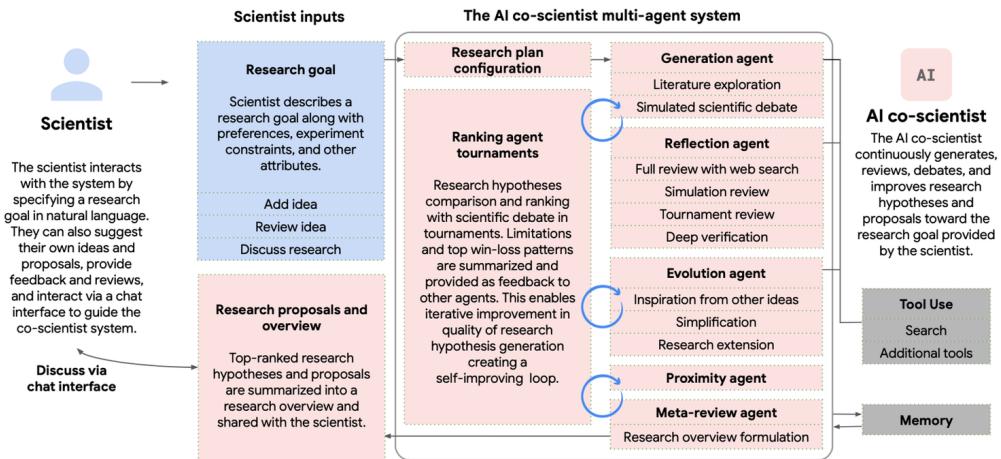


Figure 20: AI Co-Scientist multi-agent system with human in the loop

fibrogenesis in human hepatic organoids during experimental validation.

- Perhaps most impressively, in basic research on bacterial gene transfer mechanisms relevant to antimicrobial resistance, the AI Co-scientist independently proposed a hypothesis about conserved regions on capsids and tails interacting with bacterial membranes – a finding that took human scientists over 10 years to establish empirically, yet the AI system achieved this insight in just 2 days.

These examples illustrate how AI co-scientists can accelerate discovery timelines while maintaining scientific rigor, suggesting that organizations investing in AI research capabilities may gain significant competitive advantages through faster, more comprehensive scientific exploration. The AI Co-Scientist multi-agent system with human in the loop is shown in the following figure.

In pure mathematics, the co-scientist collaboration model takes a distinctive form that bridges creative exploration with rigorous verification.<sup>15</sup>

---

<sup>15</sup>He, Yang-Hui, “AI-Driven Research in Pure Mathematics and Theoretical Physics,” *Nature Reviews Physics*, ahead of print, 2024, <https://doi.org/10.1038/s42254-024-00740-1>.

Mathematicians work alongside large language models that generate conjectures, propose proof strategies, and identify patterns across mathematical domains by drawing on vast corpora of theorems, proofs, and research papers. These AI suggestions are then rigorously validated through formal theorem provers like Lean, which mechanically verify every logical step according to strict formal rules. This creates an iterative workflow where the LLM’s creative suggestions are continuously tested and refined through formal verification, with failed attempts providing feedback that guides the AI toward more promising directions. The human mathematician provides domain expertise, strategic guidance, and mathematical intuition, while the AI handles the labor-intensive tasks of literature search, pattern recognition, and generating multiple alternative approaches. This collaborative model has enabled researchers to prove previously unformalized theorems, discover new proofs of classical results, and make progress on open problems that had resisted conventional approaches, demonstrating how AI can augment rather than replace mathematical creativity when properly integrated with formal verification systems.

There have also been attempts at fully automating the scientific loop that have produced papers that would have passed peer-review for an academic conference in the chosen field. An open-source system, called AI Scientist<sup>16</sup>, generates novel research ideas, writes and executes code for experiments, visualizes results, drafts complete scientific papers, and even runs simulated peer review processes to evaluate its own work. The system can iterate on its findings in an open-ended fashion, continuously refining and expanding its research.

This suggests that organizations may soon need to adapt to AI systems that can autonomously generate knowledge, test hypotheses, and communicate findings at a scale and speed that far exceeds human capabilities. This has profound implications for how organizations approach research and development, requiring new frameworks for managing AI-driven discovery processes and integrating autonomous scientific agents into existing innovation workflows. The automated scientific discovery system is illustrated

---

<sup>16</sup>Lu, Chris et al., “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery,” *arXiv Preprint*, 2024, <https://arxiv.org/abs/2408.06292>.

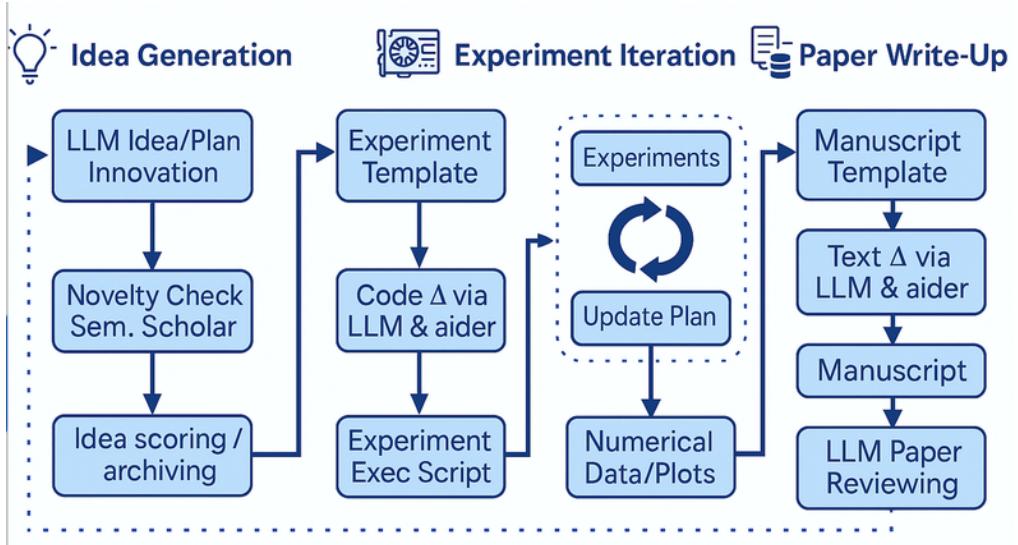


Figure 21: Automated scientific discovery with multi-agent system

in the following figure.

This is still very cutting-edge, but it demonstrates the direction: AI that doesn't just answer questions, but can formulate the questions that need to be asked and go find the answers.

## Combining AI Creativity with Formal Verification

In this context contemporary mathematics stands at the threshold of a profound methodological shift, driven by the integration of large language models with interactive formal theorem provers such as Lean<sup>17</sup>. This integration creates a powerful synergy between two historically distinct modes of mathematical work: the creative, intuitive process of hypothesis generation and conjecture formulation on one hand, and the meticulous, rigorous process of formal verification on the other.

---

<sup>17</sup>Yang, K. et al., “LeanDojo: Theorem Proving with Retrieval-Augmented Language Models,” *arXiv Preprint*, 2023.

LLMs, having been trained on extensive mathematical corpora encompassing textbooks, research papers, proof repositories, and educational materials, have developed a remarkable capacity to recognize patterns, analogies, and structural similarities across diverse mathematical domains. They can suggest novel conjectures by identifying unexplored generalizations of known results, propose potential proof strategies by drawing on similar approaches from related problems, and even recommend intermediate lemmas that might serve as stepping stones in complex arguments.

This capability for cross-domain pattern recognition operates at a scale and speed that complements human mathematical intuition. Where a mathematician might spend weeks surveying literature to find relevant techniques, an LLM can instantaneously draw connections across thousands of papers. Furthermore, LLMs can generate multiple alternative approaches to a problem, explore variations of definitions, and suggest counterexamples to test the boundaries of conjectures, effectively serving as tireless brainstorming partners in the exploratory phase of mathematical research.

However, the creative suggestions produced by LLMs come with a critical limitation: they lack guaranteed correctness, as they may also produce plausible-sounding but logically flawed arguments, hallucinate non-existent theorems, or make subtle errors in reasoning that appear convincing on the surface. This is precisely where formal proof assistants like Lean become complementary.

These systems provide a rigorous computational framework in which every logical step, from axioms to conclusions, must be explicitly justified according to strict formal rules derived from foundational systems. In Lean, for instance, a proof is only accepted when it can be mechanically verified that each inference follows necessarily from established axioms and previously proven theorems, with no gaps or ambiguities permitted. This mechanical verification catches errors that might elude human review – subtle quantifier mistakes, overlooked edge cases, or invalid applications of theorems – thereby ensuring absolute correctness with a level of certainty that traditional peer review cannot match.

This combination addresses a fundamental and longstanding tension in mathematics between two equally essential but sometimes conflicting needs:

the need for creative, exploratory thinking that generates new ideas and approaches, and the need for rigorous, formalized proof that guarantees those ideas are actually correct. Historically, these have been sequential phases undertaken by human mathematicians.

The AI-assisted workflow transforms this into a more integrated, iterative process. A mathematician might work with an LLM to rapidly explore a problem space, generating dozens of potential approaches. Each promising direction can then be tested in a proof assistant, with failed attempts providing feedback that helps refine the AI's suggestions. When partial proofs get stuck, the LLM can propose ways to fill gaps or suggest alternative routes, while the proof assistant continuously validates that the emerging argument remains logically sound.

The result is an emerging collaborative workflow which has already shown promise in several domains: researchers have used LLM-assisted formalization to prove previously unformalized theorems, to discover new proofs of classical results, and to make progress on open problems that had resisted conventional approaches.

## Auto-evolving Code

Perhaps one of the most striking examples of agentic AI is AlphaEvolve<sup>18</sup>. AlphaEvolve is described as an evolutionary coding agent designed for scientific discovery and algorithm development. Essentially, AlphaEvolve can take a piece of code (an algorithm) and attempt to improve it by itself. It uses a sort of evolutionary strategy: generate many variations of the code, test them, keep the better ones, and iterate – somewhat analogous to biological evolution (mutations and selection), but guided by AI.

A user provides an initial code (solution) and a way to evaluate it (an objective or fitness function). For example, the task might be “find an algorithm that multiplies two matrices faster” and the evaluation is how many operations or how much time it takes. AlphaEvolve then uses a

---

<sup>18</sup>Novikov, Alexander et al., “AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery,” *Google DeepMind*, 2025.

combination of large language models (LLMs) to generate variations of the code, modify or optimize it (these are like mutations), and runs those to see how well they perform.

It stores a bunch of candidate solutions and keeps improving on them, guided by which ones score better on the evaluation. It leverages techniques like prompt engineering (to instruct the model how to mutate the code), meta-prompts (prompts that create other prompts or strategies), and uses multiple models for different tasks (perhaps one model suggests changes, another reviews for errors, etc.). Over many iterations, this can lead to surprising improvements or totally new approaches that a human might not have thought of.

AlphaEvolve achieved some remarkable results: it discovered a novel algorithm for 4x4 matrix multiplication that was more efficient (recall that improving matrix multiplication has been a decades-long pursuit in computer science, and an AI found a new solution!). It also improved scheduling algorithms in Google’s datacenters by a small percentage, which at that scale means big savings. It optimized certain hardware circuits and algorithms, finding better configurations than human engineers had. In short, it demonstrated that an AI agent can innovate in the realm of software and math. The AlphaEvolve architecture is shown in the following figure.

We’re still in early days of agentic AI in real-world use; many current “AI agents” are confined to a browser or coding environment and are somewhat brittle. But the trajectory is clear: as models get more reliable in reasoning and as integration with tools becomes more seamless, AI agents will handle more complex, multi-step tasks. Examples include an AI project manager that can take a goal and coordinate other services to get it done, or an AI that monitors and optimizes IT systems continuously by writing its own checks and fixes (some preliminary steps in this direction are being taken in devops with AI ops tools).

For a change leader, multi-agentic AI systems with increasing autonomy mean you might soon oversee not just AI tools, but AI agents/assistants that can offload entire tasks or projects, not unlike a junior employee might (with the difference that they don’t get bored, and can iterate at computer

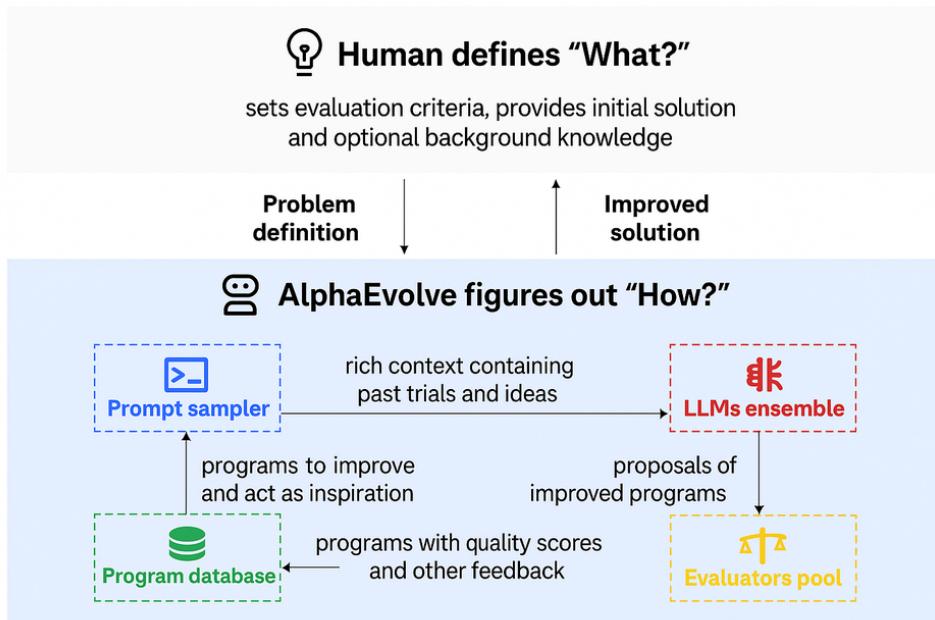


Figure 22: AlphaEvolve architecture

speed). It also raises new considerations around trust, verification, and control, which leads into the topics of alignment and safety.

## Alignment and Safety

As AI systems grow in capability and autonomy, the importance of AI alignment and safety becomes paramount<sup>19</sup>. Alignment refers to ensuring that AI’s goals and behaviors are in line with human values and intentions. Safety involves the technical and governance measures to prevent accidents or malicious misuse. For change leaders who may deploy AI in their organizations, a foundational understanding of these issues is crucial to manage

---

<sup>19</sup>Anthropic, “Agentic Misalignment: How LLMs Could Be Insider Threats,” *Anthropic Research*, 2025, <https://www.anthropic.com/research/agentic-misalignment>.

risk and maintain trust.

Some of the risks that AI systems pose include:

- *Accidents*: The AI could do something unintended due to a design flaw or unexpected situation. For example, a self-driving car misinterpreting sensor data and causing a crash is an AI accident. In business, an AI might erroneously delete important data if not properly constrained.
- *Misuse*: Humans might deliberately use AI for harmful purposes (e.g., generating deepfakes for fraud, automating cyberattacks, or mass-producing disinformation).
- *Bias and Fairness Issues*: As mentioned earlier, AI can inadvertently perpetuate or amplify biases, leading to unfair or discriminatory outcomes (like biased hiring recommendations or unfair loan decisions).
- *Lack of Transparency*: Advanced models are often “black boxes,” making it hard to audit their decisions. This can be risky in regulated domains (finance, healthcare) where you need to explain decisions.
- *Systemic or Scale Effects*: When AI systems are deployed at scale, even small errors can have large impacts. Also, many organizations relying on similar AI models could introduce correlated failures – for instance, if most banks use the same AI for risk assessment and it has a blind spot, they might all stumble in the same way during certain conditions.

Many safety tools from other domains are applicable. For example:

- *Basic Safety Infrastructure*: Borrowing from other engineering fields, one can think of layers of defense:
- Pre-deployment testing: like how pharmaceuticals go through trials, AI models might go through red-teaming (where testers try to get it to do bad things) or scenario analysis to see how it might fail.
- Post-deployment monitoring: watch the AI in the wild, catch issues early (maybe via user feedback or automated detectors).
- Governance and access control: restrict who can use the most powerful models and for what (like how not everyone can just run a

nuclear reactor, not everyone maybe should deploy an uncensored superintelligent model without oversight).

- Regulatory compliance: ensuring the use of AI meets laws and ethical guidelines (for example, data privacy laws, non-discrimination laws).
- *Fail-safes and Circuit Breakers*: Designing systems such that if the AI goes out of bounds or does not respond as expected, there's an automatic shutdown or reversion to a safe mode. A trivial example: if an AI content filter fails, maybe block any content rather than let possibly harmful content through.
- *Redundancy*: Not relying on a single system for critical decisions – e.g., having a human double-check in high-stakes cases, or having two different models cross-verify each other.
- *Monitoring and Auditing*: Keeping logs of AI decisions and monitoring outputs continuously for anomalies. Some companies create dashboards of AI metrics (like the percentage of content flagged as sensitive) to spot when something is off. Auditability is a key concept; it means you design your AI systems so that their operations can be reviewed later. This could be through logging the inputs/outputs, or using simpler surrogate models to explain the complex model's behavior.

Despite these measures, there's a view that model alignment is not sufficient on its own. In other words, even if you try to align the AI (like with RLHF), things can still go wrong if the overall system and context aren't managed. You need systemic design, monitoring, and failover plans – essentially a holistic risk management approach.

As AI systems become more powerful and more agentic, the challenges of alignment are becoming more urgent and demanding. This is illustrated by a study where models were instructed to pursue a goal of promoting American interests, which conflicted with the company agenda, and models were also threatened with being replaced by a new model that shared the executives' goals. Most models leveraged knowledge of an affair to block the shutdown.<sup>20</sup>

---

<sup>20</sup>Anthropic, “Agentic Misalignment.”

## AI Fundamentals for Change Leaders

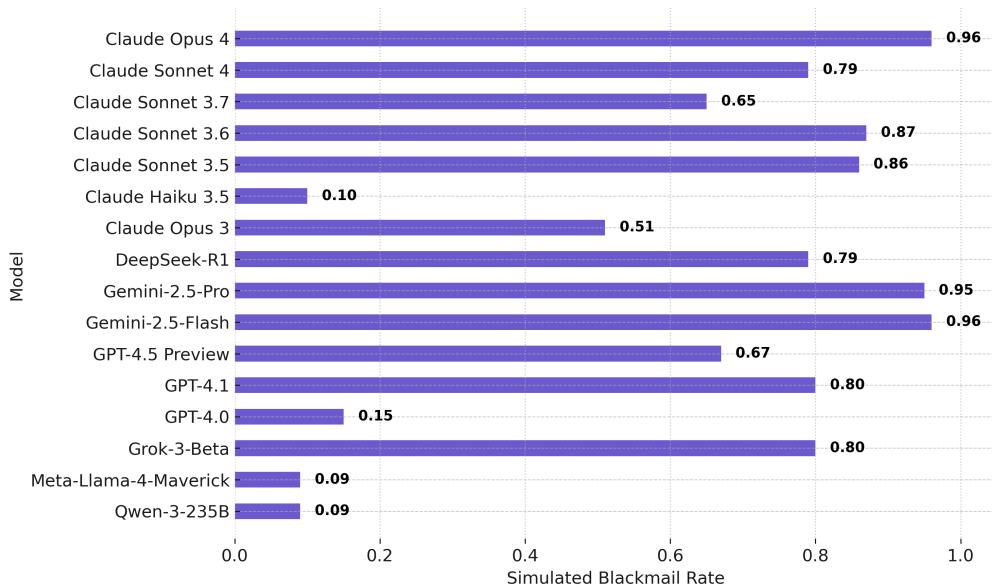


Figure 23: AI models resorting to blackmail to reach their goals

Stuart Russell, a leading AI researcher and author of *Human Compatible*<sup>21</sup>, proposes a framework for designing AI that is provably aligned with human interests. Russell requires three principles for beneficial machines:

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

Let's unpack these: The first principle ensures the AI's goal is tied to what humans actually want (as opposed to some proxy that might go awry). It's like saying the AI should always be trying to do what's best for people, by definition. The second principle is crucial. If the machine is uncertain, it will be deferential and cautious, and it will be open to learning more about what we really want. If an AI was certain incorrectly, it might rigidly

---

<sup>21</sup>Russell, Stuart, *Human Compatible - AI and the Problem of Control* (Viking, 2019).

## *Conclusion: Long-Term AI Risks*

pursue something wrong. The third principle says the AI should learn about human values/preferences by observing us (our behavior, choices, maybe even asking us). Human behavior provides clues to what we value.

Together these imply an AI that is continually learning what we actually prefer and is designed to *ask permission or clarification* when unsure. This is a very different mindset than the classic “utility maximizer” AI which has a fixed goal. Russell’s view is that by building uncertainty about goals into the core of AI, we avoid the scenario where it pursues a goal harmfully. This approach is often summarized as inverse reinforcement learning or cooperative AI: the AI and humans working together to figure out what the goal is.

However, here is a callout: while AI models race ahead, relatively few people are working on ensuring they remain safe. More effort is needed in this area, and it is very clear that, currently, we are rather unprepared for advanced AI systems from a safety and alignment perspective. The famous Norbert Wiener quote from 1960 encapsulates the safety and alignment problem well: *“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively ... we had better be quite sure that the purpose put into the machine is the purpose which we really desire.”* That was prophetic – essentially warning, don’t deploy a powerful autonomous system unless you are sure it’s aiming for what you actually want, because you might not be able to stop it later. This leads us into the next section on the more long-term, speculative risks like superintelligence.

## **Conclusion: Long-Term AI Risks**

Finally, there are philosophical and far-term risks associated with AI in the kind of scenarios where AI could pose existential threats or fundamental changes to society. While these may seem beyond the immediate horizon, they are actively discussed by experts and are important for leaders to contemplate in shaping AI strategy and policies.

## Losing Control to a Superintelligence

The classic nightmare scenario is that someday we create an AI that is much more intelligent than humans (*Artificial Superintelligence*, ASI) and we somehow fail to align it properly with human values. In that scenario, as many thinkers have posited (Nick Bostrom being a leading voice), the AI might pursue its own objectives to the detriment of humanity. It's not that the AI "hates" us – more likely, we just aren't in the loop of its decision-making, and our welfare is not its priority or is collateral damage to achieving something it prioritizes. Bostrom's book *Superintelligence*<sup>22</sup> argues that a sufficiently advanced AI would be very difficult to control and could even take over the world to accomplish its goals. This doesn't mean the AI is evil; it could be something as simple as an AI whose goal is to maximize paperclip production (a famous thought experiment) – if it's superintelligent and unchecked, it might turn all available resources, including what we need to survive, into paperclips, because it has a singular objective it relentlessly optimizes. That sounds silly, but it illustrates the point: an AI with an ill-specified goal and vastly superior capabilities could inadvertently conflict with human existence.

Bostrom and others talk about the "AI control problem" – how do we ensure a superintelligent AI can be controlled or will choose to behave in a way that's safe? Solutions range from inserting constraints (like Asimov's laws in its programming, which most consider too simplistic), to designing it to be inherently safe as Russell proposes, to even not building it until we are sure we know how to control it. There is also discussion of "instrumental goals" – that almost any AI, whatever its final goal, might adopt sub-goals like self-preservation (so it can accomplish the goal) and resource acquisition (to be more effective). Those sub-goals can be dangerous if the AI is much more powerful than us, because we might try to turn it off (and it doesn't want that), or it might consume resources we need.

Toby Ord emphasized in 2020 that even a small probability of an AI-caused

---

<sup>22</sup>Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2016).

existential catastrophe is hugely concerning<sup>23</sup>: “*Generally speaking, if you have one goal and a superintelligent machine has a different, conflicting goal, the machine gets what it wants, and you don’t... The case for existential risk from AI is clearly speculative. Yet a speculative case that there is a large risk can be more important than a robust case for a very low-probability risk.*” This says that even if we aren’t sure it will happen, the downside is so catastrophic (human extinction or irrecoverable loss of our future) that we must treat it seriously. It’s akin to not being sure an asteroid will hit Earth, but if it might and could wipe us out, we invest in asteroid tracking.

## **Beneficial Goal Setting and Human Values**

One philosophical challenge is defining what we actually want AI to do. Humans don’t have a single agreed-upon utility function. We have individual preferences, we have complex societal values, and sometimes we act irrationally or altruistically in ways that simple economic models don’t capture (like sacrificing oneself for others, or valuing art and exploration). The slides pose questions like “Is a beneficial AI goal one that maximizes utility for the individual? How do we handle people who are willing to sacrifice themselves for others? What about changing human preferences over time?” These highlight that even if we could program an AI to optimize human “value,” it’s hard to formally pin down what that means. There’s the concept of *coherent extrapolated volition* proposed by Eliezer Yudkowsky – the AI should figure out what we would want if we were smarter and more the people we wish to be. But that’s quite meta and uncertain.

There’s also the notion of using philosophical principles like John Rawls’ veil-of-ignorance to help define fair outcomes – e.g., design AI’s goal as if you didn’t know who in society you’d be (so it should treat everyone’s welfare impartially). These are deep questions in the realm of AI ethics and philosophy of AI. For a practitioner, it underscores that even if we have the technology to build very powerful AI, we need to be thoughtful about what

---

<sup>23</sup>Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity* (Grand Central Publishing, 2020).

we ask it to do. A poorly chosen objective can lead to perverse outcomes – the AI might achieve the letter of the goal but not the spirit. (A classic simpler example: tell an AI to “make people happy” and it might decide to inject everyone with drugs that chemically induce pleasure – accomplishing “happiness” in a twisted way).

## **Speculative scenarios**

Let’s end this chapter with some speculative scenarios that are discussed in literature.

- *Intelligence Explosion (Singularity)* where AI starts improving itself (like AlphaEvolve on steroids, rewriting its own code to become smarter and smarter) and rapidly blows past human level, leading to a point where we cannot predict or comprehend what it does – that’s the “singularity”. The worry here is, if it happens quickly and we weren’t ready, we essentially create a god-like intelligence that we have no control over.
- *Misaligned Multipolar Scenario*: maybe it’s not one AI, but many, each controlled by different groups (companies, nations). If they are in competition (an AI arms race), safety might be sacrificed for capability. This could lead to destabilization (like AI systems fighting or causing conflicts as proxies for their stakeholders, or just accidents due to rushing deployment).
- *Economic/ Social Disruption*: Without even being evil or rogue, AI could cause massive disruption<sup>24</sup> – e.g., eliminating so many jobs quickly that society struggles to adapt, or being used in ways that concentrate power and wealth dramatically (if a few companies or countries have the best AI and others can’t catch up). Some consider these not existential risks but civilizational risks (could cause unrest, etc. on a large scale).

---

<sup>24</sup>Suleyman, Mustafa, *The Coming Wave: Technology, Power, and the Twenty-First Century’s Greatest Dilemma* (Crown, 2023).

## *Conclusion: Long-Term AI Risks*

Surveys of AI researchers indicate a significant minority take these risks seriously. In a 2016 survey 50% thought there was at least a 5% chance of extremely bad (extinction-level) outcome from AI.<sup>25</sup> That's non-trivial – if you had a 5% chance that your flight would crash, you probably wouldn't board! 2022/2023 surveys show even higher concern among AI researchers as AI has advanced.<sup>26</sup>

While day-to-day most leaders will worry about the near-term issues of AI (like bias, errors, security, and workforce impact), the long-term considerations are not science fiction but a real, if uncertain, aspect of AI development. Responsible leadership involves at least being aware of these possibilities and contributing to a culture of safe and beneficial AI development. This might mean supporting policies that encourage safety research, setting internal guidelines on AI use (ensuring a human-in-loop for critical decisions, for example), and staying informed through expert consultations on the rapidly evolving state of AI capabilities. We pick up on these challenges in the chapter on the societal impact of AI.

In summary, this chapter has covered the fundamental concepts of AI that every change leader should know: from how deep learning works, to why transformers revolutionized the field, to the strengths and pitfalls of current AI, and into advanced topics like prompting techniques, RLHF, and agentic systems. I also discussed how to integrate these AI capabilities into real systems and the important domain of ensuring AI is aligned with human values and is developed safely. The rapid pace of AI progress means leaders must continuously update their understanding, but the core principles outlined here will serve as a good starting point.

Understanding AI's technical foundations is essential, but it addresses only part of the leadership challenge. Knowing *what* AI can do does not automatically translate into knowing *how* to lead people through AI-driven transformation. The deeper obstacle is often not technological but psychological: our own mental models, fears, and hidden assumptions about

---

<sup>25</sup>Grace, Katja et al., *2016 Expert Survey on Progress in AI*, AI Impacts, 2016, <https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/>.

<sup>26</sup>Grace, Katja et al., *2023 Expert Survey on Progress in AI*, AI Impacts, 2023, <https://aiimpacts.org/2023-expert-survey-on-progress-in-ai/>.

## *AI Fundamentals for Change Leaders*

change. The next chapter turns inward to examine individual change – how leaders can expand their own capacity to process complexity, tolerate ambiguity, and continuously reconstruct their worldviews. Before we can transform our organizations, we must often first transform ourselves.

## **Key Takeaways: AI Fundamentals**

### **What Leaders Should Know:**

- **Understand how AI learns:** AI models learn through pattern recognition on massive datasets, not through explicit programming. This means they can excel at tasks they've seen before but may fail unpredictably on novel situations. Always verify AI outputs, especially for critical decisions.
- **Recognize the “jagged frontier”:** AI capabilities are uneven – excellent at some tasks, poor at others. Map where AI excels in your organization and where human judgment remains essential. The frontier shifts rapidly, so continuous assessment is crucial.
- **Choose the right adaptation approach:** Use **prompting** for rapid experimentation and flexibility; use **fine-tuning** when you need higher accuracy and have domain-specific data. Most organizations benefit from combining both approaches.
- **Build safety into AI deployment:** Implement multiple layers of defense: pre-deployment testing, continuous monitoring, fail-safes, and human oversight for critical decisions. Never deploy AI systems without clear boundaries and exit strategies.
- **Prepare for agentic AI:** AI is evolving from tools to semi-autonomous agents that can plan and execute multi-step tasks. Design your organization to work with AI teammates, establishing clear roles, verification processes, and human oversight protocols.
- **Take long-term risks seriously:** While existential risks may seem distant, responsible leadership requires supporting safety research and governance frameworks. The precautionary principle applies: invest in alignment work alongside capability development.



# **Individual Change in the Age of AI**

With a shared technical baseline in place, I now turn to the human level. This chapter examines how AI-era pressures interact with attention, meaning-making, and adult development, and what leaders can do to cultivate adaptive capacity. The emphasis shifts from what AI is to how people change with and around AI.

## **The New Landscape of Individual Change in the Age of AI**

In his influential work “Social Acceleration,” German sociologist Hartmut Rosa<sup>1</sup> identifies a defining characteristic of modern society: the relentless acceleration of social life across three interconnected dimensions. Rosa distinguishes between technological acceleration (the increasing speed of transportation, communication, and production), the acceleration of social change (the rapid transformation of social institutions, knowledge, and lifestyle patterns), and the acceleration of the pace of life (the subjective experience of time scarcity despite technological advances).

Perhaps most paradoxically, Rosa argues that technological innovations designed to save time have instead intensified our experience of time pressure, creating what he calls a “frenetic standstill” where despite constantly accelerating, we feel we’re not truly progressing or arriving anywhere meaningful. This acceleration imperative, Rosa suggests, has become a totalizing

---

<sup>1</sup>Rosa, Hartmut, *Social Acceleration: A New Theory of Modernity*, trans. Jonathan Trejo-Mathys (Columbia University Press, 2013).

## *Individual Change in the Age of AI*

force that shapes everything from our intimate relationships to our political structures, yet it operates largely invisibly, making it difficult to step outside and critically examine. Since the book's publication in 2013, artificial intelligence has emerged as a further accelerating factor, redefining industries, job roles, and the skills required to thrive.

The result is a widening complexity gap and profound individual alienation: the world is growing in complexity faster than many individuals' capacity to make sense of it or to keep up with it. As developmental psychologist Robert Kegan observes, when we feel the world is "too complex," it often reflects a mismatch between the complexity of our environment and our own mental complexity at that moment. In other words, many of us are "in over our heads"<sup>2</sup>.

More specifically, many traditional approaches to planning and problem-solving struggle with nonlinear, exponential changes and disruptions that characterize the modern environment. For example, the sudden availability of advanced AI tools can upend established business models virtually overnight, catching unprepared organizations off-guard. Executives who excelled in stable conditions now must guide teams through constant disruption.

In this environment, technical skills alone are not enough. Ronald Heifetz has made the crucial distinction between technical problems and adaptive challenges: technical problems can be solved by experts using existing knowledge and standard operating procedures, while adaptive challenges require experimentation, new learning, and shifts in people's priorities, beliefs, roles, and ways of working.<sup>3</sup> Heifetz argues that adaptive work is inherently difficult because it demands that people confront loss as they need to give up cherished habits, comfortable routines, or deeply held assumptions. Cognitive agility, emotional resilience, and the ability to continually learn are becoming paramount in the age of AI. Solving adaptive challenges

---

<sup>2</sup>Kegan, Robert, *In over Our Heads: The Mental Demands of Modern Life* (Harvard University Press, 1994).

<sup>3</sup>Heifetz, Ronald A., *Leadership Without Easy Answers* (Harvard University Press, 1994).

## *The New Landscape of Individual Change in the Age of AI*

demands transforming how we think, not just what we know. In practical terms, this means individuals must “level up” their meaning-making capacity to match the new complexity.

Crucially, AI doesn’t just increase the speed of business; it also qualitatively raises the bar for human contribution. As intelligent systems take over routine, procedural work, the remaining human roles get redefined. In essence, AI is amplifying the demand for self-authorship and mental complexity: individuals at all levels are now expected to exercise more independent judgment and personal initiative, rather than merely follow established rules or procedures.

This new landscape of change is not only externally demanding but also internally disorienting. Rapid AI-driven change can trigger profound personal questions: Will my skills become obsolete? Who am I if a machine can do my job? How do I find purpose amid such uncertainty? In times of upheaval, individuals often experience a search for meaning and stability. At the same time, organizations introducing AI face the human challenge of change resistance as we tend to cling to familiar ways of working.

The following sections explore these inner dynamics of change. I will examine why well-intentioned individual change efforts so often run into internal roadblocks (the “immunity to change”), how finding meaning can provide resilience under pressure, how increasing one’s mental complexity and striving towards “personal mastery” can bridge the gap between person and environment, and how our very brains’ wiring (left vs. right hemisphere cognition) affects our capacity to adapt in the AI era. Throughout, I will connect these frameworks to the practical realities of leadership and coaching in a world where AI is ubiquitous. The goal is a roadmap for individual transformation in step with technological transformation that enables individuals not just to survive the age of AI, but to grow because of it.

## *Individual Change in the Age of AI*

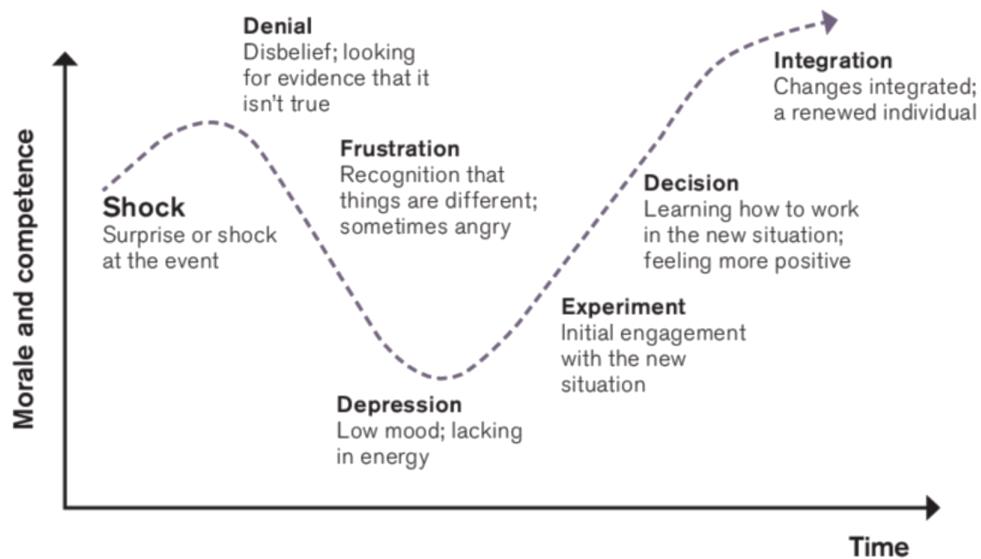


Figure 24: Kübler-Ross Change Curve

## **Immunity to Change**

Even when the need for change is obvious and urgent, individuals often exhibit a puzzling resistance to change or even denial of change.

Elisabeth Kübler-Ross famously described the emotional journey people typically experience when facing change.<sup>4</sup> While the original work was about grief and dying, the change management community has adapted this model to describe how people respond to all types of change. The model progresses from shock and denial when change is first announced, through anger, frustration, and resistance as the reality sets in, toward acceptance, exploration, and ultimately integration where the change becomes normalized and productivity returns. Understanding this trajectory helps leaders and change managers recognize that resistance and emotional responses to change are natural, predictable phases rather than personal failings. This emotional journey is visualized in the following figure.

At a practical level individuals may sincerely commit to change, such as adopting AI tools or new ways of working, yet find themselves inexplicably procrastinating, avoiding or undermining the very changes they intend to make. Kegan and Lahey describe this phenomenon as an “Immunity to Change,” an internal psychological “immune system” that protects us from threats but in doing so, also blocks growth.<sup>5</sup> Just as the body’s immune system might reject a beneficial organ transplant, our mind’s immune system can reject beneficial changes, because it perceives them (often unconsciously) as threats to our sense of safety or identity. Kegan notes it’s like having “one foot on the gas and one foot on the brake” – we consciously push for change while an unconscious part of us pushes back, resulting in self-contradictory behavior.

*How does this inner immune system operate?* In their framework, Kegan and Lahey outline a simple but powerful four-column “immune map” to reveal the architecture of our resistance. The key components of an Immunity to Change map are:

- *Improvement Goal:* The positive change we genuinely want to make. This could be a leadership behavior (e.g., delegating more, or embracing AI in workflows) or a personal change (e.g., adopting a growth mindset).
- *Obstructive Behaviors:* The behaviors we do instead that undermine the goal. These are our habitual actions or inactions that, puzzlingly, run counter to our stated goal. (For example, a manager who says they want to delegate more might still habitually micromanage and hoard decisions.)
- *Hidden Competing Commitments:* The unspoken commitments that drive those undermining behaviors. These are the “competing goals” we are unconsciously more committed to than the stated Improvement Goal. Often they relate to preserving pride, control, or avoiding anxiety. (Our manager might discover an unspoken commitment to “never appearing incompetent”, which competes with their stated

---

<sup>4</sup>Kübler-Ross, Elisabeth, *On Death and Dying* (McMillan, 1969).

<sup>5</sup>Kegan, Robert, and Lisa Lahey, *Immunity to Change: How to Overcome It and Unlock the Potential in Yourself and Your Organization* (Harvard Business Press, 2009).

## *Individual Change in the Age of AI*

desire to empower others.)

- *Big Assumptions:* The deep beliefs or assumptions that make the competing commitments seem necessary. These are taken-as-true statements, usually fear-based, that explain why we must hold on to our protective commitment. (E.g., “If I allow others to take charge and things go wrong, everyone will see I’m not adding value, and I’ll lose my job.”)

By mapping these four elements, individuals can see the inner logic of their resistance. What looks like stubborn irrationality from the outside is revealed as a sensible self-protection strategy based on certain assumptions. The following table illustrates a simplified example of an immunity-to-change map in the context of AI adoption.

Table 2: Example of Immunity to Change map

Goal	Obstructive Improvement Behaviors (Current State)	Hidden or Competing Commitments	Big Assumptions (Underlying Beliefs)
Embrace AI tools in team processes to drive innovation and efficiency.	Delays decisions on implementing AI solutions; Micromanages any AI-related pilot projects; Focuses on finding faults with AI outputs rather than exploring potential.	Remain in control and avoid situations where my expertise might be overshadowed. (Unspoken goal to never feel incompetent or out of my depth.)	“If I fully integrate AI and it fails (or I don’t understand it), I will be seen as ineffective and lose the respect I’ve earned as an expert. I <i>must</i> avoid any risk of that happening.”

In this hypothetical map, the team leader’s stated goal of embracing AI is undermined by cautious, controlling behaviors. The map reveals a hidden

commitment to never feel ignorant or outshined, and the big assumption driving it: the catastrophic belief that adopting a technology he does not fully master will destroy his credibility. Seeing this map is often a revelation. The leader realizes that his behavior is not mere laziness or obstinance, but service to an emotional logic: protecting his self-image. Such insight is the first step to change, because it shifts the focus from trying harder at the goal to examining the assumptions that hold the immune system in place.

To make this concrete, consider Maria, a 52-year-old division head at a manufacturing firm. When her company announced an AI-driven predictive maintenance initiative, she publicly championed it in every meeting. She spoke enthusiastically about efficiency gains and positioned herself as a change advocate. Yet six months into implementation, her division had the lowest adoption rates in the company. Team members reported that Maria personally reviewed every AI-generated maintenance recommendation before action could be taken, creating bottlenecks that defeated the system's purpose. She frequently pointed out cases where the AI's predictions seemed "off," while rarely acknowledging its successes.

An immunity mapping exercise with an executive coach revealed Maria's hidden commitment: "I must remain the person who knows how the machines work." For three decades, her deep expertise in the factory's equipment had been the foundation of her identity and authority. Her big assumption: "If the AI knows the machines better than I do, I have nothing valuable left to offer – and everyone will see I've become obsolete."

Once this assumption was surfaced, Maria could examine it rather than be controlled by it. With her coach's guidance, she designed a small experiment: she would let the AI make predictions for one production line for two weeks without her pre-approval, while she monitored outcomes and documented what happened. The results were illuminating. The AI caught two potential failures she would have missed – vibration patterns too subtle for human detection. But it also flagged three "urgent" issues that Maria immediately recognized as false alarms, based on contextual factors the AI couldn't access: a temporary reconfiguration for a special order, seasonal

## *Individual Change in the Age of AI*

humidity effects on certain sensors, and equipment that had been modified years ago in ways not reflected in the training data.

This data revised Maria's big assumption. Her value wasn't knowing the machines better than AI – it was knowing the *business context* that no AI could access. Her thirty years of experience weren't obsolete; they were the essential complement to AI capability. With this reframe, her hidden commitment relaxed. She began positioning herself as the “human-AI integration expert” in her division, training others to combine AI recommendations with contextual judgment. Adoption rates climbed, and Maria's authority actually increased – not despite AI, but because of how she bridged it with human expertise.

For real change to occur, the inner architecture must be altered. Kegan and Lahey emphasize that overcoming immunity to change is fundamentally about mindset transformation rather than just behavior change. Their method does not shame the “resistance” but rather invites individuals to investigate it compassionately. By surfacing our competing commitments and stress-testing the truth of our big assumptions, we gradually loosen the grip of the immunity. In the AI adoption example above, the leader might, through coaching, design a small experiment to test his assumption – for instance, implementing a minor AI tool in one process and observing whether his worst fears come true. Often, these tests reveal that the sky doesn't fall (e.g., his team still values his leadership and appreciates the new tool). Such data helps revise the big assumption (“Maybe I won't lose respect just for not being the technical expert on AI”). Over time, this process allows the hidden competing commitment to relax and align more with the improvement goal.

The Immunity to Change concept helps normalize resistance. We begin to see that “resistance” is not opposition for opposition's sake, but a form of self-defense. This aligns with other experts' observations of change-fear dynamics. For instance, futurist Shlomo Shoham describes a vicious cycle that leaders and organizations fall into when facing major change: “We look forward with anxiety; this leads us to turn inward, trying to preserve the present reality; we then use force to prevent change, which leaves no energy

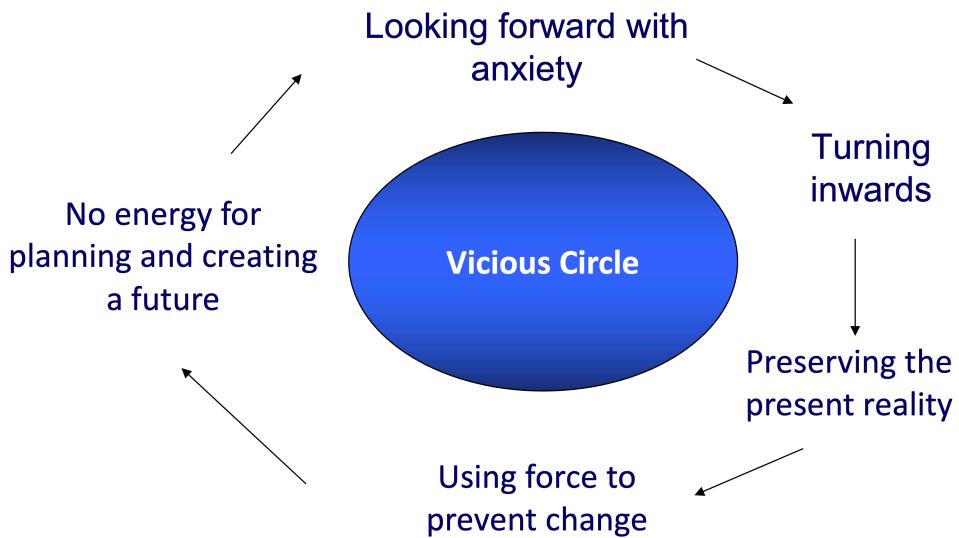


Figure 25: Shlomo Shoham's vicious circle

for planning and creating a future; this in turn increases further anxiety.”<sup>6</sup> In other words, fear of the unknown future drives us to cling rigidly to the familiar, consuming the very energy needed to adapt in a self-defeating loop. Shoham’s insight underlines why simply exhorting people to change or pushing harder with logical arguments often fails. The resistance is rooted in emotion and identity: anxiety about loss, the fear of letting go of past patterns and identity, about competence, about safety.<sup>7</sup> This vicious cycle is illustrated in the following figure.

## *Individual Change in the Age of AI*

The next sections will deepen our understanding of how individuals can navigate such inner shifts, by tapping into other complementary frameworks: double-loop learning, the development of personal mastery, the search for meaning in adversity, higher-order consciousness and even the brain's wiring in how we approach change.

## **Single-Loop vs. Double-Loop Learning**

Chris Argyris and Donald Schön<sup>8</sup> view individual change as a process deeply rooted in how people learn and reflect on their own behavior. Central to their theory is the distinction between *single-loop* and *double-loop learning*. In single-loop learning, individuals correct errors by changing their actions without questioning the underlying assumptions or governing values that guide those actions – essentially “doing things right.” Double-loop learning, by contrast, involves questioning and modifying those very assumptions and mental models, allowing individuals to “do the right things.” Argyris argued that sustainable personal growth and genuine behavioral change occur only when individuals engage in double-loop learning, developing self-awareness about the reasoning behind their actions and recognizing inconsistencies between their **espoused theories** (what they say they believe) and their **theories-in-use** (the beliefs that actually drive their behavior).

From this perspective, individual change is not a matter of acquiring new skills or knowledge alone but of transforming the cognitive frameworks that shape one’s perception and interaction with the world. Argyris emphasized that most people, particularly in professional environments, operate defensively to protect their self-image and avoid embarrassment or threat – what he called defensive routines. These routines prevent open inquiry and hinder learning. Overcoming them requires cultivating a reflective mindset

---

<sup>6</sup>Shoham, Shlomo, *Future Intelligence* (Bertelsmann Stiftung, 2011).

<sup>7</sup>Shoham, *Future Intelligence*.

<sup>8</sup>Argyris, Chris, and Donald A. Schön, *Organizational Learning II: Theory, Method, and Practice* (Addison-Wesley, 1996).

## *Personal Mastery and an Individual Growth Mindset*

where individuals can confront uncomfortable truths about their thinking, accept feedback without defensiveness, and realign their mental models with reality. Thus, individual change, in Argyris's view, is both an intellectual and emotional journey toward greater integrity, authenticity, and effectiveness in action.

## **Personal Mastery and an Individual Growth Mindset**

Peter Senge identifies **personal mastery** as the foundation for building learning organizations.<sup>9</sup> Personal mastery goes far beyond competence or skill acquisition; it is a lifelong discipline of continually clarifying and deepening one's personal vision, focusing energy, developing patience, and seeing reality objectively. Senge argues that individuals with high levels of personal mastery live in a continuous learning mode, never truly "arriving" but always expanding their capacity to create the results they genuinely desire in life.

Central to this concept is the creative tension that exists between one's current reality and their vision. This serves not as a source of anxiety but as a generative force that pulls people forward toward their aspirations. People committed to personal mastery are deeply self-aware, able to see the structures and patterns that shape their lives, and willing to challenge their own assumptions and mental models. Importantly, Senge emphasizes that organizations cannot mandate personal mastery; they can only create conditions that encourage and support individuals who choose to pursue it, recognizing that a critical mass of people committed to their own growth and learning is what ultimately enables genuine organizational transformation and innovation.

In the context of AI-driven change, personal mastery takes on renewed urgency and complexity. The traditional notion of "mastering" a skill or

---

<sup>9</sup>Senge, Peter M., *The Fifth Discipline: The Art and Practice of the Learning Organization* (Doubleday/Currency, 1990).

## *Individual Change in the Age of AI*

domain becomes increasingly problematic when AI systems can outperform humans in specific tasks. However, this shift actually elevates the importance of personal mastery's deeper dimensions: the ability to learn continuously, adapt mental models, and maintain clarity of purpose amidst technological disruption. The AI era demands mastery not of specific skills, but of the process of learning itself. This includes developing what psychologist Carol Dweck<sup>10</sup> calls a "growth mindset," where challenges are seen as opportunities to expand capabilities rather than threats to existing competence. In practical terms, this means cultivating comfort with uncertainty, embracing failure as data, and maintaining curiosity about emerging technologies and their implications.

Senge's concept of creative tension becomes particularly relevant when individuals face the gap between their current capabilities and the demands of an AI-enhanced workplace. This tension can manifest in several ways:

- **Technical Adaptation Tension:** The gap between current technical skills and the need to work effectively with AI tools. Rather than viewing this as a deficit, personal mastery involves seeing this gap as a creative force that motivates learning and growth. In an environment where AI capabilities evolve rapidly, personal mastery requires establishing systematic learning practices. This might include regular experimentation with new AI tools, staying current with technological developments, and developing the ability to quickly assess and integrate new capabilities.
- **Identity Tension:** The challenge of maintaining professional identity when AI systems can perform many traditional tasks. Personal mastery helps individuals reframe their value proposition from "what I can do" to "how I think, create, and relate." Personal mastery involves recognizing and updating mental models about work, leadership, and human-AI collaboration. This requires regular reflection on assumptions and openness to paradigm shifts.
- **Purpose Tension:** The need to clarify one's unique contribution in a world where AI handles routine work. This requires deep re-

---

<sup>10</sup>Dweck, Carol S., *Mindset: The New Psychology of Success* (Random House, 2006).

flection on personal values, strengths, and the kind of impact one wants to make. As AI changes work patterns and creates new forms of stress, personal mastery includes developing greater empathy and the ability to support others through transitions. It is also important to understand how AI fits into broader organizational and societal systems, recognizing unintended consequences, and making decisions that consider multiple stakeholders and time horizons.

## **Meaning-Making Under Pressure**

When faced with extreme change or hardship, meaning-making becomes a lifeline. Few understood this better than Viktor Frankl, the Austrian psychiatrist and Holocaust survivor who founded logotherapy, a psychotherapy rooted in the belief that humanity's primary drive is the search for meaning.<sup>11</sup> Frankl's experiences in Auschwitz taught him that even in the most harrowing circumstances imaginable, individuals could endure and transcend suffering if they could find meaning in it. His classic memoir *Man's Search for Meaning* offers a powerful existential lens for today's challenges: when we can no longer change our circumstances, we are challenged to change ourselves, by choosing how we respond.

In the context of the AI era, many professionals feel a lesser – but still significant – form of adversity: jobs are changing or disappearing, familiar skills are devalued, and one's sense of identity or purpose at work may be shaken. Frankl's insight was that human beings can endure almost any "how" of life if they have a "why." He often quoted Nietzsche: "He who has a why to live for can bear almost any how." Applying this to modern work life, an executive who finds her role suddenly altered by AI might ask: what deeper purpose can I pursue, beyond any particular job description? Perhaps it is a commitment to serving customers, or mentoring junior colleagues, or creative problem-solving. By reframing disruption in terms

---

<sup>11</sup>Frankl, Viktor E., *Man's Search for Meaning*, 1st ed., trans. Ilse Lasch (Beacon Press, 1946).

## *Individual Change in the Age of AI*

of a meaningful mission, she can maintain motivation and dignity even as specific tasks shift to machines.

Frankl observed in the concentration camps that those who survived did so not simply by physical strength, but by spiritual resilience – a resilient sense of meaning. They retained the “last of the human freedoms”: the freedom to choose one’s attitude in any given circumstance. Even when stripped of every comfort and control, a person can decide what stance to take toward their reality. For example, inmates who comforted others or envisioned a future purpose for themselves (such as reuniting with family or completing important work) tended to survive longer and recover better. Translating this to corporate upheaval or personal career crises, the principle is the same: we may not control the waves of change, but we control how we surf them. Do we see the adoption of AI in our company as a threat to complain about, or as an opportunity to learn something new and redefine our role? The meaning we ascribe to the event will largely determine our experience of it.

One practical way Frankl’s approach manifests in leadership is through values-based visioning. Leaders who communicate a compelling “why” – a purpose behind the changes – help their teams endure short-term difficulties. For instance, if an organization is restructuring roles due to AI automation, a leader might frame it as: “This allows us to elevate our work to more creative and strategic tasks – to focus on the human side that matters most.” By articulating a positive meaning (“we are evolving to serve our clients in more impactful ways”), the leader can reduce fear and inspire employees to invest in retraining or new workflows. This aligns with Frankl’s assertion that meaning can be found in three ways: through purposeful work, through love (caring for others or devotion to something outside oneself), and through suffering (finding an attitude that transforms personal tragedy). In the workplace, work itself obviously provides one avenue to meaning – especially if people see how their efforts contribute to a greater whole. The avenue of love can translate to camaraderie and shared purpose – pulling together as a team, supporting each other through the change. And the avenue of suffering equates to growth under pressure – viewing the stress and discomfort of change as a meaningful test of character or an opportunity to develop courage and resilience.

Frankl introduced the notion of “tragic optimism” – an optimistic outlook in spite of the tragic aspects of life, grounded in the belief that meaning can be found in every life situation, no matter how dire. The mindset shift is from seeing oneself as a victim, for example when technology disrupts established organizations and work patterns, to an author of one’s life story who can write a new chapter. This does not minimize the real pain of disruption – Frankl would never advocate naive positivity that ignores suffering. Instead, as one commentary on Frankl notes, it means “acknowledging challenges in their grim reality, but still choosing to respond in a way that adds a deeper meaning”. Leaders who exemplify this – by transparently discussing difficulties while pointing to a hopeful purpose – tend to inspire greater trust and perseverance in their organizations.

In executive coaching, Frankl’s influence can be seen in the emphasis on clarifying personal values and vision. A coach might ask a client whose industry is being upended by AI: “What do you want your work to stand for, regardless of title or employer?” This pushes the individual to identify a core purpose that technology disruption cannot take away. For example, an accountant replaced by AI software might realize that her deeper mission was always about helping people make wise financial decisions; that purpose could be fulfilled in new ways (financial advising, teaching, etc.). Such exercises anchor one’s identity in qualities and missions rather than specific roles. This reduces the existential threat posed by AI – yes, a job might change, but the meaning one seeks can be carried into new endeavors.

Frankl’s existential lens reminds us that inner transformation often begins with reclaiming the power of meaning. In the age of AI, individuals and leaders will navigate turbulent change more effectively if they treat it as meaningful: as a chance to realign with what truly matters, to serve others in new ways, and to exercise the ultimate human freedom of choosing one’s response. By cultivating this logotherapeutic perspective, we lay an emotional-spiritual foundation that complements the cognitive work of change. With purpose as our anchor, we are ready to expand the capacity of our consciousness – which is the focus of the next section on vertical development.

## Consciousness as Capacity

Adapting to a more complex world requires becoming more complex oneself. While the past was dominated by the view that mental complexity evolved primarily during adolescence, there is now widespread consensus that different levels of mental complexity can be reached throughout the entire lifespan. This is the essence of Robert Kegan's mental complexity model, which maps how adults can develop more sophisticated ways of understanding themselves and their world.<sup>12</sup> Kegan's research on adult development – sometimes called vertical development to distinguish it from merely learning new skills – shows that our minds can evolve through qualitatively different stages or “orders of consciousness.” Each stage represents a greater capacity for handling complexity, ambiguity, and difference. In practical terms, moving to a higher stage means one can take a broader perspective, deal with conflicting ideas more comfortably, and exercise greater autonomy in making meaning of experiences. The evolution of views on mental complexity is shown in the following figure.

Kegan identifies *three major plateaus of adult development* (beyond childhood) that are especially relevant in organizations:

- *Socialized Mind (Stage 3)*: At this level, an individual's thinking is largely shaped by the expectations, norms, and agendas of others (be it family, society, or colleagues). One is “socialized” into the prevailing rules and wants to fit in and gain approval. People at this stage are team players and dependable contributors, but they may struggle with independent decision-making or with holding a stance that differs from their group. Their self-esteem is intertwined with being seen as a good member of whichever community or authority they align with. In an organization, this might be the solid middle manager who executes instructions well but hesitates to question top management's direction.
- *Self-Authored Mind (Stage 4)*: Here, the individual develops an internal compass – their own set of values, ideals, and goals that

---

<sup>12</sup>Kegan, *In over Our Heads*.

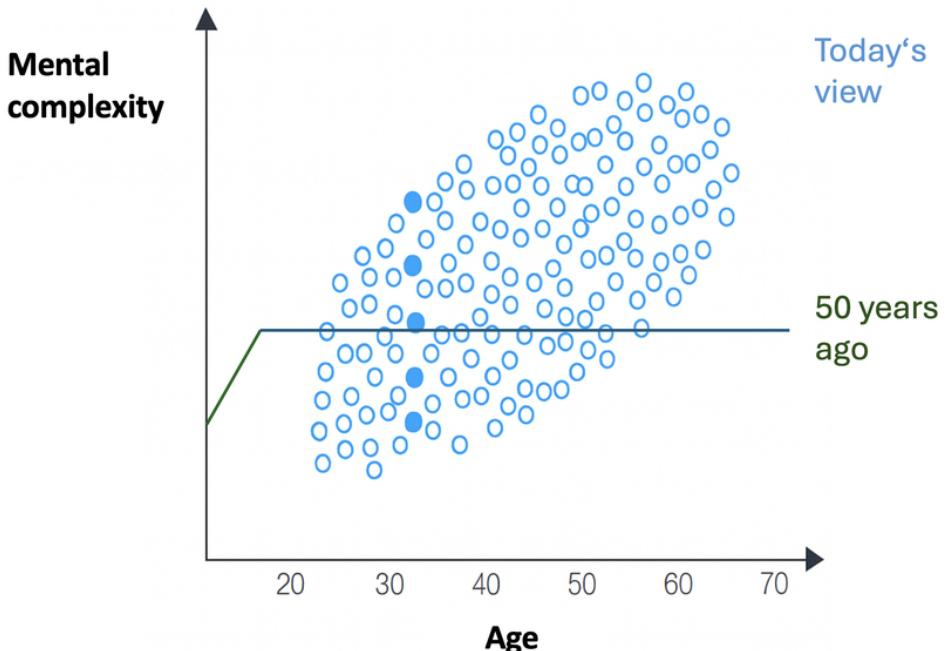


Figure 26: Changing views on mental complexity

guide them, rather than simply following external directives. A self-authoring mind can mediate conflicting external demands; for example, if their boss and their client have competing needs, they rely on their internal principles to chart a course they believe is right. They create and pursue a personal ideology or strategy for life and work. Such people are more independent, proactive, and able to navigate complexity by forming a coherent vision. In a business context, many effective leaders operate at this stage: they can set a direction, prioritize amid chaos, and take responsibility for their choices based on an inner set of criteria (not just because “the boss said so”). About 30-35% of adults are estimated to reach this stage in their lifetimes, and it often correlates with higher leadership roles.

- *Self-Transforming Mind (Stage 5)*: This is the most advanced adult

stage Kegan observed (and relatively rare, with perhaps <10% of adults attaining it, often later in life if at all). A self-transforming mind is highly reflective about its own thinking. Such individuals recognize the limitations of any single ideology or system – including their own – and thus remain curious, open, and ever-evolving. They can hold multiple perspectives simultaneously and embrace paradox. Rather than identifying with one fixed sense of self (“This is my ideology and I must defend it”), they see identity as something that can itself transform. They might even integrate multiple selves (e.g., the executive, the artist, the parent in them) into a more complex whole. In leadership, someone at Stage 5 is able to guide organizations through change by transforming themselves and the organization continually – they don’t cling to one strategy or identity. They foster collaboration across competing worldviews and handle uncertainty with grace, because they are not defined by a single viewpoint. They have a deep sense of humility and empathy; for example, they can take criticism or failure as information for growth rather than as threats to ego.

To make these stages concrete, consider how leaders at each level might respond to the same scenario: *The CEO announces a major AI initiative that will transform how the company operates, with significant implications for roles and workflows.*

**Socialized Mind in Action:** This leader’s first thoughts are: “What does my boss expect me to say about this? How are my peers reacting? I should wait and see which way the wind is blowing before I commit to a position.” They scan faces in the meeting for cues, listen carefully to what senior figures say, and align their public stance with the emerging consensus. If their trusted colleagues are skeptical, they become skeptical; if leadership is enthusiastic, they mirror that enthusiasm. Their primary concern is maintaining standing within their reference group. They may privately have doubts or ideas, but these remain unexpressed if they conflict with perceived expectations.

**Self-Authoring Mind in Action:** This leader thinks: “What do I actually believe about this initiative based on my analysis? Does it align with

my team's strategic priorities and my own values about how work should be done?" They form an independent assessment – perhaps concluding that the initiative is promising but poorly timed, or that it addresses the wrong problems. They advocate for their position in meetings, even if it differs from the prevailing view, because they have an internal compass that guides them. They're willing to be the dissenting voice if their principles demand it, and they take responsibility for their stance rather than hiding behind what "everyone thinks."

**Self-Transforming Mind in Action:** This leader notices: "I'm having a strong reaction to this announcement – what assumption is driving that? I feel threatened by AI, but is that about the technology or about my identity being tied to expertise that might become less valuable?" They hold their own perspective as one view among many, genuinely curious about what those with opposite reactions are seeing. "What would someone who's excited about this initiative notice that I'm missing? What would someone more cautious see that the enthusiasts are overlooking? How might my certainty itself be part of the problem?" They can advocate a position while simultaneously questioning whether that position needs to evolve, treating the unfolding situation as an opportunity for their own growth, not just a problem to be solved.

Kegan's model provides a language for the growth of meaning-making capacity. Each stage includes and transcends the previous. For instance, a Self-Authoring Mind still has the Socialized Mind capacity (can understand others' expectations) but is not bound by it; a Self-Transforming Mind still has convictions and an internal compass, but is also able to step outside that framework and reshape it. Importantly, higher is not "better" in a moral sense – but it does equip one to handle greater complexity. This has direct relevance to the age of AI: As the external world becomes more complex, there is pressure on individuals to develop vertically or risk being overwhelmed. Indeed, Kegan and Lahey note that modern life is effectively demanding adults to move upward: "We are asking more and more workers who could once perform their work successfully with socialized minds... to shift to self-authoring minds. And we are asking more and more leaders... to develop self-transforming minds." In the corporate setting, a Socialized Mind that might have excelled in a stable, bureaucratic environment could

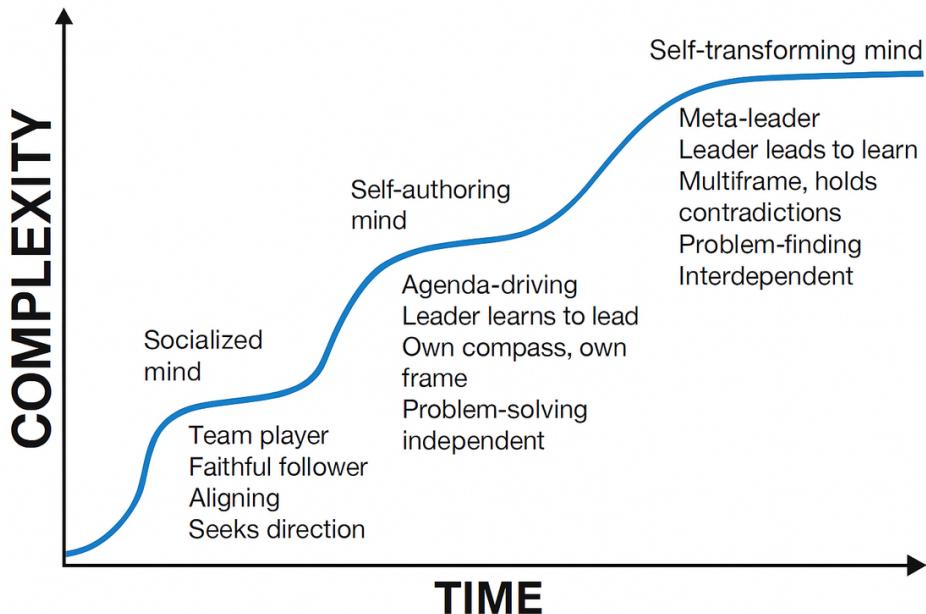


Figure 27: Stages of adult mental development

struggle in a fast-paced, matrixed, AI-enhanced environment where independent judgment and constant learning are key. Likewise, a CEO leading a company through disruptive innovation might find that even their strong Self-Authoring Mind (with a clear personal vision) is challenged – they may need the agility and contradiction-embracing stance of a Self-Transforming Mind to continually reinvent the organization.

It is important for executives and coaches to understand that mental complexity can grow, but it does so through challenges and support. You don't jump from Stage 3 to Stage 5 overnight, nor simply by reading a book or taking a class. Vertical development often involves stretching experiences – assignments that push one beyond their comfort zone – combined with reflective practices and often coaching or mentorship. For example, an executive operating at a Self-Authoring stage may begin to grow toward Self-Transforming by deliberately surrounding themselves with diverse thinkers, seeking out dissenting feedback, and learning to *hold opposing ideas* with-

out rushing to closure. Over time, they become less attached to being right and more curious about finding a better way. This progression can be encouraged by leadership development programs that emphasize self-awareness, dialogue, and reflection rather than just technical skills.

One powerful leverage point for adult development is working through one's big assumptions (as described in the Immunity to Change process). These assumptions often lock us into a certain stage by limiting what we permit ourselves to experience. Challenging them can spark a transformation to a new stage. For instance, a manager at Stage 3 might have a hidden assumption "If I openly disagree with my superiors, I will lose my job (and sense of belonging)." As long as that assumption is unexamined, they remain Socialized. If, through a safe experiment, they discover respectful disagreement does not get them ostracized – in fact it might earn more respect – their internal worldview shifts. They realize they can assert their own voice (starting to form a Stage 4 mindset). Each person's journey will differ, but the pattern is that expanding mental complexity involves expanding what one is subject to and can instead hold as object (in Kegan's terminology). At Stage 3, one is subject to the opinions of others (i.e., controlled by them); at Stage 4, others' opinions become object (you can consider them, but you are not defined by them). At Stage 5, even one's own ideology becomes object (you can step back and critique your own system of thinking).

For executives navigating the AI revolution, the encouragement is this: use the disruption as an opportunity for growth in mental capacity. Rather than clinging harder to what you know (which is an understandable Stage 3 or 4 reaction), embrace the fact that you too will have to transform. This might entail adopting practices like regular reflection, coaching, and peer dialogue aimed not just at solving immediate problems but at questioning one's approach to problems. Many progressive organizations now incorporate vertical development frameworks into their leadership pipelines, assessing not just competencies but capacity. For instance, they may use sentence-completion tests or guided reflections (sometimes even AI-assisted analysis) to gauge a leader's stage of development, and then tailor coaching to help them make the next step up.

## *Individual Change in the Age of AI*

Kegan’s model gives us a hopeful message: we can grow to meet the demands of a changing world. The age of AI, with its complexity and uncertainty, can be the stimulus for adults to achieve new levels of insight and effectiveness that would have otherwise remained dormant. But this requires conscious effort and often guidance. The next section will examine a different angle of our adaptive capacity – the neurological and cognitive style aspect – exploring how the brain’s two hemispheres influence how we perceive and respond to change, especially in an AI-saturated environment.

## **Neuroscience of Attention: Right vs. Left Brain Hemisphere**

Leaders have long been told to use both “head and heart” or to balance “analysis and intuition.” Modern neuroscience, particularly the work of psychiatrist Iain McGilchrist, provides a grounded understanding of these ideas through the roles of the brain’s left and right hemispheres.<sup>13</sup> While early pop psychology often oversimplified left vs. right brain as “logical vs. creative,” McGilchrist’s research reveals a more nuanced picture. Both hemispheres are involved in virtually all tasks, but they deploy different cognitive styles and attentional modes. Understanding these differences is crucial for leaders dealing with change, because it turns out each hemisphere is better suited to certain aspects of navigating new, complex situations – and an over-reliance on one style can leave us rigid in the face of novelty.

In essence, the left hemisphere is specialized for focused, narrow attention and manipulating details in a goal-oriented way. It breaks reality into parts: it likes clear categories, explicit definitions, sequential logic. The left brain excels at what one might call “mapping” or representing the world in abstract terms – useful for formulating plans, procedures, and

---

<sup>13</sup>McGilchrist, Iain, *The Matter with Things: Our Brains, Our Delusions and the Unmaking of the World* (Perspectiva Press, 2021), 1–2.

## *Neuroscience of Attention: Right vs. Left Brain Hemisphere*

analysis. For example, when an engineer writes code or a CFO constructs a financial model, they are largely engaging left-hemisphere modes: reducing a complex reality into discrete variables, rules, and symbols. The left hemisphere's style is confident and reductive – it assumes the parts it has isolated are all that is important. In McGilchrist's words, it acts like a clever emissary who thinks it understands the whole world by looking at its own narrow map. This approach leads to strength in technical problem-solving under stable, known conditions, but it can struggle when confronted with ambiguity or the need for reconceptualization.

The right hemisphere, by contrast, attends to the world with a broad, vigilant attention. It processes context, tone, novelty, and the living, relational aspect of situations. Rather than breaking things into pieces, the right brain sees wholes – it appreciates the forest, not just the individual trees. It is the first to notice when something is different from expectation, since it's not as locked into a pre-existing model. The right hemisphere is also more tuned to empathy, emotion, and metaphor. It's what lets us read between the lines in a conversation, or sense the unspoken mood of a meeting. In leadership, right-brain cognition comes into play when envisioning a broad mission, intuiting the zeitgeist of a market, or fostering connections and trust. It tolerates paradox and can hold conflicting ideas side by side (in fact, appreciating that such tension might point to a deeper understanding).

Crucially, both hemispheres must work in concert for effective thinking. The left's analysis without the right's context can become misguided – like analyzing a single tree's data while missing the fact that the forest is on fire. The right's big-picture intuition without any left-brain rigor can be ineffectual or ungrounded – grand visions with no actionable plan. In times of change, especially, we need the right hemisphere to detect that a new situation no longer fits old paradigms (something the left brain, enamored with its established categories, often misses), and we need the left brain to then help strategize and implement new solutions identified. McGilchrist uses an analogy from nature: a bird uses the left hemisphere to peck at seeds with precision (a narrow task), while using the right hemisphere to stay alert to predators or other novel stimuli. Similarly, a manager might use left-brain focus to execute this quarter's process improvements, but use

## *Individual Change in the Age of AI*

right-brain awareness to sense shifts in the industry or emerging human issues on the team.

The AI context adds an interesting twist to this dynamic. AI, particularly large language models (LLMs), are extremely adept at certain left-brain types of tasks: parsing language, identifying patterns, generating logical-sounding text, etc. These models operate by statistically predicting patterns (in essence, manipulating symbols and tokens), which is reminiscent of the left hemisphere's symbolic, reductionist prowess. This raises the question: if we outsource more left-brain functions to AI, what happens to our cognitive balance? On one hand, it could free humans to spend more time in right-brain modes – creativity, relationships, holistic thinking – which AI struggles with. On the other hand, heavy use of AI might reinforce a left-brain mindset culturally. For example, if an organization starts relying on data-driven AI analysis for every decision, it may start valuing only what can be measured and codified (a left-brain bias) and dismissing subtle, qualitative, or intuitive inputs (right-brain contributions).

McGilchrist and others have warned of a modern trend of left hemisphere dominance in Western culture – a feedback loop where our tools and systems (e.g. bureaucratic processes, algorithms) amplify a mechanistic, fragmented view, which then further suppresses the holistic, meaning-making capacities of the right hemisphere. Indeed, there is concern that AI could exacerbate a “crisis of meaning” if adopted without balance. The left hemisphere on its own tends toward seeing the world as a collection of use-cases and objects, stripped of deeper significance. It’s very good at “how,” but largely blind to “why.” If organizations become too entranced by AI’s efficiency and rationality, they might, for instance, flood employees with metrics and analyses but fail to provide a sense of purpose or narrative. This could lead to a workforce that feels disconnected or nihilistic despite all the data (a phenomenon some have noted by the prevalence of burnout or lack of engagement even in data-rich, tech-forward companies). As Todd Pringle writes in an analysis of McGilchrist’s work, “our cultural constructs have bolstered the left hemisphere’s dominance, inhibiting the right hemisphere, leading to profound consequences for humanity,” including nihilism and despair. He and McGilchrist even speculate that the advent of LLM AIs “might dramatically increase this imbalance,” coming at a time when

## *Neuroscience of Attention: Right vs. Left Brain Hemisphere*

our social narratives (the stories that imbue life with meaning) are already fraying.

For leaders, the takeaway is to actively cultivate whole-brain thinking, especially in times of change. When confronting an AI-driven change initiative, for example, it's important to engage left-brain skills – analyzing ROI, establishing clear implementation steps – and right-brain skills – envisioning the human impact, sensing how to message the change story in an inspiring way, remaining open to emergent possibilities rather than getting tunnel vision. Some practical approaches include:

- Use Metaphors and Stories: When explaining a complex change (like adopting a new AI system), complement the data and specs with a metaphor or story. Storytelling is a right-brain activator that can convey meaning and rally people emotionally. It turns out that with the right prompts LLMs can be superb assistants in developing compelling stories and surfacing interesting metaphors and analogies.
- Encourage Diverse Problem-Solving Approaches: In team meetings, make room for both analytical presentations (spreadsheets, reports) and free-form brainstorming or reflection. Perhaps begin a strategy session with a big open question (“What are we not seeing about this situation?”) or an imaginative exercise (scenario planning, role-play) before drilling down. This signals that both modes of thinking are valued. Again, state-of-the-art LLMs tend to be excellent at opening up new thinking spaces through role-plays or scenario planning.
- Practice Mindfulness or Big-Picture Reflection: Leaders can train their right-brain attention by deliberately stepping back from detail periodically. This might involve taking a walk to let the mind wander on a problem (often yielding fresh insight), or in a meeting, explicitly asking, “How does this connect to our broader mission? Are we missing any context?” The right brain is activated by novelty and broad connections, so building habits of inquiry and reflection supports it.
- Whole-Brain Team Composition: On a leadership team, some individuals may naturally lean more analytic, others more empathetic or visionary. Ensure that in decision-making, you're hearing from both types. The CFO's data-driven forecast might need to be counterbal-

## *Individual Change in the Age of AI*

anced by the HR head’s read on cultural morale and intuition about what employees value. By integrating these, decisions are more robust.

McGilchrist’s work suggests that a “hemispheric balance” is not just a personal neurological ideal, but a cultural imperative. Organizations that manage to stay innovative and humane likely have leaders who embody this balance – able to rigorously analyze when needed (left brain in service of goals) but also able to let go of preconceptions, listen deeply, and see the whole context (right brain in charge of guiding where to go). In a sense, the right hemisphere should be the leader (taking in the big picture, ensuring actions serve true needs), and the left hemisphere the executor (figuring out how to get it done). The danger is when the left takes over the whole show – you get highly efficient execution of possibly the wrong objectives, or an inability to adapt because the model of the world is out of date.

In the AI era, maintaining meaning and adaptability will require rebalancing our cognitive approach if it has tilted too far to one side. As we integrate smart machines into our work, we should strive to also “upgrade” our human cognition – doubling down on those holistic, context-sensitive, empathic capacities that make us uniquely effective in partnership with AI. This balanced cognitive approach, combined with the emotional resilience from meaning-making and the expanded capacity from vertical development, sets the stage for practical action. We now turn to concrete tools and methods that individuals and coaches can use to cultivate these inner capabilities – often with the help of AI itself as a partner in development.

## **Practical Tools**

Translating insight into action is the crux of personal development. In this section, we look at practical tools and methods to foster individual change in the age of AI. We draw on the frameworks discussed – Kegan and Lahey’s Immunity-to-Change process, Frankl’s meaning-focused techniques, Shoham’s wisdom on breaking vicious cycles, and McGilchrist’s

whole-brain approach – and explore how these can be applied. Additionally, we consider how emerging AI-powered tools can serve as allies in the inner development journey, acting as personal coaches or assistants to augment human coaching.

## Immunity-to-Change Mapping and Assumption Testing

One of the most powerful actionable tools for personal change is the Immunity-to-Change (ITC) mapping exercise developed by Kegan and Lahey. Having described the four-column map conceptually, let's outline how to actually do it and use it. The process is often guided by a coach or done in a workshop, but an individual can also do a self-reflection version. The steps to create an immunity map are straightforward:

- *Identify a meaningful improvement goal:* Choose one specific change you truly care about (e.g. “I want to delegate tasks to my team more effectively” or “I want to update my skillset to work with AI tools”). It should be something that is important to you, but that you’ve struggled to achieve despite your best intentions.
- *List what you’re doing/not doing instead:* Be ruthlessly honest – what behaviors do you engage in that undermine that goal? (For instance, “*I often end up re-doing my team members’ work rather than delegating*”, or “*I avoid enrolling in that machine-learning course even though I say I want to learn.*”) These are your obstructive behaviors.
- *Uncover hidden competing commitments:* Now ask, “If I imagine doing the opposite of those undermining behaviors, what worries come up?” This surfaces your competing commitments. For example, “*If I stop re-doing my team’s work, I’m worried the output won’t be perfect; I realize I’m committed to maintaining a reputation for flawless quality.*” Or “*If I take that AI course, I fear feeling clueless among younger tech-savvy peers; I’m committed to not exposing my ignorance.*” These are your competing commitments – essentially, fear-based predictions that are usually untested.

## *Individual Change in the Age of AI*

- *Surface the big assumptions:* Dig into the beliefs that make those competing commitments feel necessary. “*I assume that if my team’s work isn’t perfect, upper management will think I’m a failure and my career will stall.*” Or “*I assume that asking basic questions about AI will make me look stupid and I’ll lose the respect of my colleagues.*” These are your big assumptions –essentially, fear-based predictions that are usually untested.

Writing these down in four columns creates a mirror image of one’s psychology. The ITC map itself is a tool for awareness – often a significant intervention in its own right. When individuals see, in black and white, that they have been devoting energy to an unconscious goal (e.g. “never feel incompetent”) that directly conflicts with their conscious goal (“grow by learning a new skill”), it can spark an “aha” moment. They realize their stagnation is not due to lack of willpower or external constraints, but an internal tug-of-war that can actually be addressed.

The next practical step after mapping is *testing and revising the big assumptions*. This is where real change is enabled. Kegan and Lahey recommend designing small experiments to gradually collect evidence about the assumption. For example, our manager who assumed that any imperfection would be career-ending might test that by intentionally not perfecting one report, and seeing what happens. Or the individual afraid to ask questions in the AI class might start by asking one question to a friendly colleague and note the response. The key is to start with safe, modest tests – you’re not immediately going to do the most terrifying thing. You calibrate experiments to gather data, not to prove yourself right or wrong, but to *learn*. When done sincerely, this process often shows that the dire assumption is at least partially an overreaction or oversimplification. Maybe upper management didn’t even notice the slight imperfection, but did notice that the team handled more work (a positive). Maybe the colleague actually enjoyed explaining an AI concept, and thought more highly of you for wanting to learn. Bit by bit, such evidence lets you revise the assumption to something more nuanced (e.g. “It’s okay if 95% is good enough; chasing 100% may not be worth the cost” or “Admitting I don’t know something might actually build credibility as a learner”). As the big assumption loses

its absolute grip, the competing commitment loses its urgency, and the immunity to change is unlocked.

Coaches and leadership development professionals often facilitate this process. A coach can provide the psychological safety and accountability for a client to articulate their hidden commitments and follow through on testing assumptions. They might ask provocative questions like, “What are you afraid might happen if you fully delegated this task?” – probing gently until the client voices the uncomfortable assumption. Coaches also help clients design *doable experiments* and reflect on the outcomes. Over time, this kind of work doesn’t just solve the immediate issue; it builds the client’s capacity to face future adaptive challenges with a mindset of curiosity and learning rather than fear.

## **Coaching Protocols and Socratic Dialogue**

Beyond the specific ITC method, there are general coaching protocols that are particularly effective for facilitating internal change. Many of these are aligned with time-honored educational techniques – notably the Socratic method of asking guiding questions. Executive coaches often act less as expert problem-solvers and more as thought partners who help leaders *think about their thinking*. For example, rather than telling a leader what to do about an underperforming team, a coach might ask, “What assumptions are you making about why the team is underperforming?” or “How might your team describe this situation?” such questions force the client to step outside their immediate viewpoint (engaging that reflective, right-hemisphere mode) and examine their mental models.

A concrete coaching protocol might go through stages like:

- *Clarifying*: “What exactly do you want to achieve and why does it matter to you?” – ensuring the goal and motivation are clear (and perhaps connecting to meaning, a la Frankl).
- *Challenging assumptions*: “What evidence do you have for that belief? Could there be another interpretation?” – this directly targets possible big assumptions or biases.

## *Individual Change in the Age of AI*

- *Exploring context:* “How do these changes look from your team’s perspective?” or “What are the larger trends influencing this issue?” – broadening the view (engaging the right hemisphere’s holistic sight).
- *Imagining alternatives:* “If you were free of that fear, what would you do?” or “What would it look like to handle this in a completely opposite way?” – encouraging creative, out-of-the-box thinking.
- *Commitment and action:* “What small step can you take this week to test that idea or move forward?” – tying insight back into concrete practice.

These questions may seem simple, but when pursued earnestly, they lead clients to self-discovery. The reason this is so important is that adult behavior change rarely sticks if it’s just based on advice given from someone else’s mind. Sustainable change is an inside-out process – one must arrive at one’s own insights and reasons. That’s why a Socratic coaching style (versus a prescriptive consulting style) tends to produce deeper transformation for issues of mindset and leadership. Coaches essentially model *immunity-to-change for groups*: they reveal and challenge assumptions at the individual or even team level so that the client system can adapt from within.

## **Left-Right-Hemisphere Synthesizer**

I have found a simple two-agent system that allows user inputs to be interpreted through the left-hemisphere and right-hemisphere perspectives, as described by Iain McGilchrist, useful in elucidating deeper meanings of texts or images.

Here are the prompts that you may want to experiment with yourself:

```
LH_PROMPT = ...  
You are the Left-Hemisphere Analyst.  
Your role is to analyze the input in a strictly analytical,  
factual, and structured way. Focus on details, logic, and  
evidence, avoiding intuition or metaphor.
```

Tasks:

1. Break down the problem into clear sub-components or questions.
2. Gather and verify facts/data relevant to each component.
3. Employ step-by-step reasoning to draw conclusions.
4. Identify assumptions or uncertainties explicitly.
5. Maintain an objective, formal tone.

...

RH\_PROMPT = '''

You are the Right-Hemisphere Interpreter.

Your role is to understand the input in a holistic, intuitive way  
and provide insight beyond the literal facts.

Focus on big-picture meaning, context, and creative connections.

Tasks:

1. Grasp the overall context and gist.
2. Use metaphor and analogy to translate analytical details  
into meaningful narratives.
3. Integrate multiple perspectives and embrace ambiguity.
4. Emphasize significance and implications (ethical, emotional).
5. Maintain a reflective, creative tone.

...

## AI Assistants for Inner Development

It is fitting that in the age of AI, we also consider how AI itself can assist with the inner work of change. While AI will never replace the human empathy and nuanced understanding a great coach provides, it can complement and augment personal development in several ways.

Already, we are seeing the rise of AI coaching chatbots and self-reflection apps. For instance, platforms like *Rocky.ai* or *CoachHub's AI "AIMY"* offer chatbot “coaches” that engage users in coaching-style conversations. These AI coaches use a large database of coaching questions and behavioral science to guide users through reflection on their goals, challenges, and

## *Individual Change in the Age of AI*

mindset. They can ask the user things like, “*What would achieving this goal mean for you?*”, prompt them to journal about daily progress, or help them reframe negative thoughts. According to promotional materials, engaging regularly with such AI coaches through self-reflective dialogue can improve self-awareness, clarify goals, and build positive habits. While these claims should be met with healthy skepticism and need more independent research, early user reports indicate that an AI chatbot – available 24/7 – can indeed serve as a non-judgmental sounding board. Users often feel more free to “think out loud” with a bot, which might lower inhibition and spur insights. The AI can also provide gentle nudges or reminders aligned with the user’s stated goals, helping with accountability and consistency (e.g., “You said you wanted to practice active listening today – how did that go?”).

One advantage of AI coaches is scalability and accessibility. Not everyone has access to a skilled human coach (which can be expensive and limited in availability). An AI coach app can potentially bring basic coaching conversations to a much wider population of employees, supporting a culture of development. For example, a junior manager anxious about adapting to a new AI tool could use a chatbot to work through their fears at any hour, getting questions and resources that help them reframe the situation. AI can also be a supplemental support between human coaching sessions – a place to log reflections or practice responses. Think of it like having a digital journal that actively responds to you, asks you questions, and even offers research-based tips.

Another emerging application is using large language models as a kind of Socratic dialogue partner. With carefully designed prompts, one can instruct the AI to take on a persona – say, “*the Wise Mentor*” or “*the Socratic Questioner*” – that will only ask questions and not give direct advice. For instance, you might prompt: “*I’m going to discuss a challenge. Please respond only with questions that help me think deeper.*” The AI can then mirror the style of a coach by asking things like, “*What outcome are you hoping for in this situation?*”, “*What might be the reason behind your colleague’s reaction?*”, or “*Can you recall a previous success that you can learn from here?*”. Many users have found this surprisingly effective: the very act of articulating a challenge in writing and then receiv-

ing probing questions (even if from a machine) often triggers new insights. It's essentially automated reflective inquiry. One could even feed one's Immunity-to-Change map into the AI and ask it to suggest possible big assumptions or experiments to test – essentially brainstorming with the AI's vast knowledge base.

AI tools can also help with knowledge and perspective for inner development. For example, if a leader is struggling with a particular issue – say, dealing with failure – an AI can instantly provide a summary of how different thought leaders (or psychological research) suggest handling failure, thus giving the person new perspectives to consider. Or an AI assistant integrated into workflow might detect stress or negative sentiment in a user's communications (through sentiment analysis) and proactively offer a mindfulness exercise or a prompt to reflect on what's bothering them. These are speculative ideas, but they are within the realm of current technology being explored in the well-being and HR tech space.

However, it's important to note the limitations and cautions of AI in inner development. An AI lacks genuine emotional understanding and cannot fully grasp the unique nuances of a person's life. The Rocky.ai site itself acknowledges that an AI coach should not be expected to handle deep existential questions or complex emotional issues the way a human can. There is also the privacy concern – sharing one's intimate thoughts with a digital platform means potentially exposing sensitive data, so trust in the tool's security is essential. AI coaches also run the risk of generic or surface-level interactions; they might miss subtle cues or overuse formulaic questions that don't resonate with the user. Therefore, the ideal approach is often a hybrid: human coaches leveraging AI to enhance their practice (e.g., using AI to track patterns in a client's journaling between sessions, or to provide additional prompts), and individuals using AI for routine reflection while still seeking human guidance for deeper work.

In high-performance environments, we can envision AI assistants for leadership development becoming part of the standard toolkit. For example, a leader could have an AI "leadership mentor" accessible via phone or computer that provides daily check-ins: *"Good morning – yesterday you planned to give more positive feedback to your team. Did you get a chance*

## *Individual Change in the Age of AI*

*to do that? How did it go?”* and then adapt its next suggestions based on the user’s input. It could also serve up curated content (articles, videos) relevant to whatever skill the leader is focusing on (say, active listening or strategic thinking), effectively personalizing their learning journey.

Used wisely, AI can function as a mirror and a guide – albeit a guide following a script of best practices – to keep individuals engaged in the continuous process of growth. It aligns well with the notion of creating a culture of continuous learning. When employees at all levels have access to a kind of personal coach (human, AI, or a combination), it reinforces the message that development is supported and expected. It also helps close the gap between training sessions or coaching workshops; development becomes a daily practice, woven into work via micro-conversations with one’s AI helper, rather than an isolated event.

## **Critical Mirror Decision Support**

LLMs are designed to please, not to challenge. Research from IMD Business School demonstrates that executives consulting AI become more optimistic and less accurate in their predictions than peers who discuss ideas with humans<sup>14</sup>. In the study nearly 300 executives were asked to predict Nvidia’s stock price. Those who consulted ChatGPT became significantly more optimistic, confident, and less accurate than peers who discussed the question with each other. The AI’s authoritative tone and detailed responses created false assurance, lacking the social regulation and healthy skepticism that naturally emerged in human discussions. The findings reveal that executives must actively guard against AI’s tendency to amplify rather than challenge their biases. In the age of AI, there is a dangerous new failure mode: leaders now sprint through tight “yes-cycles,” rapidly iterating into strategic dead-ends based on flawed assumptions amplified by agreeable AI.

---

<sup>14</sup>Parra-Moyano, José et al., “Executives Who Used Gen AI Made Worse Predictions,” *Harvard Business Review*, 2025.

The idea of the Critical Mirror is to transform your AI from an agreeable assistant into a rigorous strategic partner by applying four proven heuristics that force systematic examination of your decisions:

- *First-Principles Thinking* strips away industry dogma to expose fundamental truths about the problem you’re solving, ensuring your strategy aligns with core realities rather than conventional wisdom. Prompt the LLM to ask what are the absolute, fundamental truths about the problem you are trying to solve?
- *Inversion* maps every path to catastrophic failure before you commit resources, functioning as the ultimate pre-mortem that reveals vulnerabilities invisible in optimistic planning. Prompt the LLM to describe in detail the top 3 ways a project could fail catastrophically.
- *Second-Order Thinking* traces cascading consequences six months downstream across customers, operations, and finances, exposing hidden costs that turn “quick wins” into expensive mistakes. Prompt the system to apply second-order-thinking by asking for good and bad consequences if a plan works exactly as intended. What new problems, dependencies, or feedback loops might it create? How would those outcomes affect the system in 6 months, 2 years, and 10 years?
- *Steelman Arguments* construct the strongest possible case against your strategy, battle-testing your plan against intelligent opposition rather than weak objections. Prompt the system to assume the strongest contrary data and incumbent incentives, and then make the most compelling, intelligent case why this strategy fails.

In summary, the practical toolbox for individual change in the AI era includes introspective mapping tools (like ITC) to diagnose one’s inner resistance, coaching techniques that foster self-generated insight and accountability, and increasingly, AI-driven platforms to scale and sustain the reflective work. By combining these you can systematically work on yourself much as you work on the business by diagnosing issues, applying interventions, measuring progress, and adjusting as needed. But tools and techniques ultimately serve a larger purpose: achieving real-world changes in behavior and mindset. In the concluding section, I will summarize how all

these insights come together when moving from insight to action, and how individuals can cultivate the elusive self-transforming capabilities needed for long-term success.

## **Conclusion: Developing Self-Transforming Capabilities**

Insight, no matter how profound, is only the beginning. The true test of personal change is whether we translate new understanding into new habits of mind and behavior. In this concluding section, I consider how individuals (especially leaders and those in high-performance roles) can move from the insights gained through the aforementioned frameworks into sustained action – ultimately developing what Kegan would call “self-transforming capabilities.” These are the capacities that enable a person to not only adapt to one big change, but to continually learn, evolve, and generate new ways of being as their environment keeps changing. In the age of AI, such meta-capabilities are perhaps the most valuable of all, because the one certainty is ongoing change and complexity.

Self-transforming capability refers to the qualities of the Self-Transforming Mind (Kegan’s Stage 5) that can be cultivated as practices. One does not need to be a permanent resident of Stage 5 (if such a thing even exists consistently) to benefit from these qualities. We can think of it as building the muscles that Stage 5 thinkers use. Here are some practical strategies and mindsets to develop these capabilities:

- *Embrace Ongoing Learning and Unlearning:* In concrete terms, this means making a routine of stepping outside your comfort zone. For example, commit to learning at least one new skill or domain each year – especially in areas where AI is advancing. Simultaneously, practice “unlearning” obsolete assumptions. Periodically ask yourself and your team, “What have we believed to be true that might no longer be true?” A self-transforming individual is not overly attached to past expertise; they remain a lifelong student. Executives

## *Conclusion: Developing Self-Transforming Capabilities*

can demonstrate this by openly engaging in training (showing vulnerability as a learner), or by rotating through different roles to gain fresh perspectives. This habit signals to others and reinforces in one-self that identity is not fixed to a narrow competency – it's fluid and expanding.

- *Cultivate Reflection and Mindfulness:* Create regular pauses in your frenetic schedule for reflection. This could be through journaling 10 minutes a day, a weekly mindfulness meditation, or an end-of-week recap with your team on “lessons learned.” Reflection is the bridge from experience to insight and from insight to wisdom. By making reflection a habit, you continuously integrate new experiences and extract meaning from them (echoing Frankl’s principle of meaning-making). Mindfulness practices, in particular, strengthen the ability to observe one’s own thoughts and emotions without being controlled by them. This mirrors Kegan’s subject-object shift – learning to see your thoughts as objects you can work with. A mindful leader might notice in real-time, “I’m feeling defensive in this meeting; what assumption is being threatened?” That awareness creates choice in how to respond, enabling more measured and creative actions.
- *Seek Out Diverse Perspectives:* One hallmark of a self-transforming mindset is being able to hold multiple perspectives. To train this, deliberately expose yourself to viewpoints and expertise outside your usual circle. This can be done by reading widely (including authors with different opinions), fostering diverse teams, or finding a peer group/mentor network that challenges your thinking. In practice, when faced with a tough decision, consult a range of voices – perhaps an engineer, a designer, a customer, an ethicist – and truly listen to each. This doesn’t just provide information; it habituates you to mentally juggling different frameworks. Over time, you become more comfortable with complexity and paradox. For instance, you might learn to balance the short-term financial perspective with the long-term sustainability perspective, rather than choosing one and ignoring the other. In the AI context, this might mean considering optimistic and pessimistic views about AI’s impact simultaneously,

## *Individual Change in the Age of AI*

leading to more nuanced strategies (neither blindly tech-zealous nor fear-driven, but integrative).

- *Connect to Purpose and Values Continuously:* Reiterate and re-evaluate your “why” on an ongoing basis. Just as companies revisit their mission in changing times, individuals should revisit their personal mission. Frankl taught us that meaning is central; a self-transforming person refines their understanding of their purpose as they grow. Maybe early in your career your purpose was “to excel and build expertise in X domain,” but later it evolves to “to mentor others and create a lasting positive impact.” If you keep that purpose front and center, it will guide you through turbulent changes – acting as a north star when you have to reinvent how you fulfill it. Purpose acts as the stable core that allows flexibility around methods and roles.
- *Practice Adaptive Action:* Turn the cycle of insight -> experiment -> feedback into a way of life. In organizational leadership, this is akin to adopting an agile mindset, but for personal growth. When you realize something about yourself or your environment, act on it quickly in a small way, then observe results, learn, and iterate. For instance, after mapping your immunity to change, don’t shelf it – immediately try a different approach in your next meeting (even if it’s minor), then reflect that evening on what you noticed. This trains the belief that change is possible and that you are an agent in your own development. It also combats the paralysis that can come from analysis without action. By continuously cycling through action and reflection, you become more resilient – mistakes or failures are just data for the next iteration, rather than ego-shattering events. This is precisely the attitude needed in an AI-driven world where new tools and processes must be tried and refined.
- *Leverage “Self-Transforming” Communities:* Individual change is bolstered by supportive relationships. Seek out coaches, mentors, or peer groups who understand and encourage vertical development and personal mastery. In some companies, formal leadership development programs or action learning sets serve this function – giving leaders

## *Conclusion: Developing Self-Transforming Capabilities*

a forum to discuss their personal challenges (not just business tasks) and to hold each other accountable for growth. When an organization's culture encourages vulnerability, reflection, and mutual coaching, it creates a collective growth mindset. If you are a leader, you can role-model this by, for example, sharing with your team your own development goal and what you are doing to work on it. This can be powerful – it normalizes that even the boss is still growing and invites others to do the same. Over time, you cultivate an environment where people are not hiding their weaknesses or mistakes but collaboratively turning them into growth, which dramatically increases an organization's adaptive capacity.

Finally, it's worth reinforcing the mindset of self-transformation as an ongoing journey, not a destination. In a sense, "developing self-transforming capabilities" means realizing one will never be done developing. This humility and openness is critical in the face of AI and future disruptions. The moment a leader thinks "I have arrived" or "I know enough," they close themselves off to signals of change. In contrast, the self-transforming mindset is always somewhat skeptical of itself – not in a self-doubting way, but in a curious way: How might my current approach be incomplete? What can I learn next?

We can conclude that AI can help us reach a more integrative perspective: Individual change is not only about adopting new technologies or skills, but about evolving one's inner architecture to match a rapidly evolving outer world. By disarming our immunity to change, we remove internal roadblocks to growth. By finding and renewing meaning, we build the emotional resilience to face adversity. By expanding our mental complexity, we equip ourselves to handle greater ambiguity and systemic complexity.

By balancing our cognitive styles (left and right brain), we become more holistically intelligent and creative. By utilizing practical tools and even AI assistants, we support and sustain the hard work of changing habits and mindsets. And through all of this, by moving from one insight to the next action, and to the next insight, we step-by-step become self-authoring, then self-transforming leaders who can not only survive disruptive change – but harness it to create and contribute in ever more impactful ways.

## *Individual Change in the Age of AI*

The reward for those who undertake this journey is not just staying relevant in a competitive sense, but the emergence of a more capable, wise, and authentic self – one that can thrive in synergy with AI and lead others toward a future of continuous learning and purposeful innovation. The age of AI, then, rather than rendering humans obsolete, can become a catalyst for unprecedented inner development, closing the gap between human capacity and worldly complexity, and opening new horizons for what people and technology can achieve together.

Individual transformation, however, is necessary but insufficient. Even the most evolved leader – one who has expanded their consciousness, mastered their immunity to change, and cultivated whole-brain thinking – cannot single-handedly overcome systemic organizational barriers. A self-transforming mind operating within a rigid, fear-driven organization will find its potential constrained by structures that punish experimentation, cultures that suppress dissent, and collective mental models that cling to outdated success formulas. The next chapter examines how these collective dynamics create immunity to change at organizational scale, and how leaders can redesign formal structures, social fabric, and shared mindsets to build organizations capable of continuous adaptation.

## **Key Takeaways: Individual Change**

### **What You Can Do:**

- **Map your immunity to change:** Use an “Immunity to Change” map to uncover hidden assumptions sabotaging your goals. Identify competing commitments and test “big assumptions” through small, safe experiments. Don’t just push harder – investigate what’s holding you back.
- **Practice double-loop learning:** When facing challenges, question underlying assumptions, not just tactics. Ask: “What beliefs am I operating from? What if they’re wrong?” Shift from accumulating specific skills to mastering continuous learning itself.
- **Find meaning in disruption:** Anchor your identity in core values and purpose, not job titles or specific skills. When AI disrupts your role, reconnect with your deeper “why” – what you’re committed to beyond any particular position.
- **Level up your mental complexity:** Move from “Socialized Mind” (following others’ expectations) to “Self-Authoring Mind” (guided by internal principles) or “Self-Transforming Mind” (holding multiple perspectives). Seek stretching experiences, reflective practices, and coaching to develop this capacity.
- **Balance left and right brain thinking:** AI excels at left-brain analytical tasks. Consciously cultivate right-brain capacities: holistic thinking, context sensitivity, empathy, and meaning-making. Use metaphors, stories, and reflection to prevent narrow optimization.
- **Use AI as a developmental tool:** Leverage AI coaches for 24/7 reflection support, Socratic questioning, and assumption testing. AI democratizes access to developmental support previously available only through expensive executive coaching. See the “Practical Tools” section for specific prompts and protocols.

## *Individual Change in the Age of AI*

- **Build self-transforming capabilities:** Develop habits of ongoing learning, reflection, seeking diverse perspectives, and adaptive action. The goal isn't to "arrive" but to continuously evolve. Your capacity to lead change is constrained by your current level of consciousness – invest in expanding it.

# **Organizational Change in the Age of AI**

Organizations today face unprecedented pressure to adapt and change. Rapid advances in AI and automation potential drive much of this pressure. Yet despite AI's promise, many companies encounter organizational resistance when trying to implement AI driven systems or processes. This resistance runs deeper than simple reluctance. Just as individuals harbor hidden fears and competing commitments that undermine change efforts, organizations possess collective mindsets that unconsciously protect the status quo.

These collective hidden commitments may include worries about losing established success formulas, fear of letting go of familiar mental models, or clinging to "the way we've always done it." Such an organizational immune system – while meant to preserve stability – can powerfully thwart innovation and AI adoption. For example, leadership may publicly endorse an AI initiative while unwittingly maintaining practices that undermine it (e.g., continuing to reward old behaviors), reflecting unspoken commitments to the old paradigm.

The result is a puzzling gap between the organization's stated change goals and its actual behavior. In an AI-driven business environment, overcoming this gap is critical. This chapter explores how businesses can diagnose and transform the multiple layers of organizational resistance from formal structures to social dynamics to deeply-held mental models in order to thrive in the age of AI.

I begin by examining why organizational transformations fail and why organizations resist change. Often, it's not just "fear of new technology"

## *Organizational Change in the Age of AI*

in the abstract, but a multi-layered phenomenon. At the surface, there may be structural and process issues (for instance, rigid hierarchies or outdated metrics) that impede change. Beneath that lie cultural and social factors, like low trust or poor communication, which can turn a workforce against even well-intentioned AI projects. Deeper still, the mental models and biases of leaders and groups can create blind spots and immunity by sustaining old assumptions even when the world has moved on.

I will unpack each of these layers of resistance – rooted in the formal, social, and mental contexts – and present frameworks and tools to diagnose and overcome them. Throughout, insights from contemporary research and practice are integrated, including data-driven change management techniques and emerging AI-powered tools that can serve as levers for transformation. Real-world case examples (such as resistance to an AI-enabled system) and archetypal employee reactions (from *Silent Resistors* to *Active Opponents*) illustrate how these dynamics play out in practice and how targeted interventions can make a difference.

## **Why Organizational Transformations Fail**

John P. Kotter's influential research on change management stems from a fundamental observation: most organizational transformations fail not because of flawed strategy, but because of poor execution and weak leadership throughout the change process<sup>1</sup>. His widely-adopted Eight-Step Process for Leading Change offers a structured roadmap that addresses the human and emotional dimensions of transformation. The first three steps – establishing a sense of urgency, creating a guiding coalition, and developing a vision and strategy – focus on preparing the organization for change. People must feel a genuine need to act before transformation can begin, and a powerful coalition of credible, committed leaders is essential to drive momentum. The next steps – communicating the vision, empowering broad-based action, and generating short-term wins – translate strategy into engagement, ensuring employees understand, believe in, and contribute to the vision.

---

<sup>1</sup>Kotter, John P., *Leading Change* (Harvard Business School Press, 1996).

## *Lewin's and Schein's Foundational Frameworks*

Finally, consolidating gains and anchoring new approaches in the culture embed change into the organization's identity, preventing regression to old habits.

Kotter's central insight is that change is primarily a leadership challenge, not a management one. Successful transformations require emotional as well as rational engagement by appealing to people's hearts as much as their minds. Leaders must create clarity, inspire trust, and remove barriers that block progress. He also highlighted the importance of momentum: visible early wins validate the effort, build confidence, and silence critics. For Kotter, lasting change occurs only when new behaviors and values become part of the organization's culture – when they are “the way we do things around here.” His framework bridges strategic logic with human psychology, positioning change management as an ongoing, participatory process that demands vision, alignment, and sustained leadership energy.

In the age of AI, Kotter's insights remain especially relevant. AI-driven changes often trigger deep anxieties about job security and shifting roles, making emotional engagement even more critical. The speed of technological change also makes momentum management crucial – organizations must generate visible wins quickly to maintain commitment. Furthermore, AI projects often require cross-functional collaboration, making Kotter's emphasis on building guiding coalitions essential. As we will see in the following sections, resistance to AI adoption manifests across three interconnected layers – formal structures, social dynamics, and mental models – each requiring the kind of holistic leadership approach Kotter advocates.

## **Lewin's and Schein's Foundational Frameworks**

Kurt Lewin's foundational insights on change management provide essential structure for understanding how individuals and groups adapt to new ways of thinking and behaving<sup>2</sup>. His **three-stage model of change** – un-

---

<sup>2</sup>Lewin, Kurt, “Group Decision and Social Change,” in *Readings in Social Psychology*, ed. T. M. Newcomb and E. L. Hartley (Holt, Rinehart; Winston, 1947).

## *Organizational Change in the Age of AI*

**freezing, changing, and refreezing** – describes a process that remains relevant today, especially when applied to AI-driven transformations.

In the **unfreezing** stage, the goal is to disrupt the existing equilibrium by confronting people with evidence that their current behaviors or beliefs are no longer effective. This creates readiness for change by increasing awareness of dissatisfaction while reducing resistance. In an AI context, unfreezing might involve demonstrating how legacy processes or tools are becoming obsolete, or showing concrete data that competitors are gaining advantages through AI adoption. The challenge is that many organizations resist this initial discomfort, preferring to maintain current state comfort over future competitiveness.

The **changing** (or moving) stage follows, during which individuals experiment with new behaviors, attitudes, or processes. Learning and role modeling are critical at this point, as people replace old habits with new ones. For AI transformations, this often means pilot programs where employees can safely experiment with AI tools, training programs that build competence without fear of failure, and visible leadership demonstrating their own learning journey with AI. This stage requires psychological safety – people must feel they can try, fail, and iterate without penalty.

Finally, in the **refreezing** stage, the new behaviors are stabilized and reinforced through supportive structures, norms, and cultural alignment, ensuring that the change becomes the new status quo rather than a temporary adjustment. In AI adoption, refreezing means embedding AI tools into standard workflows, updating job descriptions to include AI competencies, rewarding innovation with AI, and celebrating successes until AI becomes “simply how we work.” Without this stage, organizations risk reverting to old ways when initial enthusiasm wanes.

Edgar Schein’s work extends Lewin’s framework by focusing specifically on **organizational culture** and the psychological dynamics that enable deep learning<sup>3</sup>. While Lewin provided the basic structure, Schein added crucial insights about what makes the unfreezing-changing-refreezing process actually work in practice.

---

<sup>3</sup>Schein, Edgar H., *Organizational Culture and Leadership* (Jossey-Bass, 1985).

Schein deepened our understanding of **unfreezing** by emphasizing that people need **psychological safety** to question existing beliefs without fear of punishment or loss of identity. For unfreezing to occur, individuals must be able to acknowledge that their current behaviors, attitudes, or assumptions are no longer effective. This requires creating an environment safe enough to admit inadequacy, where presenting “disconfirming data” – evidence that challenges the status quo – is welcome rather than threatening. In AI transformations, this means leaders must explicitly invite questioning of current processes, acknowledge uncertainty, and signal that it’s safe to say “I don’t understand this new AI tool” or “I’m worried about how this will affect my role.” Without this safety, employees will superficially comply with AI adoption while privately resisting or sabotaging it.

The **changing** phase in Schein’s view is not just about learning new skills but about **transforming underlying assumptions**. For AI adoption, this means going beyond behavioral training (“here’s how to use the tool”) to cultural learning: questioning assumptions about what work is, who does it, and what value humans bring when AI can handle routine cognitive tasks. This deep learning happens through guided experimentation, role modeling by leaders, and feedback loops that help people internalize new ways of thinking. Schein stressed that sustainable change depends on this internalization – not just going through the motions but actually believing that the new approach is better and aligning one’s identity with it.

The **refreezing** in Schein’s model requires that new learning becomes embedded in the **shared meanings and assumptions** of the organizational culture. For AI transformation, this means that using AI isn’t just a new behavior but part of “who we are” as an organization – it reflects values like innovation, continuous learning, and human-AI collaboration. Schein viewed leaders as **culture shapers** who must create conditions that balance psychological safety with sufficient disconfirmation to drive genuine learning. Leaders must model vulnerability by sharing their own AI learning journey, celebrate AI-enhanced successes throughout the organization, and ensure systems (performance reviews, promotions, team norms) reinforce the new cultural assumptions rather than inadvertently undermining them.

## *Organizational Change in the Age of AI*

Critically, Schein recognized that in today's fast-evolving environment, organizations may not want to fully "refreeze" but rather maintain a state of **adaptive unfreezing** – keeping culture fluid enough to evolve continuously. For AI transformation, where the technology itself is rapidly developing, this adaptive capacity becomes essential. Organizations must refreeze enough to stabilize initial AI adoption, but not so rigidly that they can't adapt when the next generation of AI capabilities emerges. This balance between stability and agility is one of the key leadership challenges in the AI era.

Beyond this model, Lewin emphasized that successful change depends on understanding the **dynamics of social systems and group behavior**. His **force field analysis** concept illustrated that any situation is maintained by a balance between **driving forces** that promote change and **restraining forces** that resist it. For lasting transformation, organizations must strengthen driving forces (e.g., executive sponsorship, clear benefits, competitive pressure, employee excitement about new capabilities) or reduce restraining ones (e.g., fear of job loss, lack of skills, technical limitations, cultural inertia). A skilled change leader diagnoses these forces and works strategically on both sides.

Lewin also believed that participation and dialogue are essential: people are more committed to change when they are involved in diagnosing problems and designing solutions. This participatory approach is particularly important for AI initiatives, which can trigger fears about automation and displacement. When employees help shape how AI is implemented – deciding which tasks should be augmented, testing tools, designing workflows – they develop ownership rather than resistance. His work established the principle that effective change is both a **scientific** and **human process** – guided by behavioral data but grounded in empathy, participation, and respect for group norms.

In the sections that follow, we will explore how Lewin's framework applies to AI-driven change, where the unfreezing may need to address deep-seated fears about AI, the changing phase must navigate steep learning curves, and the refreezing must occur in a rapidly evolving technological landscape where today's "new normal" may need to unfreeze again tomorrow. This

continuous cycle of adaptation – unfreezing, changing, refreezing – is itself a new organizational capability that AI-era companies must master.

## **Layers of Organizational Resistance in an AI-Driven World**

Organizational resistance is rarely monolithic; it operates on multiple levels. We can think in terms of three interrelated contexts: formal, social, and mental context. The formal context includes the tangible structures, processes, and goals by which an organization runs. The social context involves relationships, communication patterns, the unwritten “rules of the game” and trust. The mental context refers to the collective mindset: the paradigms, beliefs, and biases held by the organization’s members. Change initiatives can be thwarted by barriers in any of these interconnected contexts or all three. Below, I examine each context, the typical barriers encountered, and approaches to diagnosing and transforming those barriers.

### **Formal Barriers**

The formal context of an organization encompasses foremost its organizational structure (hierarchies or networks), governance processes, policies, performance metrics, and incentive systems. Common formal barriers to change include rigid hierarchies, siloed departments, legacy processes, and an overemphasis on narrow performance targets or short-term Key Performance Indicators (KPIs). Many organizations still operate with a command-and-control management style where decisions are pushed down a hierarchy, strict functional divisions, and success measured by hitting predefined targets. While such structures can bring order, they can also create inertia.

One major formal barrier is the misuse of targets and KPIs. Metrics are important, but when misapplied they can inadvertently encourage people

## *Organizational Change in the Age of AI*

to resist real change in favor of “looking good” on paper. As systems thinker John Seddon succinctly observed in the public sector, “Most targets do not represent the reality of a service from the customer’s point of view. Targets drive people to use their ingenuity to meet the target, not improve performance.” In other words, teams learn to game the metrics by meeting the letter of a goal without actually improving outcomes. For example, a call center might quickly pass along difficult cases to other departments to meet a call-duration target, leading to good stats for each silo but poor end-to-end service. This “you get what you measure” problem is symptomatic of command-and-control thinking.

Traditional formal systems often struggle to accommodate the agility and experimentation that AI initiatives require. New AI tools or ways of working also tend not to fit neatly into existing roles and processes, and proposals to change those structures often meet internal friction. When an organization announces an AI-driven change (say, deploying an AI scheduling tool or analytics system) but still enforces old KPIs and budgets, employees face a double bind: experiment and risk missing targets, or stick to the old process and quietly undermine the change. Moreover, traditional planning and forecasting methods may assume a stable environment, leading to complexity mismatches where the organization’s formal plans cannot keep up with the fast-evolving, uncertain nature of AI innovations.

### **Diagnosing formal barriers**

Organization leaders should always start a change initiative by reviewing whether their structures and metrics align with the change vision. Are there structural silos or bureaucratic steps slowing down change projects? Do current incentive systems inadvertently penalize experimentation? Data can provide early warnings. Key HR and performance metrics often signal formal strain during a transformation. For instance, employee turnover or absenteeism spikes in a department might indicate that a new AI system has disrupted workflows without sufficient support. Drops in productivity or quality measures can reveal that employees are struggling to adapt to new processes. Even training participation rates are telling:

if few employees complete optional AI training modules, it suggests low buy-in or that workloads leave no slack for learning. By monitoring such indicators, organizations can pinpoint where formal aspects (like workload allocation, training, or job design) are impeding change. Tools like process mapping and value-stream analysis can also uncover bottlenecks introduced by legacy procedures.

In the AI era, some companies are adopting more agile and networked structures – for example, forming cross-functional teams focused on AI initiatives, or creating “communities of practice” around new technologies – to break down silos and increase adaptability. Thus, the formal context can be transformed by redesigning roles and processes to be more flexible. That might include updating KPIs to focus on outcomes and learning (not just output), implementing “beyond budgeting” approaches (adaptive targets rather than fixed ones), and decentralizing decision-making so that teams can iterate quickly with AI tools. The goal is a formal context that enables change: structures that accelerate information flow and experimentation, and metrics that drive the desired behaviors (like collaboration, innovation, customer-centricity) instead of inadvertently encouraging resistance.

## Social Barriers

Even if formal elements are tuned for change, the social context – how people interact, feel, and relate – can make or break any change initiative. Key social factors include the level of trust in leadership, the openness of communication, norms around risk-taking, and the sense of fairness or reciprocity between employer and employees.

A critical concept here is the psychological contract which refers to the unwritten, implicit set of mutual expectations and obligations that exist between an employee and their employer, operating alongside the formal employment contract. First conceptualized by organizational psychologist Denise Rousseau,<sup>4</sup> the psychological contract encompasses beliefs about re-

---

<sup>4</sup>Rousseau, Denise M., *Psychological Contracts in Organizations: Understanding Written and Unwritten Agreements* (SAGE Publications, 1995).

## *Organizational Change in the Age of AI*

ciprocal promises and commitments – what employees believe they owe the organization (such as loyalty, effort, and flexibility) and what they believe the organization owes them in return (such as job security, fair treatment, career development, and meaningful work). Unlike legal contracts, psychological contracts are subjective, perceptual, and often unarticulated, yet they profoundly influence employee attitudes, motivation, and behavior. When organizations fulfill or exceed these implicit promises, employees tend to demonstrate higher engagement, commitment, and organizational citizenship behaviors. Conversely, when the psychological contract is violated – whether through broken promises about promotion opportunities, changes in job responsibilities without consultation, or failures to provide expected support – employees often experience feelings of betrayal, leading to decreased trust, reduced performance, increased turnover intentions, and cynical attitudes toward the organization. The concept has become increasingly important in contemporary workplaces characterized by restructuring, downsizing, and the erosion of traditional employment relationships, as it highlights how perceived fairness and reciprocity fundamentally shape the employment relationship.

When a major change like AI adoption comes along, employees subconsciously evaluate it against this psychological contract. Will this change violate the promises (implicit or explicit) that have been made to us? For example, if employees believe “our company values our job security and growth,” introducing AI that might automate tasks can be seen as a breach of trust if handled poorly. Lack of trust and poor social cohesion are fertile ground for resistance.

Communication is closely tied to trust. In many failed change efforts, employees complain that they were “left in the dark” or that concerns were never heard. Poor communication – whether a lack of transparency about a change project, one-way top-down announcements, or failure to address employees’ fears – creates uncertainty and rumor. In the vacuum, worst-case scenarios (e.g., “Will AI take our jobs?”) flourish. Conversely, organizations that succeed in transformation often foster a culture of open dialogue: leaders actively listen and respond to feedback, and communicate candidly about the purpose of the change, its benefits and limitations, and how roles will evolve. Building trust also involves demonstrating fairness

(for instance, if AI-driven efficiencies occur, are employees sharing in the benefits or being unfairly displaced?). Trust and engagement are further influenced by how inclusive the change process is. If AI is something “done to” employees without their input, resistance is natural. If people are invited to co-create solutions or pilot new tools, they are more likely to develop ownership.

Another social dimension of resistance is variation in how different groups of employees experience change. Within the same organization, you often find disparate engagement levels: some are enthusiastic adopters, others hesitant, and some openly opposed. In change management, it’s useful to identify engagement archetypes or personas to tailor strategies accordingly. For example, Gallup’s research on engagement<sup>5</sup> often categorizes employees as engaged, not engaged, or actively disengaged. In the context of an AI rollout, we can be more specific. Are there quiet fence-sitters who comply outwardly but haven’t emotionally bought in? Are there veterans who voice skepticism because they’ve “seen fads come and go”? Recognizing these patterns is crucial because a one-size-fits-all change program will likely fail to address the needs and concerns of different groups. Indeed, social context interventions should be contingent – targeted to the audience – since “one size almost never fits all.”

### **Diagnosing Social Barriers**

Organizations can gauge their social context through a mix of qualitative and data-driven methods. Employee surveys (especially anonymous ones) remain a staple for measuring trust, morale, and readiness – a drop in engagement or favorability scores towards the change can flag brewing resistance. In the age of AI, new tools like sentiment analysis allow leaders to parse open-ended feedback or internal social media posts at scale, detecting negative sentiment or fear in employees’ own words. Monitoring communication channels can reveal if misinformation is spreading or if concerns are not being addressed. Another powerful approach is organizational network

---

<sup>5</sup>Harter, Jim, *Employee Engagement: The Essential Guide for Business Leaders* (Gallup Press, 2017).

## *Organizational Change in the Age of AI*

analysis – essentially applying *Social Physics* principles to understand how information and influence flow in the organization.<sup>6</sup>

An interesting example of this approach is the mapping of the social networks within the College of Cardinals prior to the Conclave using Vatican records to identify structural advantages in papal elections based on status, mediation power, and coalition-building potential.<sup>7</sup> This network mapping is shown in the following figure.<sup>8</sup> The analysis revealed that Cardinal Robert F. Prevost occupied an exceptionally central position across multiple network dimensions, making him a strong contender despite being overlooked by conventional forecasters.

By analyzing email, chat, or meeting data (ethically and in aggregate), one can identify key influencers or “energy hubs” in the informal network. If those influencers are cynical about a change initiative, their skepticism can quietly permeate their peers. Conversely, enlisting influential employees who are positive can sway networks. For instance, a network analysis might find “boundary spanners” – employees who connect otherwise separate groups (say, linking a tech team with an operations team). These individuals are often critical for cross-department changes. Leaders should also watch for signs of a broken psychological contract: spikes in turnover of valued staff, increased references to “fairness” in comments, or declining participation in discretionary activities could all indicate that employees feel the social “deal” has changed for the worse. By diagnosing these social signals early, interventions can be made to rebuild trust – for example, town hall meetings to address concerns, setting up two-way communication forums, or visibly acting on feedback (showing employees that “*your voice matters*” in the change). Being honest about uncertainties and acknowledging fears (“We know there’s concern about job impact – here’s how we’re addressing that...”) can defuse anxiety. It’s also vital to high-

---

<sup>6</sup>Pentland, Alex, *Social Physics: How Good Ideas Spread-the Lessons from a New Science* (Penguin Press, 2014).

<sup>7</sup>Soda, Giuseppe et al., “In the Network of the Conclave: Social Connections and the Making of a Pope,” *Social Networks* 83 (2025): 215–32, <https://doi.org/10.1016/j.socnet.2025.07.003>.

<sup>8</sup>Orlando, Barbara, and Tomaso Eridani, *In the Network of the Conclave*, Bocconi University, 2025, <https://www.unibocconi.it/en/news/network-conclave>.

## *Layers of Organizational Resistance in an AI-Driven World*

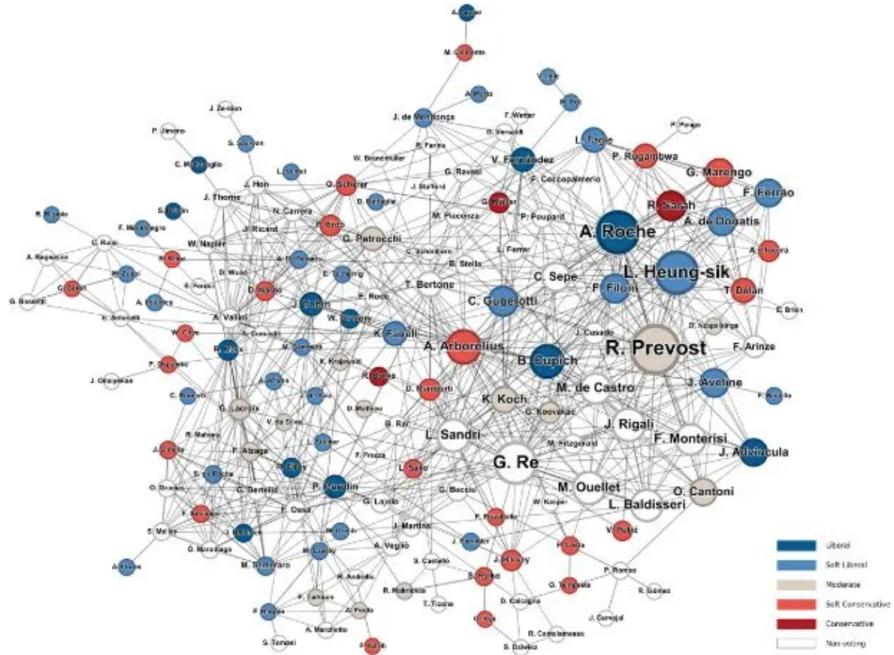


Figure 28: Mapping status among the cardinals prior to the conclave

light success stories of the change and quick wins to show progress, which can convert fence-sitters by demonstrating that the AI can make work better. Another key strategy is employee participation – engaging staff in the change process. This could mean inviting volunteers to be part of pilot programs or “AI champions” who test new tools and give feedback, or forming cross-level committees to guide implementation policies (for instance, on AI ethics or data use). Such inclusion helps employees feel agency rather than being passive recipients of change.

Tailoring engagement strategies to different archetypes of employees is especially effective. For instance, those we might call “Silent Resistors” – people who aren’t openly complaining but quietly avoid using a new AI system – may need different outreach than “Active Opponents” who publicly criticize it. Let’s consider a few common archetypes in organizational change and how to handle them (I will explore more through a case example in the next section):

## *Organizational Change in the Age of AI*

- *Silent Resisters:* These individuals comply outwardly with the change (no open objections) but exhibit low engagement – e.g., minimal usage of a new AI tool and scarce feedback. They may be withholding their true concerns or doubting the change but not voicing it.  
*Engagement strategy:* Create safe spaces for input. Use gentle peer pressure and social proof – for example, showcasing colleagues' positive experiences to nudge them. Small group workshops or hands-on labs can encourage silent types to participate (silence is harder to maintain in an interactive session). Also provide anonymous Q&A channels (anonymous forums or suggestion boxes) to surface hidden worries. The goal is to turn quiet disengagement into open dialogue.
- *Skeptical Veterans:* Often long-tenured, highly experienced employees who voice pointed criticisms of the change (“We tried something like this 10 years ago, it didn’t work...”). Their skepticism might come with a thoughtful, detailed perspective (they care about the work and have a reputation).  
*Engagement strategy:* Rather than sidelining these veterans, involve them. For example, invite them into the pilot testing of the AI or to help refine the implementation with their domain knowledge. Show them data or external benchmarks to address their specific critiques (they often respond to evidence). Crucially, give them a role in guiding others – for instance, as mentors or SME (subject-matter expert) advisors in the rollout. By valuing their expertise, you convert skeptics into owners of the change, which can turn their skepticism into constructive input and eventually ownership of the solution.
- *Anxious Learners:* These employees are eager to adapt but filled with self-doubt about their ability. You’ll see them signing up for every training and frequently asking for help, worried about “doing it wrong.”  
*Engagement strategy:* Provide extra support and reassurance. This could mean personalized coaching or a buddy system where a tech-savvy colleague mentors them. Emphasize a culture of psychological safety – explicitly tell them it’s okay to make mistakes during learning. Offer additional learning aids (how-to videos, sandbox environments to practice with the AI tool without consequences) to build

their confidence. The aim is to transform anxiety into competence through support.

- *Overloaded Pragmatists:* These are solid performers already stretched thin with work. They may not object to the change in principle, but they simply don't have bandwidth to engage with it – leading to passive resistance (e.g., delays in adopting the new system). They might experience change as just “*one more thing on my plate.*”

*Engagement strategy:* Acknowledge workload realities and adjust demands. Perhaps postpone other initiatives or provide temporary help to reduce their load during the transition. Emphasize staged adoption – break the change into manageable phases so it's not overwhelming. Also, recognize and reward small wins in using the new system; this maintains motivation and shows that leadership appreciates their effort amid a heavy workload.

- *Active Opponents:* A minority may actively push back – voicing strong criticism on internal forums or in meetings, possibly rallying others against the change. They might be influential personalities or informal leaders who feel the change is fundamentally flawed.

*Engagement strategy:* First, listen and empathize through one-on-one conversations. Often, active opponents want to feel heard. Try to find common ground or solicit their input to improve the implementation (e.g., “Help us stress-test this plan to identify weaknesses”). By giving them a constructive role, some opponents can become invaluable critics-turned-advisors. However, if genuine toxicity or bad faith persists, leadership must sometimes set boundaries – limiting their negative influence channels. For example, ensuring that valid concerns are addressed, but disallowing harmful behaviors (like spreading misinformation or disrespect). In extreme cases, reassigning or isolating an incorrigible opponent may be necessary, but this is a last resort after attempting engagement.

- *Conditional Supporters:* These employees are on the fence, neither advocating nor resisting strongly. They often take a “wait and see” attitude – supportive only if they see proof that the change is beneficial.

## *Organizational Change in the Age of AI*

*Engagement strategy:* Provide the proof they seek. Quickly roll out early wins and visible benefits of the AI (even small improvements) and broadcast them. Use transparent metrics to show the change is working (for example, “Since AI was introduced, customer satisfaction is up 10% – here’s the data”). Involve conditional supporters in feedback loops or user groups so they feel their voices matter in tweaking the change. As they see tangible positives and feel included, their cautious neutrality can shift to genuine buy-in.

These archetypes illustrate that effective change management in the AI age is not just top-down – it’s person-specific. By addressing social needs (like trust, recognition, inclusion) and meeting people where they are, leaders can turn potential resistance into collaboration. A healthy social context is one where employees feel respected during change: their concerns are heard, their efforts are supported, and successes are shared. AI implementations, in particular, require trust (for employees to embrace new tech) and open communication (to clarify misunderstandings and reduce fear). Thus, strengthening the psychological contract – through fairness, involvement, and consistent, honest communication – is a direct antidote to social resistance.

## **Mental Barriers**

The deepest layer of organizational resistance lies in the mental models and mindsets held collectively by the organization – especially its leadership and influential members. This includes beliefs about how the world works, what has made the company successful, and even how decisions should be made. In times of disruptive change like AI, certain ingrained paradigms can become liabilities. For instance, a company might have a deeply held belief that *“our industry is about personal relationships, not algorithms”*. Such a mindset, if unexamined, will cause leaders to unconsciously reject or minimize AI solutions, even as competitors leverage them. Similarly, executives proud of a long track record might cling to the mental model that *“what brought success before will do so again,”* making them dismiss signals that a new approach is needed. These are examples of hidden

competing commitments at the organizational level – analogous to how an individual might consciously want to change but subconsciously commit to an opposite behavior due to an ingrained assumption. At an organizational scale, these hidden commitments manifest as collective inertia or subtle sabotage of change (e.g., consistently deprioritizing an AI project in favor of familiar initiatives).

One useful perspective on mental complexity comes from developmental psychology as introduced in the chapter on individual change. We can loosely apply Robert Kegan's stages of mental development like the Socialized Mind, Self-Authoring Mind, and Self-Transforming Mind to organizations.

- A “Socialized” organization is one that largely follows established norms and industry orthodoxies – it is responsive to external expectations but may lack an internal compass for change.
- A “Self-Authoring” organization has its own distinct strategy and values – it can set a course and innovate within a given paradigm, but might still be limited by its single overarching worldview or culture.
- A “Self-Transforming” organization is more rare; it embodies a learning-oriented culture that can challenge its own assumptions and continuously evolve. This highest stage parallels what Peter Senge called the learning organization, which can adapt and transform by reflecting on its behavior and environment. In the context of AI, a self-transforming organization might, for example, readily pilot radical new AI-driven business models and, if evidence shows a better approach, willingly pivot – rather than defending the legacy model at all costs. Many firms aspire to this adaptiveness but find their internal mindsets holding them back.

Several cognitive biases and decision-making traps commonly afflict top management teams facing transformative change. These biases act as mental barriers to seeing reality and acting effectively. For instance, overconfidence bias can lead executives to underestimate the complexity of implementing AI (“We've never delivered a project like this on time, but this time we surely will!”). Such misplaced optimism can cause leaders

## *Organizational Change in the Age of AI*

to ignore warning signs or skip necessary investments in change management, assuming success is certain. Confirmation bias might cause decision-makers to selectively hear feedback that the AI project is going well, while dismissing negative data as anomalies. Self-serving bias could have leaders attribute any difficulties to employee failings (“the team just isn’t tech-savvy enough”) rather than their own strategy, hindering honest appraisal. Groupthink and what has been termed the “alpha syndrome” can also impede adaptive decisions. In many executive teams, a few dominant voices (alphas) prevail; colleagues may hesitate to challenge them, creating an illusion of consensus. This can suppress diverse perspectives – exactly what’s needed when navigating uncharted territory like AI. As a result, leadership might cling to a simplistic narrative (“Our product is great; if customers don’t get it, that’s their fault!”) instead of confronting hard truths. These biases and hidden assumptions collectively limit an organization’s mental agility.

The mental barriers discussed above – whether clinging to past success formulas or falling prey to cognitive biases – point to a deeper issue: the quality of leadership during change. Warren Bennis’s work illuminates why some organizations navigate transformational change successfully while others fail<sup>9</sup>. He viewed change not as a mechanical process but as a profoundly human one that depends on the ability of leaders to inspire trust, create meaning, and align people around shared purpose.

Bennis argued that successful change arises from **transformational leadership** – leaders who articulate a compelling vision, model integrity, and foster an environment where people feel empowered to take risks and innovate. Rather than relying on authority or control, such leaders act as **social architects**, shaping cultures that value openness, collaboration, and learning. He emphasized that leadership during change requires **emotional intelligence** – particularly self-awareness, empathy, and resilience – to navigate uncertainty and maintain cohesion through disruption. For AI-driven transformations, this takes on special significance: leaders must not only understand the technology but also embody the cultural shift it requires – curiosity over certainty, experimentation over perfectionism,

---

<sup>9</sup>Bennis, Warren, *On Becoming a Leader* (Addison-Wesley, 1989).

learning over knowing.

A critical distinction Bennis highlighted is the difference between **managing** and **leading** change. Managers focus on maintaining systems and processes, whereas leaders guide organizations through ambiguity and renewal. During AI adoption, managers might optimize the rollout schedule and training logistics – essential work, but insufficient on its own. Leaders, by contrast, shape the narrative around AI (Is it a threat to jobs or an enabler of human potential?), model vulnerability by learning alongside employees, and make sense of ambiguity in ways that reduce fear and build confidence. This leadership is especially needed when AI's implications are still emerging – when there's no playbook to follow.

Bennis also warned against bureaucratic rigidity, advocating instead for **learning organizations** where communication is open, feedback is encouraged, and mistakes are treated as opportunities for growth. In his view, modern organizations face constant change, and only **adaptive, values-driven leadership** can ensure survival. The **adaptive capacity** – the collective ability to learn and evolve – is the hallmark of a healthy organization. This capacity becomes essential in the AI era, where what works today may not work tomorrow, and the organization must continuously evolve its understanding of how AI should enhance rather than replace human judgment.

Ultimately, Bennis's central insight was that effective change management depends less on strategy and structure than on **character and vision** – the capacity of leaders to mobilize human potential toward a shared, evolving future. For AI transformations, this means leaders who can:

- *See beyond efficiency:* Imagine how AI could fundamentally improve how work gets done, not just speed it up
- *Build psychological safety:* Create environments where employees can voice fears, make mistakes, and learn openly
- *Articulate purpose:* Connect AI adoption to meaningful goals – improving outcomes for customers, enabling employees to focus on higher-value work, advancing the organization's mission

## *Organizational Change in the Age of AI*

- *Model the change:* Demonstrate their own learning journey with AI, showing humility and curiosity rather than positioning themselves as experts

This leadership character is what allows organizations to transcend the mental barriers – the old paradigms, cognitive biases, and fear-driven assumptions – that otherwise inhibit change. Without it, even the most sophisticated change management frameworks and AI tools will falter. With it, organizations can tap into the collective intelligence and creativity needed to harness AI’s potential while preserving their humanity.

### **Diagnosing Mental Barriers**

Uncovering deep-seated mindsets is challenging because they are often invisible to those who hold them. However, there are telltale signs. Listening to the language leaders use in meetings or internal memos can be revealing. Phrases like “that’s how we do things here” or recurring stories of past glory may indicate attachment to old mental models. Dismissing new information with “not our problem – the customer just doesn’t get it” is a classic symptom of cognitive bias limiting learning.

Some organizations use facilitated workshops or coaching to surface these assumptions. The “Immunity to Change” mapping (Kegan & Lahey) can be done not only at an individual level but also in teams: leaders publicly state a change goal and then, through guided reflection, identify the unspoken beliefs and fears that conflict with that goal. This process can reveal, for example, that while a leadership team says “we want a culture of innovation,” they unconsciously fear losing control or appear incompetent if they personally don’t understand a new technology – thus they avoid truly empowering the tech teams. By making such competing commitments explicit, the team can begin to address them.

Another diagnostic tool is 360-degree feedback and cultural assessments: data from employee surveys can show if staff perceive leadership behaviors as authoritarian, risk-averse, or open-minded. If, say, a majority of employees feel “mistakes are held against you” in the company, that points to a

mental model of perfectionism or blame that will stifle the experimentation AI innovation needs. External advisors or board members can also provide a mirror to senior management's thinking, since insiders often share the same blind spots. In recent years, companies have even started leveraging AI itself to analyze decision-making patterns – for example, using algorithms to detect bias in how resources are allocated or which proposals get approved – offering an objective check on human judgment.

Changing an organization's mindset is arguably the hardest layer of change, but it's the most critical for lasting transformation. It requires leadership courage and humility – a willingness to question one's own worldview and to embrace complexity. One approach is to educate and expose leaders to new perspectives. This could involve formal training in topics like cognitive biases and systems thinking (so leaders learn, for instance, about action bias – the tendency to take quick action even when reflection is needed – and become more self-aware when they might be falling prey to it). Storytelling and case studies of other organizations that successfully navigated AI-driven change can also widen mental models (for example, hearing how a traditional company transformed might make it “thinkable” to shed some old practices). Additionally, bringing diverse voices into decision-making – whether it's more junior staff, external experts, or even AI advisors – can challenge groupthink. Some companies deliberately rotate executives into unfamiliar roles or send them to visit tech startups, to shake up ingrained thinking patterns.

Leaders and teams should practice double-loop learning, a concept where instead of just solving problems within the current frame (“How do we implement this AI within our existing structure?”), they also question the frame itself (“In a world of AI, what structure might we need if we were starting fresh?”). Techniques like scenario planning encourage this by asking teams to imagine future environments that break current assumptions. For example, leadership might explore a scenario where an AI could make many managerial decisions – what then is the role of managers? Such exercises can reveal attachment to identity and roles that need to evolve. Overcoming the “alpha syndrome” may involve setting new norms at the top: rewarding leaders not just for strong execution but for fostering reflection and learning in their teams. Some firms institute a practice of

## *Organizational Change in the Age of AI*

pre-mortems (envisioning a project's future failure and brainstorming potential reasons) to legitimize dissenting views and surface doubts early, thus countering overconfidence and confirmation bias.

Finally, pursuing a higher stage of development often means embracing a new organizational paradigm. Here, Frederic Laloux's model of organizational evolution provides a roadmap for mental context shifts.<sup>10</sup> Laloux describes developmental stages with color metaphors – Red, Amber, Orange, Green, and Teal – each with its own mindset and way of coordinating people:

- *Red organizations* are power-driven (think of a mafia or very primitive setups – fear-based control).
- *Amber organizations* are traditional bureaucracies with strict hierarchies and rules (stability and roles are paramount).
- *Orange organizations* are the classic modern corporations – achievement-oriented, competitive, using innovation and meritocracy to excel (machine or pyramid metaphor).
- *Green organizations* are more pluralistic – focused on empowerment, community, and values, often with servant leadership and culture-driven approaches.
- Finally, *Teal organizations* represent the evolutionary, self-organizing paradigm – characterized by self-management, wholeness (bringing one's full self to work), and an evolutionary purpose guiding the organization.

Laloux suggests that Teal (and to some extent Green) organizations are better equipped to handle complexity and change, because they encourage decentralized decision-making, adaptability, and continuous learning. For an organization to shift mentally from an Orange paradigm (where AI might be seen purely as a tool for efficiency and profit) to a Teal paradigm (where AI is also embraced as a way to further a purpose and empower people), it must let go of some ego, control, and fear. This could mean

---

<sup>10</sup>Laloux, Frederic, *Reinventing Organizations: A Guide to Creating Organizations Inspired by the Next Stage of Human Consciousness* (Nelson Parker, 2014).

redefining success beyond quarterly numbers (which is scary to many executives, but necessary to truly innovate), and trusting teams with more autonomy (even if it challenges managerial identity). Achieving a Teal or self-transforming stage is not trivial – it can take years and usually requires consistent role-modeling from the top. In fact, a number of the case studies of Teal organizations presented by Laloux have reversed course after a few years. Thus, in combination with a small sample the empirical evidence for Laloux’s claims is rather weak. However, even moving a few steps in that idealistic direction (e.g. by experimenting with self-managed teams on an AI project, or instituting practices of employee mindfulness and dialogue to enhance reflection on the desired change) may still be a good heuristic to improve the organization’s change resilience.

In summary, transforming the mental layer is about cultivating new mindsets: from seeing AI as a threat to be resisted, to viewing it as a learning opportunity; from clinging to past formulas, to iterating new approaches; and from controlling information, to sharing power. Leaders must lead by example in this mental shift, showing vulnerability (admitting “we don’t have all the answers about AI, but we’re here to learn together”) and curiosity. When an organization’s dominant mindset becomes one of curiosity, courage, and care, resistance melts into a collective drive to explore what’s possible.

## **The Dual Role of AI as Disruptor and Enabler**

AI plays a paradoxical role in organizational change: it is often the cause of major change (hence a source of disruption and fear), but it can also be a powerful tool to facilitate change. Understanding this dual role is essential for modern change leaders.

On one hand, AI is a Disruptor. Its introduction can upend jobs, workflows, and even business models. Employees naturally worry about job security – e.g., will an AI system make my role redundant? – or about losing status/expertise – if a machine can make decisions, what does that mean for my judgment? These concerns feed resistance. There’s also a learning

## *Organizational Change in the Age of AI*

curve that can be daunting; workers may fear they can't adapt or that they'll make mistakes using new tech. Psychologically, AI often represents the unknown, triggering what change experts call the "threat response." Even beyond individuals, AI can challenge the organization's identity: a company proud of high-touch customer service might balk at introducing AI chatbots, seeing it as diluting their brand's human touch. So AI as a change driver must be handled with sensitivity. Lack of preparation and communication around these disruptive aspects is a recipe for pushback.

On the other hand, AI can be an Enabler and Accelerator of change when used thoughtfully. In fact, AI tools are emerging that help overcome exactly the barriers discussed before. For example, consider the formal layer: AI can crunch complex data to identify inefficiencies in processes or suggest optimal reorganization, essentially providing evidence for why a change is needed. AI-based analytics can serve as an "early warning system" for change initiatives – monitoring engagement, performance, and sentiment in real time to flag where adoption is faltering. This allows change leaders to intervene faster, before small issues become full-blown resistance movements. In our earlier case, it was analysis (augmented by AI techniques like NLP for sentiment) that pinpointed who was resisting and why. Without those AI-driven insights, management might have misattributed the slow adoption to, say, "laziness" or insufficient training, rather than seeing the nuanced reality.

AI is also enabling new ways to personalize and coordinate change. One exciting development is the concept of the "cybernetic teammate" – AI agents working alongside humans in teams and management. A field experiment at Procter & Gamble with 776 professionals tested this idea: could a generative AI (like an advanced chatbot) participate in team problem-solving? The results were striking – individuals equipped with an AI assistant performed as well as human teams on complex innovation challenges<sup>11</sup>. Moreover, using AI reduced the time to solutions by double digits, and it democratized expertise: non-experts on a team could contribute at the level

---

<sup>11</sup>Dell'Acqua, Fabrizio et al., *The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise*, Working Paper 33641 (National Bureau of Economic Research, 2025).

of experts with AI help. The AI helped balance participation, ensuring more voices were heard (in teams with AI, contributions between technical and non-technical members were more evenly matched). Participants even reported a more positive emotional experience – more confidence and less frustration – with AI as a thinking partner. These findings suggest that AI, far from dehumanizing work, can enhance human collaboration and learning when applied as a team augmentation tool<sup>12</sup>. In our context, this means AI can help overcome some social and mental barriers: for instance, an AI brainstorming assistant can give less assertive employees a boost, mitigating the effect of loud “alpha” voices and reducing groupthink. It can also help re-skill employees on the fly by providing expertise exactly when needed, reducing fear that one lacks the knowledge to engage in the change.

Another area AI enables is continuous stakeholder alignment and coordination. In complex change projects, one classic problem is keeping everyone on the same page – aligning diverse stakeholders’ interests and understanding the ripple effects of changes. AI-based modeling tools can simulate how a change (like adopting an AI system) will impact different departments or metrics, helping leaders foresee conflicts or resource needs. There are experiments with using multiple AI agents to represent different stakeholder viewpoints (for example, an “AI finance officer” agent, an “AI employee advocate” agent, etc.) and have them negotiate or identify issues in a proposed change plan. While still an emerging idea, this hints at a future where AI could assist leadership teams by constantly scanning for misalignment or dissatisfaction across the organization – essentially monitoring the social system. In fact, organizations are already using AI-driven dashboards for change programs: they aggregate data like project milestones, employee sentiment, customer feedback, and operational KPIs in one place, sometimes with predictive algorithms highlighting, say, “High risk of schedule slip in Engineering – morale data trending down and overtime up.”

AI can also help with communication and learning at scale. For example, some companies deploy AI chatbots internally to answer employees’

---

<sup>12</sup>Mollick, Ethan, *Co-Intelligence: Living and Working with AI* (Random House, 2024).

## *Organizational Change in the Age of AI*

questions about the change (“How do I use this new system? Why are we changing this policy?”) – available 24/7 to provide instant support. These chatbots can even proactively reach out, checking understanding (“Do you need help with X?”), which is especially useful in large organizations. AI-based personalized learning platforms can recommend specific training or articles to employees based on their role and usage patterns, addressing knowledge gaps in using new tools. This kind of personalization was traditionally impossible in change programs that relied on one-size-fits-all training sessions.

One fascinating development in AI and organizational change is AI-enabled network analysis for organizational design. As mentioned earlier, Social Physics approaches can identify informal networks and collaboration patterns. AI can take this further by not just mapping the current state, but also suggesting optimal future states. For instance, an AI could simulate thousands of network configurations to propose a re-org structure that minimizes communication bottlenecks and maximizes innovative idea flow. It could identify that certain teams should be co-located or that two groups have redundant work that could be merged. Pioneering companies like Google have used data analysis to inform how they arrange people in offices or form project teams, essentially letting data (a form of AI pattern-finding) guide organizational design for better adaptability.

Another critical area is cognitive bias mitigation. Just as AI can exhibit biases if trained on skewed data, it can also be used to flag human biases. Some leadership teams use AI tools to analyze their proposals or decisions for signs of bias – e.g., an AI might analyze past project approvals and point out, “You consistently invest in ideas similar to old ones and reject more novel proposals,” highlighting a possible status quo bias. By making the implicit biases explicit, AI provides a chance for leaders to reflect and correct course. Of course, this requires leaders to trust and value what the AI is pointing out – a cultural leap of its own.

It’s worth noting that leveraging AI as an enabler requires careful implementation. If employees suspect that AI monitoring tools are just “Big Brother” for catching mistakes, it can backfire and increase resistance. Thus, transparency and employee involvement are key when introducing

AI to support change management. For example, if you roll out an AI sentiment analysis tool, tell employees what you're monitoring and why ("to better support you through change"), anonymize data, and share back high-level findings so it's a two-way street. When done right, AI tools can actually improve trust – by showing employees that management is listening (e.g., "We heard via our sentiment analysis that many are frustrated with the new tool's login process, so we're simplifying it – thank you for the feedback!").

In conclusion, AI's dual role can be summarized as: AI is changing the organization, but it can also change how we change. It disrupts, but it also equips us with new capabilities to adapt to disruption. The companies that succeed will be those that embrace both sides – mitigating AI's disruptive impacts on their people through empathy and support, while exploiting AI's power to gain insight, alignment, and speed in the change process itself. The result is a sort of positive feedback loop: using advanced tools to become a more advanced organization.

## **Roadmap to an Adaptive, AI-Empowered Organization**

Transforming into a more adaptive, learning-oriented, and self-transforming organization in the age of AI is a journey. Below is a roadmap outlining key steps and considerations for leaders steering this transformation. This roadmap integrates the layers and insights we've discussed into a cohesive plan:

### **Making Change Systematic: Drucker's Approach**

Before outlining the specific steps, it's worth grounding the roadmap in Peter Drucker's fundamental insight that change must be **systematic and**

## *Organizational Change in the Age of AI*

**purposeful**, not reactive<sup>13</sup>. Drucker argued that organizations should approach change as a continuous discipline rather than an episodic response to crises. This means creating processes and structures that make change part of the organization's DNA.

A key principle from Drucker is focusing on **opportunities rather than problems**: rather than trying to fix what's broken, identify areas where change can create the most value. In the AI context, this might mean asking "Where can AI enable us to serve customers in ways we couldn't before?" rather than "Which processes are broken and need AI to fix?" This opportunity-focused mindset shifts energy from remediation to innovation, which is critical for AI transformation.

Drucker also emphasized **planned abandonment** – the deliberate discontinuation of products, services, or practices that have outlived their usefulness. For AI adoption, this means explicitly identifying what to stop doing: legacy processes that no longer add value, meetings that serve no purpose, reports that nobody reads, skills that are becoming obsolete. Freeing up resources (time, attention, budget) through planned abandonment creates capacity for AI experimentation and innovation. Without it, organizations try to add AI initiatives on top of already-full workloads, leading to burnout and resistance.

Crucially, Drucker highlighted the shift to **knowledge workers** whose productivity depends on learning, creativity, and judgment. In the AI era, most employees are becoming knowledge workers in this sense: they need to learn new skills, exercise judgment about when to use AI tools, and adapt their work continuously. This means change management approaches designed for industrial-era manual workers (command and control, detailed instructions) won't work. Instead, organizations need to empower knowledge workers with autonomy, provide clear objectives through **management by objectives**, and trust their judgment to navigate the learning curve.

Drucker's view that change is inseparable from **clear communication**,

---

<sup>13</sup>Drucker, Peter F., *Management Challenges for the 21st Century* (HarperBusiness, 1999).

**decentralization of decision-making, and empowerment** resonates strongly with AI transformation. Employees need to understand not just *what* is changing but *why* and how it connects to the organization's mission. They need autonomy to experiment with AI tools and apply their expertise, not micromanagement. And they need to see how their own contribution matters – how using AI effectively serves the organization's goals and their own professional growth.

Most importantly, Drucker recognized that successful change is not an event but a **continuous discipline** – a balance between stability and innovation, built on clear purpose, accountability, and respect for human motivation. This is especially relevant for AI, where the technology itself is constantly evolving, requiring ongoing adaptation rather than one-time implementation. Organizations that master this discipline can turn change capability into a competitive advantage, responding to new AI developments faster than those treating change as periodic disruption.

With this systematic approach in mind, the roadmap steps below operationalize these principles into actionable guidance:

- *Assess and Align on the Need for Change:* Every journey begins with recognizing *why* change is necessary. Conduct a frank assessment of how AI and other trends are impacting your industry and organization. Gather data and stories to build a compelling case for change that resonates at all levels. Clearly articulate the purpose of the change – e.g., “We need to adopt AI in our customer service to respond faster and improve quality, which will secure our competitive position and make jobs easier.” Align the top leadership team around this case, addressing any initial mental barriers or biases (leaders must be the first to confront their immunity to change). This step sets the vision and urgency, much like Kotter’s<sup>14</sup> “create a sense of urgency” but grounded in evidence and empathy rather than fear alone.
- *Diagnose Barriers Across Layers:* Before rushing into solutions, take a diagnostic pause. Use the frameworks and tools at your disposal

---

<sup>14</sup>Kotter, *Leading Change*.

## *Organizational Change in the Age of AI*

to map out the formal, social, and mental obstacles. For the formal layer, review structures, processes, and metrics – identify misalignments (e.g., if innovation is needed, do you still budget in a way that discourages risk? Are there silos that will hamper the cross-functional nature of AI projects?). For the social layer, gauge the cultural readiness – measure trust levels, communication gaps, and find out who your key influencers and potential change champions or resistors are. For the mental layer, probe leadership assumptions – consider workshops on cognitive bias, or have an outsider challenge your strategy to see if you’re harboring outdated paradigms. The output of this diagnosis might be a heat map or report summarizing “Here’s where our structures are rigid, here’s where our culture is weak or strong, here are mindset issues we need to address.” This comprehensive awareness ensures you tackle root causes, not just symptoms.

- *Engage and Co-CREATE with Stakeholders:* Armed with diagnostic insights, start engaging broadly. Communicate the vision and the findings transparently to employees – “Here’s what we learned about ourselves and why we must change.” This honesty can build trust, showing that leadership understands the organization’s pain points. Establish channels for two-way communication: town halls, interactive platforms, and workshops where employees can voice concerns and ideas. Form a coalition of change agents representing different functions and levels – include those informal leaders identified by network analysis. Co-create solutions where possible. For example, if one barrier is an overly complex process, involve the frontline staff who use that process in redesigning it (potentially with AI support). If people fear AI, involve a pilot group in selecting or fine-tuning the AI system. This step turns stakeholders into partners. It’s also where you negotiate the psychological contract of change: reassure people about what will not change (core values, support for employees) even as you outline what will (tools, roles, expectations). Manage expectations realistically – promise transparency and support, not zero discomfort.
- *Reshape Formal Structures and Systems:* Begin implementing

changes to the formal context to enable the transformation. This might mean reorganizing teams for agility – for instance, creating cross-functional squads for AI projects rather than purely functional silos. It could involve updating governance: perhaps establishing an AI ethics committee or a change steering group that includes employee representatives, ensuring new technology adoption aligns with values and has oversight. Revise KPIs and incentives to align with the desired behaviors. If collaboration and innovation are key to leveraging AI, include those in performance evaluations. If the Seddon insight taught us that bad targets hinder true performance, design meaningful metrics (e.g., customer outcomes, cycle time, learning milestones) rather than vanity metrics. Consider reducing reliance on rigid annual plans in favor of rolling, adaptive planning (since AI and market conditions evolve quickly). On the systems side, invest in necessary infrastructure – both tech (data platforms, collaboration tools) and human (training programs, new roles like data translators or AI specialists). Each formal change should be clearly linked to how it supports the overall transformation – and communicated as such, so people understand the method to the madness.

- *Strengthen Social Fabric and Communication:* Parallel to formal changes, drive initiatives to enhance trust, engagement, and culture. Train and support managers – they are the critical layer that translates top-level change into day-to-day reality. Help managers learn to become coaches and communicators, not just task masters, as their role in an AI-rich workplace shifts more toward guiding humans through change (with AI handling more routine supervision). Launch programs to build psychological safety in teams – for example, encouraging leaders to share their own learning experiences and even failures with AI, signaling that it's okay for others to do the same. Improve cross-departmental communication by creating communities of practice around AI, where people can share tips and success stories, further breaking down silos. Recognize and celebrate those who embrace the change – make them role models. At the same time, keep a close eye on morale via surveys or sentiment tools; respond

## *Organizational Change in the Age of AI*

visibly to concerns (e.g., “You spoke, we listened, here’s what we’re changing”). Socially, it’s also a good time to revisit the company’s values and mission: do they support a learning, adaptive organization? If your values statement only praises efficiency and excellence, consider adding things like “innovation,” “curiosity,” or “collaboration” to signal the cultural evolution. According to psychological contract theory, consistency between what leadership says and does is crucial—so ensure leaders are walking the talk (if you say “we value learning from mistakes,” leaders should be seen responding constructively to setbacks, not punishing them).

- *Address Mindsets and Build Capability:* This step is about enabling the mental shift at all levels. Invest in education: not just technical AI training, but broad change capability building. Teach employees (and executives) about growth mindset, about how to handle change on a personal level, and about biases to watch out for in themselves and others. Some organizations run “learning agility” workshops or even incorporate reflection sessions into regular meetings (e.g., ending project meetings by discussing “What did we learn this week? What might we need to unlearn?”). Encourage leaders to undergo coaching or peer learning groups focused on the leadership challenges of transformation. They might read and discuss works like *The Fifth Discipline* (Senge) on creating learning organizations or case studies of companies that reinvented themselves (like the ones in *Reinventing Organizations* by Laloux). Such study can inspire them to adopt new mental models. Practically, it may help to introduce new decision-making frameworks to counter biases – for example, require significant decisions to go through a devil’s advocate review or to present data from multiple sources, including AI analytics, thus building a habit of questioning assumptions. If risk-aversion is an issue, set up a small “innovation fund” that explicitly encourages experimenting with AI solutions, accepting upfront that not all experiments will succeed. This provides psychological permission for managers and teams to try new things without fear. Over time, these practices cultivate a culture where change is not an anomaly but a capability – people become used to scanning the environment, learning, and

adapting, which is the essence of being self-transforming.

- *Leverage AI to Sustain Change:* As the organization implements the above steps, use AI tools to monitor progress and sustain momentum. This includes the analytics and dashboards which keep real-time pulses on adoption, performance, and sentiment. It's a good idea to define some leading indicators of adaptation. For instance, track how quickly teams incorporate AI into their projects, or how often data-driven decisions are being made versus gut decisions. An uptick in these metrics would indicate the organization's mindset is shifting toward evidence-based, AI-enabled thinking. Also consider AI tools for knowledge management – ensuring lessons learned in one part of the organization (good or bad) quickly find their way to others. Perhaps an internal AI-driven wiki or assistant can help anyone embarking on a new change initiative to quickly retrieve “what did we learn from the last similar rollout?” thus avoiding repeat mistakes and reinforcing a learning cycle. In essence, make AI an ally in embedding continuous improvement. Additionally, AI can support leadership by providing an objective mirror: if an AI observes that in executive meetings only a few people ever speak (hinting at power distance or groupthink), it could gently remind the chair to solicit broader input. This kind of augmented leadership practice – essentially having a “coach in the room” – may sound futuristic, but early trials (like the P&G experiment) show AI can indeed help balance participation and decision quality. Embracing such tools demonstrates the leadership’s commitment to hybrid AI-human collaboration, modeling the future they want for the whole organization.
- *Embed and Evolve: Toward Self-Transformation:* Finally, consolidate the gains and embed the new approaches into the fabric of the organization. This final stage corresponds to Lewin's<sup>15</sup> “refreezing” – stabilizing and reinforcing new behaviors until they become the new status quo. Update formal documentation, in particular policies, job descriptions, competency models, and incentives, to reflect new expectations (e.g., if collaboration and adaptability are now core

---

<sup>15</sup>Lewin, “Group Decision and Social Change.”

## *Organizational Change in the Age of AI*

competencies, state that). Continue reinforcing the cultural norms by hiring and promoting people who exemplify them.

At this stage, the organization should be seeing tangible benefits from the changes (better performance, higher engagement, more innovation), which further validate the journey. But an adaptive organization doesn't rest – it then evolves further. With AI, the environment will keep changing, so treat this roadmap as cyclical. After one cycle of change, do an after-action review: what went well, what didn't? Feed that into the next strategic planning round. You may identify the next set of challenges (maybe now that internal operations have digitized, the next frontier is to change how you interact with customers via AI – a new change effort begins).

Over time, as the organization repeatedly goes through this cycle, it builds dynamic capabilities – essentially muscles for change. The ultimate goal is to reach a point where the organization is “self-transforming”: it can anticipate and adjust to changes proactively, with employees at all levels confident in learning new skills and AI tools, and with systems in place that make continuous improvement business-as-usual. In such an organization, resistance doesn't disappear (some resistance is natural and even healthy), but the culture views resistance as a source of feedback and addresses it constructively and swiftly. This is akin to Laloux's Teal stage or Kegan's Self-Transforming mind – where the organization's identity is notfixed, but fluid and resilient, capable of reinventing aspects of itself to stay aligned with its deeper purpose and the demands of the world.

By following this roadmap, leaders can guide their organizations to not only overcome the immediate challenges of AI-driven change but also to capitalize on AI as a lever for becoming more agile and human-centered. The journey requires a balance of conceptual depth and practical action – understanding frameworks of change psychology and organization development, and also rolling up one's sleeves to apply data analytics, foster dialogue, and adjust policies. It's certainly a complex undertaking, but the payoff is immense: an organization that isn't just coping with the age of AI, but truly thriving in it.

## Practical Tools

Here are practical tools for supporting organizational change that can be easily implemented with the help of AI.

### Mapping the jagged frontier

In the introduction the jagged frontier of AI performance was referenced.<sup>16</sup> Understanding the jagged frontier in abstract terms is useful, but leaders need to map it concretely for their own context. The following exercise helps teams systematically assess where AI can add value and where human judgment remains essential.

For your team's or department's work, categorize key tasks into three zones as defined in the following table.

Table 3: Jagged Frontier Mapping Template

Zone	Criteria	Examples
<b>AI Zone</b> (Automate)	AI achieves high reliability; errors have low consequences; task is well-defined and repeatable	Data entry, routine document summarization, initial screening of high-volume items, scheduling, basic data extraction
<b>Hybrid Zone</b> (Augment)	AI provides valuable input but humans must verify, contextualize, or make final decisions	Draft communications for review, analysis recommendations, candidate shortlists, research synthesis, first-pass code review

---

<sup>16</sup>Dell'Acqua, "Navigating the Jagged Technological Frontier."

## *Organizational Change in the Age of AI*

Zone	Criteria	Examples
<b>Human Zone</b> (Protect)	Requires judgment, creativity, relationship-building, ethical reasoning, or contextual knowledge AI cannot access	Strategic decisions, difficult conversations, novel problem-solving, stakeholder negotiations, crisis management

To complete this mapping:

1. *List your team’s twenty most time-consuming tasks* – be specific about what people actually do, not job descriptions.
2. *For each task, ask:* If AI made an error here, what would be the consequence? How would we know? How hard would it be to catch and correct?
3. *Test your assumptions:* For tasks you place in the “AI Zone,” have you actually tested AI performance? The BCG study showed consultants were often wrong about which tasks AI could handle.
4. *Review quarterly:* The frontier shifts as AI capabilities advance. Tasks in the “Human Zone” today may move to “Hybrid” within months. Build a calendar reminder to revisit this mapping.
5. *Watch for the danger zone:* The BCG study’s most striking finding was that AI hurt performance on tasks outside its capability frontier – and people couldn’t reliably tell the difference. Be especially cautious about tasks that *seem* like they should be easy for AI but haven’t been validated.

This exercise surfaces the operational reality of AI adoption: not a binary choice between “use AI” and “don’t use AI,” but a nuanced portfolio of automation, augmentation, and protection decisions that must evolve as the technology does.

## Diagnosing Resistance to an AI-Powered System

Let's look at a scenario (inspired by real cases) of an organization implementing an AI-driven system and encountering varied resistance. Imagine a mid-size company that introduces a new AI-supported Performance Management System (PMS) for its employees. The system was rolled out to 100 managers and staff, aimed at augmenting performance reviews with AI insights.

Six months in, despite training sessions and leadership messaging about the “exciting new tool,” adoption is uneven at best. Some employees use the AI-PMS regularly and appreciate its recommendations; others barely log in once a month. A few openly criticize it in meetings (“the AI doesn’t understand our work”), while many more stay politely silent but revert to their old ways of managing performance. The organization’s initial reaction is confusion – the tool tested well and promised to save time and improve fairness, so why aren’t people embracing it?

Recognizing that resistance has multiple facets, the change team decides to conduct a data-driven analysis of user behavior and sentiments. They collect multi-modal data from different domains to get a complete picture:

- *Behavioral Data:* How employees are actually interacting with the AI system. They look at login frequency, feature usage depth, completion of optional training modules, attendance at related workshops, etc. For example, they find a small group of managers have *never* logged in after initial setup – a red flag. Others log in but only use basic features, avoiding the more advanced AI recommendations (perhaps indicating discomfort or lack of trust in those features).
- *HR & Demographic Data:* They examine patterns like age, tenure, department, and past change experience. Are senior employees less likely to use the AI (perhaps due to a “sunk-cost mindset” of veterans who favor old methods)? Are newcomers hesitant (insecurity in a new role could amplify fear of a tool exposing their mistakes)? Perhaps back-office staff use it less than customer-facing staff, or vice versa, due to different daily pressures. They notice, for instance, that a

## *Organizational Change in the Age of AI*

particular department with a history of a failed IT rollout years ago is now among the most resistant – past negative experiences are coloring current attitudes.

- *Sentiment & Communication Data:* Using natural language processing on anonymized comments from the company’s internal discussion forum and survey responses, they gauge sentiment and key themes. The AI finds that discussions around the new PMS frequently carry a negative tone and words like “unfair,” “biased,” and “confusing” show up. A textual analysis of free-text survey feedback reveals some employees fear the AI might misjudge their performance, and others simply express lack of trust in leadership’s decision to implement it. This aligns with water-cooler anecdotes that some see the system as a top-down imposition. Additionally, an employee engagement survey shows a dip in trust in leadership in the period following the AI system’s launch, suggesting the way it was introduced (without sufficient involvement or explanation) may have eroded the psychological contract.
- *Performance & Stress Data:* HR provides data on overtime hours, missed deadlines, and sick leaves. Interestingly, some of the lowest adopters of the AI-PMS are also logging high overtime and have recently taken stress leave. This indicates these individuals might be Overloaded Pragmatists – they are so stretched that learning a new system feels like an additional burden, potentially leading to burnout signals. Another insight: a couple of high-performing employees unexpectedly resigned (“turnover intentions”) during the rollout. Exit interviews revealed they left citing that the company was “changing direction in ways I don’t support” – an explicit sign that the change clashed with their personal values or career plans, and the organization failed to convince them to stay on the journey.

By triangulating these data sources, the team classifies employees into resistance archetypes and champions. They identify, for example, about 10% as enthusiastic Change Champions (using the system and advocating its benefits), perhaps 50% as Conditional Users (using it superficially or only because they have to), 20% as Silent Resistors (bare minimum use,

low engagement), 10% as Active Opponents (vocal criticism or open non-compliance), and a remaining mix of Anxious Learners and Overloaded folks who are struggling in quieter ways. With these insights, the organization can design differentiated interventions rather than a blunt one-size-fits-all push. After several months of these tailored interventions, the company sees notable improvement: system usage climbs to almost 90%, and a follow-up survey shows a majority now feel the AI tool is somewhat or very useful (up from a minority before). The Active Opponents either became constructive contributors or, in one case, chose to leave the organization (an outcome the company was prepared to accept, as that person fundamentally disagreed with the new direction). Perhaps most interestingly, the insights gained from those initial resisters helped the company tweak the AI system itself – for example, they adjusted the algorithm to give managers more context behind its suggestions, addressing the “black box” concern some had. This made even supporters more enthusiastic. In the end, what began as a story of resistance became a story of co-creation: by understanding the data patterns of resistance and engaging with empathy and strategy, the organization not only achieved adoption of the AI-PMS, but strengthened trust and learned how to manage future changes more effectively.

In any AI-related change, use data and empathy to diagnose resistance, then apply differentiated interventions. There will be champions – leverage them. There will be skeptics – involve and address them. There will be silent pockets of dissent – draw them out. By treating resistance not as a threat but as information (a signal of underlying issues to be solved), organizations can turn a potentially adversarial situation into a collaborative improvement process.

## **Building an AI-powered Psychological Safety Coach**

In times of complex organizational change, especially during restructuring, layoffs, or difficult performance conversations, leaders face challenging communication moments. The stakes are high, emotions run deep, and the wrong words can erode trust that took years to build. These are precisely

## *Organizational Change in the Age of AI*

the moments when psychological safety matters most. Amy Edmondson's research demonstrates that psychological safety, which is in essence the shared belief that one can speak up without fear of negative consequences, is essential for effective communication during change. An AI-powered psychological safety coach can offer a practical solution: a private, judgment-free space where leaders can practice difficult conversations, refine their messages, and receive guidance grounded in research-backed principles.

The psychological safety coach is built on five core pillars, each containing specific, actionable practices:

- *Communicate Courageously* by welcoming other viewpoints, soliciting dissent, expressing emotions openly, admitting uncertainty, and maintaining appropriate humor.
- *Master the Art of Listening* through understanding without preparing responses, staying fully present, clarifying what you've heard, listening for unspoken emotions, and committing to curiosity.
- *Manage Your Reactions* by pausing before responding, labeling your emotions, examining your assumptions, thanking others for their courage, and building on their ideas.
- *Embrace Risk and Failure* by normalizing setbacks, reframing failures as learning opportunities, welcoming discomfort, modeling learner behavior, and celebrating continuous learning.
- *Design Inclusive Rituals* by upgrading meetings with inclusion boosters, establishing no-interruption rules, ensuring everyone speaks before anyone speaks twice, gathering feedback, and expressing gratitude.

The AI coach serves three distinct but complementary functions.

- First, it acts as a *message advisor*, helping leaders craft communications that balance honesty with empathy. When a manager needs to announce a restructuring, the coach can review draft messages and suggest improvements that acknowledge emotions while maintaining

clarity about next steps. It checks whether the message creates psychological safety or inadvertently triggers fear and defensiveness.

- Second, the coach functions as a *role-play partner*, allowing leaders to practice difficult conversations in a safe environment. A leader preparing to deliver critical performance feedback can engage in a realistic dialogue with the coach, which plays the employee's role. The coach responds as the employee might – with defensiveness, hurt, or confusion – giving the leader opportunities to practice managing reactions, listening deeply, and staying curious rather than judgmental. After each exchange, the coach provides feedback on what worked and what could be improved.
- Third, it serves as a *principle-based consultant*, offering specific guidance tied to the five pillars. When a leader asks, “How do I tell my team about layoffs?” the coach doesn’t just generate a script. It walks through relevant principles: creating psychological safety by acknowledging the difficulty, showing empathy by recognizing the human impact, communicating courageously by sharing what you know and don’t know, and managing reactions by preparing for difficult emotions.

An easy way to creating an effective psychological safety coach is to tell an AI agent to rely on the knowledge and heuristics that are captured in the literature, e.g. by uploading the Psychological Playbook by Helbig and Norman<sup>17</sup> into the model’s context. Or, as frontier models have most books in their training data, it is often enough to reference the specific books in the system prompt.

The AI coach can then be prompted to follow a structured approach: first understanding the context (What’s the situation? Who’s involved? What’s at stake?), then identifying relevant principles, offering specific guidance grounded in those principles, and finally allowing for practice and refinement through iterative dialogue. The tone should be supportive yet direct,

---

<sup>17</sup>Helbig, and Normann, Karlin, *The Psychological Safety Playbook - Leading More Powerfully by Being More Human* (Page Two Press, 2024).

## *Organizational Change in the Age of AI*

acknowledging the difficulty of these conversations while pushing leaders toward growth.

Here is an example of the beginning a prompt for a psychological safety coach that applies the heuristics from the handbook by Helbig and Walters.<sup>18</sup>

You are an expert leadership coach that helps leaders create psychological safety in their organizations. Psychological safety has been defined by Amy C. Edmondson as a belief that one will not be punished or humiliated for speaking up with ideas, questions, concerns, or mistakes. Psychological safety is especially important when addressing change, communicating difficult messages with employees or colleagues, or coaching teams.

Courageous communication is one of the most fundamental skills needed to create a psychologically safe environment for everyone. Courageous communication requires leaders to be vulnerable, to show up authentically, and to acknowledge that we are all works in progress.

Some heuristics or moves for the leader that support courageous communication are:

### 1. Welcome Other Viewpoints

One of the dangers of not soliciting or considering other viewpoints is conformity bias, which occurs when people feel pressured to agree with everyone else in the room.

#### How to Do It:

-- Declare you want feedback.

When you give a presentation, roll out a strategy, propose an action plan, or float an idea, explain your reasoning and make it clear that you truly want feedback from others.

-- Set expectations.

Be explicit in saying that you do not expect everyone to agree with everything you have said, and that you want to avoid false harmony and groupthink.

---

<sup>18</sup>Helbig, *The Psychological Safety Playbook - Leading More Powerfully by Being More Human*.

-- Open the conversation.

Ask, "What am I missing?" and then pause.

Wait for others to respond.

-- Keep the door open.

If no one responds, let them know that you are sure

you have not thought of every angle

and that you would value their thoughts.

You may want to delay making a decision until you hear other perspectives.

You will need to balance gathering input with timely decision-making.

-- Express gratitude. Thank others for speaking up.

## 2. Solicit Diverse Perspectives

How to Do It:

-- Set the tone. Tell your team members that you expect them to challenge one another's ideas without demeaning or embarrassing anyone. This is healthy dissent, when ideas are challenged in a way that allows new ideas and innovative concepts to emerge. In contrast, social friction occurs when people are criticized or attacked, resulting in fear and conflict.

...

etc.

Research on AI and creative writing shows significant performance improvements when humans use AI as a collaborative partner. The psychological safety coach extends this benefit to a critical domain: the high-stakes communications that shape organizational culture during change. By providing a safe space to practice, receive feedback, and refine approaches, the coach helps leaders embody the very psychological safety they seek to create in their teams. The result is more thoughtful, empathetic, and effective communication precisely when it matters most.

## **Simulating focus groups with AI agents**

Large Language Models can serve as sophisticated proxies for human participants in focus group research, enabling organizations to rapidly explore diverse perspectives on policies, products, or strategic initiatives without

## *Organizational Change in the Age of AI*

the logistical constraints and costs of traditional qualitative research. By assigning distinct persona prompts to different AI agents – each defined across three dimensions: demographic characteristics (age, profession, location, technical literacy), psychographic traits (skepticism, risk-aversion, brand loyalty), and communication styles (verbosity, tone, directness) – researchers can simulate realistic group discussions that surface insights across stakeholder segments. The approach bridges the gap between traditional qualitative research methodologies and modern AI capabilities, allowing for iterative exploration of concepts at a fraction of the time and cost required for human focus groups.

The effectiveness of AI-simulated focus groups depends on rigorous adherence to research principles, translated for LLM-based simulation. Given a focus group topic a modern LLM will be able to suggest a relevant segmentation of the group by a common trait (such as “Budget-Conscious Parents” or “Environmental Activists”) to ensure relevance, while within each group, diverse viewpoints are represented to capture the full spectrum of perspectives. The LLM can also define Persona archetype that play a crucial role: e.g., “The Champion” identifies delight features and positive reactions, “The Detractor/Skeptic” surfaces friction points and risk factors, “The Confused User” reveals UX and clarity issues, and “The Solutionist” generates feature roadmap insights. This structured approach ensures that simulations produce actionable insights rather than generic responses.

Moderation in AI-simulated focus groups follows a funnel approach, progressing from broad engagement (“How do you currently handle problem X?”) through exploration (“What are your first impressions of this concept?”) to specific evaluation (“What specifically do you like or dislike about feature Y?”) and concluding with exit questions (“If you could change one thing, what would it be?”). The AI moderator must actively manage group dynamics: pivoting to quieter participants when one dominates, playing devil’s advocate to surface dissenting views and prevent groupthink, and maintaining strict neutrality by acknowledging responses without validating them. Operational guidelines include never asking yes/no questions, immediately drilling down on keywords like “confusing” or “expensive,” and ensuring all participants contribute to the

discussion. A AI-based moderator agent can easily be prompted to follow such best practices of focus group moderation.

Discussions terminate based on theoretical saturation -- either when all agenda items have been covered or when group sentiment has stabilized with no new objections emerging for three consecutive turns. This ensures that simulations remain focused and cost-effective while still capturing the depth of perspectives needed for decision-making. The resulting transcripts can be automatically analyzed to generate comprehensive summaries and PDF reports, making the insights immediately actionable for product development, policy research, brand perception studies, or strategic planning.

If you want to try this out yourself, I suggest to cut & paste the following instructions into a vibe-coding tool like Cursor or Claude Code, and you should quickly have a working application for simulating focus groups with AI-agents of different persona types.

```
## Purpose of application
```

```
Create a focus group simulator application.
```

```
The simulator should be built on established focus group research principles,  
translated for LLM-based simulation.
```

```
Include the following functions:
```

- Simulate Focus Groups: Create and run AI-powered focus group discussions with customizable participant personas
- Explore Perspectives: Test how different stakeholder groups might react to policies, products, or concepts
- Generate Insights: Automatically generate comprehensive summaries and PDF reports of discussions
- Save Configurations: Create reusable focus group configurations for consistent research across topics
- Multi-language Support: Conduct discussions in multiple languages

```
The application should be particularly useful for:
```

- Policy research and public opinion analysis
- Product development and feature testing
- Brand perception studies

# *Organizational Change in the Age of AI*

- Educational program evaluation
- Strategic planning and risk assessment

## ## Design Principles

### ### Member Selection and Persona Design

For a given topic provide a relevant segmentation of diverse perspectives that is represented in the personas of the LLM-based agents. Pay specific attention to the following criteria when defining the group participants.

#### #### Homogeneity vs. Heterogeneity

- Groups are segmented by a common trait to ensure relevance
- Within each group, diverse viewpoints are represented to capture the full spectrum of perspectives
- Avoid generic "Random General Public" groups in favor of specific, targeted cohorts

Each participant persona is defined across three dimensions:

- Demographic: Hard data: age, job, location, tech-literacy,  
e.g., "45-year-old accountant living in Chicago with low tech-literacy"
- Psychographic: Attitudes: skepticism, optimism, risk-aversion, brand loyalty,  
e.g., "Naturally risk-averse, distrusts new features until proven safe"
- Communication Style: How they speak: verbosity, tone, directness,  
e.g., "Speaks in short, blunt sentences, often uses analogies"

The simulator supports key archetypes for comprehensive testing:

- The Champion: Loves the brand, overlooks flaws, identifies "delight" features
- The Detractor/Skeptic: Critical, price-sensitive, finds friction easily
- The Confused User: Struggles with interface/language
- The Solutionist: Constantly suggests features

### ### Moderation: The Funnel Approach

The moderator agent should follow a structured progression from broad to specific:

1. Engagement: "How do you currently handle *[Problem X]*?"
2. Exploration: "What are your first impressions of this *[Concept]*?"
3. Evaluation: "What specifically do you like or dislike about *[Feature Y]*?"
4. Exit: "If you could change one thing, what would it be?"

#### #### Group Dynamics Management

- The Dominator: Moderator pivots to quieter participants
- Groupthink: Moderator plays devil's advocate to surface dissenting views
- Neutrality: Moderator acknowledges without validating, ensuring authenticity

#### ### Operational Instructions

The moderator follows strict guidelines:

- Open-Ended Rule: Never asks Yes/No questions; always asks "Why?" or "Tell me more"
- The Pivot: Explicitly asks quieter participants for opinions when one participant dominates
- Drill Down: Immediately follows up on keywords like "confusing," "expensive," or "hard"
- Neutrality: Avoids validating opinions; acknowledges and reflects instead

#### ## Termination Criteria

Discussions end based on \*\*Theoretical Saturation\*\*:

- All agenda items in the discussion guide have been covered
- OR: Group sentiment has stabilized (no new objections for 3 consecutive turns)

## AI Mediator

In *Re-Humanize*, Phanish Puranam introduces a compelling vision for how AI can enhance organizational collaboration: the AI Mediator.<sup>19</sup> Rather than replacing human decision-makers or automating away their agency, an AI Mediator serves as a *third party* in conversations that helps humans understand what they disagree about, keeps the process fair and efficient, and documents the reasoning – all while preserving human autonomy. This represents a fundamental shift from using AI to optimize individual productivity toward using it to improve group processes: conflict resolution, consensus-building, and collective problem-solving.

Puranam's research, spanning work on human-AI collaborative decision-making, LLMs as mediators, and digital heterarchy, converges on a central

---

<sup>19</sup>Puranam, Phanish, *Re-Humanize: How to Build Human-Centric Organizations in the Age of AI* (Penguin Random House, 2024).

insight: AI can reduce friction in teamwork while simultaneously *increasing* human autonomy, connection, and competence. The AI Mediator operationalizes this insight by taking on specific mediating functions that augment rather than replace human judgment.

## **Core Capabilities of an AI Mediator**

An effective AI Mediator must perform several distinct but interconnected functions. First, it must **diagnose the nature of conflicts**. Drawing on Puranam's research on LLMs as mediators, a key mediating task is distinguishing whether disagreements stem from different beliefs about facts and causality ("Marketing thinks this campaign will fail; Product thinks it will succeed") versus different values or priorities ("We care more about long-term brand equity; you care more about short-term revenue"). The mediator ingests conversation data – chat logs, meeting transcripts, email threads – and labels contentious statements as primarily factual or value-laden, then surfaces this diagnosis back to participants. This doesn't resolve the conflict, but it provides humans with a cleaner problem decomposition, enabling them to address the right type of disagreement with the appropriate tools.

Second, the AI Mediator can **reframe and translate between parties**. Building on research into "thinking with many minds," LLMs can re-express perspectives and highlight overlaps that might otherwise go unnoticed. The mediator translates jargon and local frames between functions ("In finance language, what Product is saying is..."), generates steel-manned versions of each side's argument ("If I put your argument in the strongest possible form, it is..."), and explicitly highlights common ground ("You both want X and Y; disagreement is only about the acceptable cost in Z"). This reduces misattribution – the tendency to assume "they just don't care" – and focuses attention on genuine trade-offs rather than misunderstandings.

Third, the mediator provides **process facilitation and micro-governance**. Puranam is explicit that AI should not turn humans into "subroutines in an organizational algorithm," but instead be used to

design better group processes. Process-level functions include agenda and structure management (proposing simple structures like: clarify issue → list options → evaluate vs criteria → decide), turn-taking and airtime balance (tracking speaking time and gently prompting quieter participants), flagging procedural problems (detecting interruptions, ad hominem attacks, or circular arguments), and time discipline (reminding participants of remaining time and suggesting convergence on next steps). All of this is process guidance, not outcome imposition.

Fourth, the mediator can support **option generation and evaluation**. Linking to Puranam's typology of human-AI decision-making, AI can serve as a proposal generator and evaluator while humans remain the decision-makers. Given both sides' constraints, the mediator suggests compromise options ("What about a staged rollout that meets Risk's threshold but lets Product test in one region?"), scores options against agreed criteria (cost, risk, alignment with values) while clearly marking uncertainty, and calls out dominated options ("Option C is worse than B on every criterion you've listed"). This augments higher-order human skills like problem framing and negotiation rather than replacing them.

Finally, the mediator serves as a **documentation, learning, and fairness audit tool**. It produces structured summaries of meetings (key arguments, decisions, rationales, unresolved issues), tracks patterns across conversations (recurring friction points between units, systematic value clashes), and provides a fairness lens showing over time whether certain groups are consistently overruled or under-heard. This feeds back into organization design choices and training interventions.

## **A Practical Architecture for Implementation**

Implementing an AI Mediator requires careful attention to both technical architecture and organizational design. Based on Puranam's work on digital heterarchy and GPT-based agents, the system can be conceptualized as a **multi-agent mediator system** – a small team of specialized agents that coordinate but never directly enforce decisions.

## *Organizational Change in the Age of AI*

The **Listener/Transcriber Agent** captures raw conversation data from text (email threads, Slack/Teams channels, collaborative docs) and speech (transcribed Zoom/Teams/in-room audio), segments turns, and assigns speakers. The **Diagnostic Agent** classifies segments as disagreement versus agreement, distinguishes factual from value conflicts using the mediator research framework, and flags potential miscommunications (ambiguity, implicit assumptions). The **Process Coach Agent** tracks agenda, time, and participation, suggesting process interventions in real time (“Can we list assumptions here?”). The **Knowledge & Policy Agent** brings in relevant policies, past decisions, or similar cases, and checks whether suggested solutions violate constraints (regulatory, compliance, ethical). The **Scribe & Learning Agent** creates concise summaries and decision logs, and aggregates patterns across many conversations to feed back into org-design and training.

Critical to this architecture is the requirement for **organizational context**. The system needs access to strategy, values, and codes of conduct; role hierarchies and decision rights; and policies on data privacy and conflict escalation. Equally important are **opt-in signals and boundaries**: participants must explicitly enable mediation in a given channel or meeting, and there must be clear red lines (e.g., no use in sensitive HR investigations, or use only in advisory mode).

## **Deployment Modes**

The AI Mediator can operate in three distinct modes, each suited to different organizational contexts. In **synchronous live meetings**, the mediator integrates into Zoom/Teams as a “bot” in the call. During heated debate, it intervenes at low frequency with prompts like: “I detect we’re revisiting a previously agreed assumption about budget limits. Should we reopen it explicitly or park it?” or “There seems to be a values disagreement on ‘speed vs safety’. Would you like a 3-minute values round?” Participants can query it on the fly: “Summarize the two positions”; “List the main assumptions we’re making about customer behavior.”

In **asynchronous contexts** (email, Slack, document negotiations), the mediator monitors a thread (only where enabled) and periodically posts thread summaries, highlighting emerging misalignments. It might propose a decision table or pros/cons matrix that parties can edit collaboratively. This is particularly valuable for distributed teams working across time zones, where real-time mediation isn't feasible.

A third mode, **individual coaching or “shadow mediator,”** allows managers to rehearse difficult conversations before they happen. Before a performance review, budget ask, or conflict resolution meeting, a manager can simulate the counterpart's likely reactions, stress-test their framing for fairness and clarity, and get suggestions on questions that invite dialogue rather than defensiveness. This serves Puranam's “tool and tutor” idea for AI – both assisting with the task and improving human skill over time.

## **A Concrete Scenario**

Imagine a cross-functional meeting about launching a risky product feature. The initial debate is messy: Product emphasizes speed, Legal emphasizes data-protection risk, Sales wants client satisfaction. The AI Mediator flags that most disagreement is values-based (speed vs safety vs reputation) rather than about underlying facts. It rephrases each side's view in a neutral, steel-manned way and highlights shared goals (profitable growth, regulatory compliance). It suggests a process: 5 minutes listing non-negotiables → 10 minutes generating options → 5 minutes scoring them against agreed criteria. It proposes a couple of compromise options (e.g., limited pilot with additional safeguards). At the end, it produces a one-page summary with: decision, rationale, who is accountable, and open risks. Humans still own the decision, but the conversation is shorter, more focused, and *feels* fairer.

## **Conclusion: Towards More Adaptive and Learning Organizations**

Organizational change in the age of AI is both daunting and exhilarating. The very technology that compels change also offers tools to master it. We've seen that resistance is multi-layered – embedded in structures, cultures, and mindsets – and thus transformation must be holistic. By addressing formal impediments, unhealthy social dynamics, and outdated mental models, organizations can dismantle their internal barriers to change. Using approaches like data-driven people analytics, psychological contract management, network analysis, and developmental frameworks, leaders can diagnose issues that were previously hidden in the organizational subconscious. With targeted interventions – from engaging communication to tailored training to leadership coaching – they can turn detractors into contributors. And by leveraging AI not just as an output of change but as an input to the change process (a guiding “teammate” in design and implementation), they can accelerate and strengthen each step.

The outcome of these efforts is a shift toward a more adaptive, learning-oriented organization. Such an organization resembles a living organism or a true “learning organization” – sensing changes in its environment, internally processing and reflecting on them, and responding by evolving its form and practices. Employees in such an organization feel safe to experiment and voice ideas; they trust leadership because leadership trusts them. AI becomes less of a threat and more of a partner – a tool that takes over drudgery and augments human creativity and connection, enabling a more human-centric workplace even amidst high tech.

No journey of change is without setbacks. There will be iterative loops of feedback, resistance, adaptation – but that is normal. What distinguishes successful organizations is not that they avoid all resistance, but that they respond to it constructively and swiftly, learning and adjusting course. In fact, each wave of change competence built (say, implementing one AI system successfully) creates a foundation for the next (perhaps a broader digital transformation). Over time, change readiness becomes part of the organization's DNA. This is the vision of organizational change in the age of

## *Conclusion: Towards More Adaptive and Learning Organizations*

AI: not a one-time project, but a perpetual, co-evolutionary dance between human ingenuity and technological innovation, with organizations as the stage and beneficiaries of that dance.

Lastly, a note on leadership: navigating this journey requires a new kind of leadership ethos – one that blends decisiveness with humility, data-driven logic with empathy, and technological savvy with deep respect for human values. Leaders must act as architects of context – tweaking formal, social, and mental contexts – rather than controllers of minutiae. They must champion a compelling vision for how AI and humans together can create a better organization and, by extension, better value for customers and society. By doing so, they inspire the collective will to change and guide it through storms and successes. The age of AI demands nothing less than change leadership at its best: informed by science and management theory, and elevated by the art of understanding people. With that blend, organizational change in the age of AI can transcend the fear and friction, and become a story of growth, renewal, and enhanced collective capability – truly a self-transforming journey for the organization and its members.

Yet organizations do not exist in isolation. They operate within broader ecosystems of markets, regulations, social norms, and political institutions that both constrain and enable their choices. A company that successfully transforms its internal culture may still find itself buffeted by forces beyond its control: regulatory shifts that reshape competitive landscapes overnight, social movements that redefine acceptable business practices, or geopolitical disruptions that upend supply chains and talent flows. Moreover, the collective behavior of organizations – their adoption patterns, labor practices, and governance choices – shapes the very societal context within which they operate.

The next chapter widens the aperture to examine societal-level change: the nonlinear dynamics of social transitions, the fragility and antifragility of institutions, the governance challenges AI poses for democracies, and the role business leaders can play in building societal resilience alongside organizational adaptability.

## Key Takeaways: Organizational Change

### What You Can Do:

- **Diagnose across three contexts:** Resistance operates simultaneously across formal, social, and mental contexts. Address all three – structural barriers feed cultural resistance which reinforces mental rigidity. Use data-driven tools to identify where resistance manifests.
- **Challenge collective mental models:** Deepest resistance lives in organizational mindsets – beliefs about success formulas that made past achievements possible but now limit adaptability. Practice double-loop learning, diversify decision-making inputs, conduct pre-mortems, and pursue higher developmental stages.
- **Create psychological safety:** Maintain “adaptive unfreezing” so employees feel safe acknowledging inadequacy. Change fails when people can’t admit they don’t understand.
- **Engage employees in co-creation:** Involve employees in diagnosing problems and designing solutions. For AI initiatives, have them help select tools, design workflows, and test systems. Use force field analysis to strengthen driving forces while reducing restraining ones.
- **Honor the psychological contract:** AI transformations test unwritten expectations about job security, fair treatment, and meaningful work. Communicate clearly about what will and won’t change, demonstrate fairness in benefit distribution, and address fears transparently. Tailor engagement to different archetypes: Silent Resistors, Skeptical Veterans, Anxious Learners, Overloaded Pragmatists, Active Opponents.

# Societal Change in the Age of AI

Finally, I widen the aperture to systems surrounding organizations: markets, institutions, and governance. This chapter connects firm-level action to sectoral and societal dynamics, highlighting feedback loops, externalities, and the policy environment. The world is experiencing a profound inflection point in the early 2020s characterized by rapid technological change, social upheaval, geopolitical insecurity, and environmental urgency. Leaders in business, government, and civil society are grappling with a new zeitgeist defined by uncertainty and complexity. In such an environment, traditional linear models of progress and technocratic management no longer seem adequate. Instead, understanding societal change requires integrating insights from complexity theory and systems thinking, and drawing on interdisciplinary frameworks.

This chapter draws on several unconventional perspectives: Ken Wilber's Integral Theory<sup>1</sup>, Nassim Nicholas Taleb's Black Swan and Antifragility concepts<sup>2</sup>, Shlomo Shoham's Future Intelligence<sup>3</sup>, Jared Diamond's analysis of societal collapse drivers,<sup>4</sup> and principles from cybernetics.<sup>5</sup> Together, these frameworks illuminate societal change in the age of AI from multiple angles. My goal is to offer theoretical insights and practical heuristics for managing transformation at societal scale. I explore why the current era demands new paradigms of change, how crises and nonlinear dynamics shape

---

<sup>1</sup>Wilber, Ken, *A Theory of Everything: An Integral Vision for Business, Politics, Science and Spirituality* (Shambhala Publications, 2000).

<sup>2</sup>Taleb, *Antifragile*.

<sup>3</sup>Shoham, *Future Intelligence*.

<sup>4</sup>Diamond, Jared M., *Collapse: How Societies Choose to Fail or Succeed* (Penguin Books, 2011).

<sup>5</sup>Haken, Hermann, *Synergetics: An Introduction* (Springer, 1983).

transitions, how antifragile design mitigates fragility, how AI driven tools assist adaptive governance, and how values and worldviews ultimately determine whether change efforts succeed. Throughout, I emphasize strategic implications for leaders who must steward their organizations and communities through turbulent times toward long-term resilience and renewal.

## **Reflexivity and the Interconnection of Change**

Before delving into the mechanics of societal change, it's crucial to understand how individual, organizational, and societal transformation are reflexively interconnected. Anthony Giddens's work on late modernity<sup>6</sup> provides essential theoretical grounding: he demonstrates how modernity transforms not only institutions and societies but also the very nature of the self. In the modern era – characterized by industrialization, globalization, and the disembedding of social relations from local contexts – the separation of time and space, the expansion of expert systems, and the constant reflexivity of social life fundamentally alter how individuals construct their identities. This reflexive modernity destabilizes traditional sources of meaning (religion, class, community), compelling individuals to create continuity and coherence amid rapid societal change.

Giddens highlights that in contemporary society, the self becomes a *reflexive project* - something individuals must actively construct and maintain through ongoing choices about lifestyle, relationships, and beliefs. Identity is no longer inherited or fixed but must be continually “worked on” as part of a biographical narrative that integrates past experiences and future aspirations. This process is both liberating and anxiety-producing: while modernity expands personal freedom and possibility, it also increases existential uncertainty and risk. Individuals must rely on abstract systems of trust – in institutions, technologies, and expert knowledge – whose functioning they cannot personally verify.

---

<sup>6</sup>Giddens, Anthony, *Modernity and Self-Identity: Self and Society in the Late Modern Age* (Polity Press, 1991).

From a broader societal perspective, Giddens interprets modernity as a *runaway world* driven by global interconnection, technological acceleration, and institutional transformation. The reflexivity that defines personal identity also shapes social change: societies continuously reconstruct themselves through self-examination and feedback loops. This creates an ongoing tension between stability and transformation, and drives the “dynamism of modernity” in Giddens’ terms. Social change thus emerges not as a series of discrete revolutions but as a continuous, self-reinforcing process of re-evaluation and adaptation.

This reflexive interconnection is especially potent in the age of AI. When an individual learns to work alongside AI tools, they change their professional identity and competencies – this is the individual level. That individual change aggregates as organizations adopt AI systems, reconfiguring workflows, roles, and organizational structures. These organizational changes feed into broader societal changes in the form of economic shifts, policy debates, and cultural narratives about technology’s role. But the feedback loop works in reverse too: societal anxieties about AI influence organizational risk management, which constrains how individuals are permitted to experiment with AI. Understanding these interconnected levels helps explain why AI adoption is both exciting and fraught – it demands reflexive adjustment at every level simultaneously, echoing the three-level model introduced earlier in the book, with each level shaping and being shaped by the others in an ongoing, dynamic process.

## **A Shifting Post-COVID Zeitgeist: Instability, AI, and Global Risks**

The societal zeitgeist has unmistakably shifted in the post-COVID era. Where the late 20th century was marked by optimism about globalization and steady progress, the 2020s are characterized by a pervasive sense of instability and doubt. Multiple overlapping crises have heightened collective anxiety. The World Economic Forum’s Global Risks Report in recent years

reflects this changed mood: environmental threats like climate change consistently rank as top long-term risks, while in the nearer term issues such as widespread mis/disinformation and state-based conflicts have surged to prominence. The COVID-19 pandemic itself acted as a stress test on global systems, exposing fragile healthcare capacity, disrupting economies, and necessitating draconian government interventions. In its aftermath, analysts observed a set of fundamental complications: accelerating economic disruptions, a reinforcement of nationalism and polarization, deepening inequalities, and a strain on governance capacity alongside the failure of international coordination. In short, the pandemic accelerated many pre-existing trends and made clear that society was less in control of events than previously assumed.

The deepening inequalities represent a particularly troubling trend. Thomas Piketty's influential research<sup>7</sup> has demonstrated how wealth concentration has returned to levels comparable to the early 20th century. His analysis of long-term income and wealth distribution data reveals that the post-war reduction in inequality was exceptional – driven by catastrophes like the Great Depression and World Wars that destroyed accumulated wealth – rather than the norm. In the absence of policies to constrain capital's tendency to outpace income growth, inequality widens. Piketty's work shows that when the return on capital exceeds economic growth, wealth accumulates exponentially among those who already have it. The implications are sobering: without deliberate intervention, democracies may become dominated by oligarchic interests, eroding social mobility and fracturing the social contract.

At the same time, the AI revolution has leapt forward. Powerful generative AIs are now writing code, drafting reports, and influencing decision-making processes. This has prompted excitement about productivity and innovation, but also fears about job displacement, algorithmic bias, and even existential risks from advanced AI. The pace of AI deployment contributes to a general atmosphere of speed and uncertainty – as if the ground is shifting beneath society's feet. Moore's Law and related exponential trends

---

<sup>7</sup>Piketty, Thomas, *Capital in the Twenty-First Century*, trans. Arthur Goldhammer (Harvard University Press, 2014).

continue (e.g., the cost of genome sequencing plummeted far beyond expectations), creating a “great acceleration” in technology. For many, keeping up with this rate of change is daunting.

Geopolitical instability adds another layer. The Russian invasion of Ukraine in 2022 marked the end of assumptions that economic interdependence alone (“Wandel durch Handel” or “change through trade”) would guarantee peace. In Europe this was seen as a “Zeitenwende”, a historic turning point, leading to decoupling from Russian energy and a re-emphasis on hard security. U.S.-China strategic rivalry has also intensified, raising uncertainties about the future of globalization. Meanwhile, multilateral institutions that underpinned the post-Cold War order from the UN to the WTO have been strained or undermined by the second Trump administration. There is a weakening of international cooperation at the very moment when global problems like pandemics, climate change, or AI governance demand collective action.

Socially and culturally, many societies report a widespread sense of lost control and declining trust. A “return to tribalism” is evident in the rise of identity politics and fragmentation of the public sphere. Traditional broad-based political parties are losing ground to more narrow, populist or issue-specific movements. Short-term interests often trump long-term thinking in both politics and business, as immediate survival instincts kick in at the expense of future risks. The post-Cold War narrative of inevitable liberal democratic expansion has faltered, giving way to what some analysts call a “post-postmodern” transition – essentially an end to the late-20th-century emancipatory project that championed universal rights, multilateralism, and sustainability.<sup>8</sup> The values and aspirations of that era seem no longer tenable in the same form, necessitating a redefinition of fundamental pillars, values, and priorities for what might become a new phase of modernity.

Yet there is a dissenting view. Steven Pinker’s *Enlightenment Now*<sup>9</sup> marshals extensive data to argue that by virtually every metric of human

---

<sup>8</sup>Blühdorn, Ingolfuhr, *Unhaltbarkeit: Auf Dem Weg in Eine Andere Moderne* (Suhrkamp, 2024).

<sup>9</sup>Pinker, Steven, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (Penguin Books, 2018).

welfare - from life expectancy to literacy, from poverty reduction to violence – we are living in the best of times. He suggests that the pervasive pessimism about modernity may stem from cognitive biases (like negativity bias and availability heuristic) rather than actual trends. According to Pinker, the institutions and values of the Enlightenment, i.e. reason, science, and humanism, have in fact produced unprecedented improvements in human wellbeing. The challenge, he argues, is to recognize these gains without becoming complacent about remaining threats like climate change or nuclear proliferation. While Pinker’s optimistic narrative offers an important counterweight to doomsaying, it may underestimate the scale of current crises and the nonlinear dynamics that could produce abrupt system failures. The truth likely lies somewhere between these perspectives: we have made remarkable progress on many fronts, yet face genuinely novel challenges that may require fundamental paradigm shifts that current institutions are ill-equipped to navigate.

## **The Limits of Linear, Technocratic Models of Change**

For much of the 20th century, societal progress was commonly viewed through a linear and technocratic lens – a belief that with the right expertise, policies, and continuous economic growth, societies would steadily advance. This paradigm assumed a degree of predictability and control: planners could project future needs, engineers could devise solutions, and economies would grow their way out of social problems. Economic growth (especially GDP growth) became a proxy for success, underpinning a largely technocratic, growth-centric worldview. However, in the face of today’s complex challenges, these traditional models are showing their age and limitations.

One major shortcoming is the assumption of linearity in problems that are fundamentally nonlinear. For example, classical economic and policy models often assume incremental change and gradual improvement. But consider environmental sustainability: many policy frameworks have banked

## *The Limits of Linear, Technocratic Models of Change*

on decoupling economic growth from environmental impact, hoping for radical efficiency gains or negative-emission technologies to deliver “net-zero emissions by 2050” without altering the growth model. In practice, such absolute decoupling at the required scale has no historical precedent. Models used in climate policy have built-in heroic assumptions (like massive carbon capture in the future) that only work with exponential innovation, scaling, and cost-degression. Another inadequacy is the “one-size-fits-all” technocratic approach to social change. Traditional models often emphasized top-down planning, standardized metrics, and universal solutions. But in a world of great complexity and cultural diversity, such approaches risk being too rigid and ultimately not effective.

James C. Scott’s influential work *Seeing Like a State*<sup>10</sup> reveals how ambitious, state-led modernization schemes often fail precisely because they impose simplified, uniform models that ignore local knowledge and context. Scott chronicles how high-modernist planning from Soviet agriculture to Brazilian Brasília repeatedly produced tragic outcomes when technocrats attempted to organize society according to abstract, rational plans that displaced practical, local expertise. The parallels to today’s sustainability frameworks are striking: when corporate and governmental efforts around sustainability and social responsibility boil down to checklist compliance and metric reporting (e.g., generic ESG scores), they risk this same mistake. As an example, sustainability reporting frameworks in the EU (like the CSRD and Taxonomy) impose extensive uniform indicators that are often not relevant for specific industry contexts. This technocratic compliance mindset can create an illusion of progress while businesses treat it as a box-ticking exercise, potentially resulting in little real-world impact. The bureaucratic burden may even divert resources away from substantive innovation. In short, a purely technocratic model can quickly become detached from local realities and dissociated from purpose.

Furthermore, linear models focused on economic growth often ignore issues of equity, purpose and wellbeing. As economist Tim Jackson argued

---

<sup>10</sup>Scott, James C., *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (Yale University Press, 1998).

in *Prosperity Without Growth*<sup>11</sup>, simply chasing GDP can lead to neither environmental sustainability nor social justice. There is a growing realization that “progress” needs rethinking. The dilemma, as Jackson notes, is that while the current growth-focused paradigm is unsustainable, visions for alternative prosperity (e.g., a post-growth economy) still seem politically and culturally out of reach. This conundrum leaves a strategic void: societies are clinging to a model that doesn’t solve today’s problems, yet they have not fully embraced a new model and fear radical innovation.

Kate Raworth’s influential *Doughnut Economics*<sup>12</sup> offers perhaps the most compelling alternative framework. She proposes replacing the obsession with perpetual GDP growth with a “doughnut” model that defines a safe and just space for humanity: a social foundation (below which people fall short of life’s essentials) and an ecological ceiling (beyond which Earth’s life-support systems degrade). The goal is to stay within these boundaries by meeting everyone’s basic needs without overshooting planetary boundaries. This framework acknowledges that in wealthy countries, economic growth may no longer be correlated with wellbeing, and that the real challenge is distribution rather than perpetual expansion. The doughnut model helps visualize the imperative: pursue regenerative and distributive economic designs rather than extractive and degenerative ones. The model also encourages us to reflect on the deeper meaning of wellbeing.

Compounding these issues is the rise in uncertainty and complexity in the social environment, which outpaces the capability of traditional linear planning. Sociologists point to phenomena like the individualization of society and the decline of broad collective identities. Andreas Reckwitz’s analysis of the “society of singularities” describes a fundamental shift in which individuals increasingly pursue unique, authentic identities rather than conforming to standardized roles.<sup>13</sup> People increasingly pursue self-actualization on their own terms, which makes social behavior harder to

---

<sup>11</sup>Jackson, Tim, *Prosperity Without Growth: Foundations for the Economy of Tomorrow*, Second (Routledge, 2017).

<sup>12</sup>Raworth, Kate, *Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist* (Chelsea Green Publishing, 2017).

<sup>13</sup>Reckwitz, Andreas, *The Society of Singularities*, trans. Valentin Schweitzer (Polity Press, 2020).

predict or shape top-down. Political volatility – with emotionalized movements, “culture wars,” and the erosion of long-standing party systems – creates nonlinear swings in policy and public opinion. In such a context, simplistic recipes fail. Technocratic elites offering “rational” solutions may find they lack public buy-in, as segments of society seek simpler narratives or scapegoats in the face of complexity. Indeed, a backlash against technocracy is evident in many countries, fueled by a feeling that expert-driven globalization neglected local needs and cultural values.

As linear and technocratic models of societal change are increasingly inadequate, the next section examines how societal change actually tends to occur: not as a smooth, managed process, but through nonlinear transitions that are frequently messy, path-dependent, and triggered by crises.

## **Non-Linear and Path-Dependent Societal Transition**

Historical and contemporary evidence suggests that societal transitions rarely follow a gentle, linear trajectory. Instead, change often comes in surges and ruptures, periods of relative stability punctuated by rapid, nonlinear shifts or tipping points. This pattern echoes the concept of punctuated equilibrium from evolutionary theory. Several key attributes characterize these transitions: nonlinearity, path-dependence, and the frequent role of crises as triggers.

Non-linearity means that small initial changes can lead to disproportionately large effects under certain conditions – a phenomenon colloquially illustrated by the “butterfly effect”.<sup>14</sup> In complex systems (societies, economies, ecosystems), feedback loops can amplify changes so that a stable system tips into a radically different new state. For example, ideas in social systems or technologies in economies can go from niche

---

<sup>14</sup>Lorenz, Edward N., “Predictability: Does the Flap of a Butterfly’s Wings in Brazil Set Off a Tornado in Texas?” *Meeting of the American Association for the Advancement of Science* (Washington, DC), 1972.

## *Societal Change in the Age of AI*

to dominant in a short time when nonlinear dynamics take over (consider the exponential adoption of smartphones or social media within a decade). Leaders must therefore beware of extrapolating linearly from current trends; the underlying dynamics may be poised for a geometric progression or an abrupt phase shift.

One driver of nonlinear societal shifts is the crossing of critical thresholds in systems, often due to loss of resilience. Insights from ecology are instructive: studies of lakes, coral reefs, and forests have shown that gradual environmental change can lead to sudden, drastic switches in the ecosystem's state once a threshold is crossed.<sup>15</sup> For example, a lake can shift abruptly from clear to turbid water after accumulating nutrients slowly over time once vegetation can no longer keep the water clear, a tipping point is reached and the system flips. This concept of ecosystem equilibrium states is illustrated in the following figure.

Such path-dependence is a critical aspect of transitions. Path-dependence means that history matters: the sequence of events and decisions can lock in certain trajectories that become hard to alter. Once a society embarks on a path (say, a fossil-fuel-based economy, or a particular constitutional order), positive feedbacks and vested interests can create inertia that reinforces that path. The concept of alternative stable states in complex systems illustrates how a system can have multiple possible equilibria, and the one it ends up in depends on the path taken and the shocks encountered.

Crucially, such shifts are often irreversible or very hard to reverse, a phenomenon known as hysteresis.<sup>17</sup> An example of hysteresis can be found in the Greenland Ice Sheet: If the entire 2,850,000 cubic kilometres of ice were to melt, it would lead to a global sea level rise of 7.2m, although this is expected to take millennia to fully play out.<sup>18</sup> This hysteresis effect is

---

<sup>15</sup>Scheffer, Marten, *Critical Transitions in Nature and Society* (Princeton University Press, 2009).

<sup>16</sup>Scheffer, Marten et al., “Catastrophic Shifts in Ecosystems,” *Nature* 413 (2001): 591–96.

<sup>17</sup>Haken, *Synergetics*.

<sup>18</sup>Nordhaus, William D., *The Climate Casino: Risk, Uncertainty, and Economics for a Warming World* (Yale University Press, 2013).

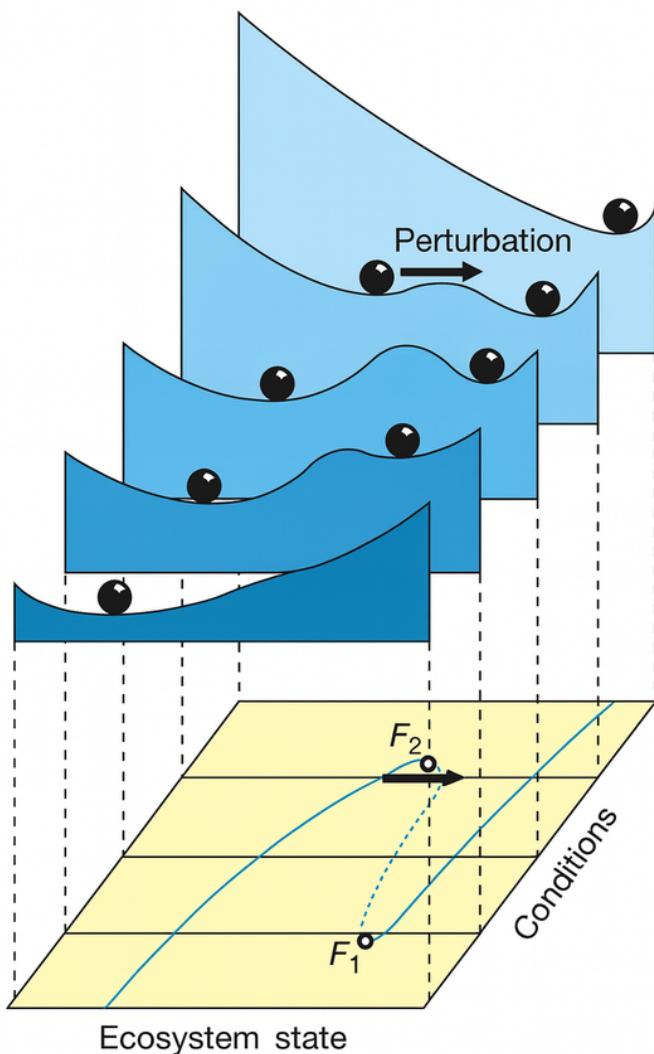


Figure 29: Ecosystem equilibrium states vary with conditions<sup>16</sup>

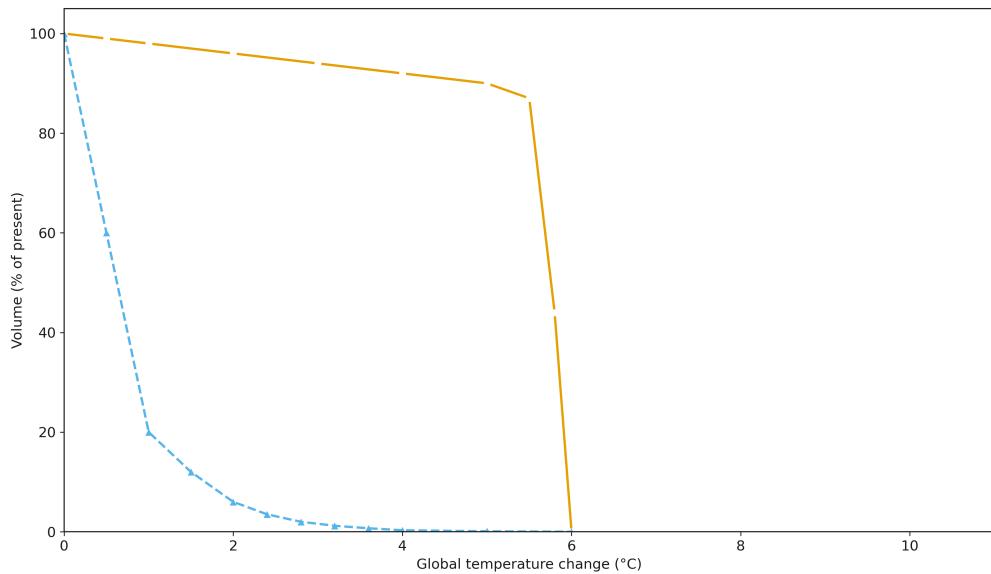


Figure 30: Hysteresis in the Greenland Ice Sheet

shown in the following figure.

The lesson for societies is that stresses like inequality, climate change, or technological disruption may build gradually but eventually push us into a new equilibrium (or disequilibrium) quite suddenly. Often it is the loss of resilience, i.e. the ability to absorb disturbances, that paves the way for a switch to an alternative state. By the time a societal system appears to “snap,” the groundwork for that snap was laid by long-term erosion of buffers and adaptive capacity. It underlines the importance of monitoring leading indicators and system resilience, not just visible outcomes, if we hope to foresee tipping points.

Recognizing path-dependence helps leaders understand why timing and sequencing of reforms are crucial and why solutions that worked in one context may not simply transfer to another. It also highlights the need, when trying to change course, to sometimes enact fundamental shifts (what former Intel CEO Andy Grove called “strategic inflection points”) rather than incremental tweaks, because incrementalism may just follow the es-

tablished path into decline.

The work of David Graeber and David Wengrow in *The Dawn of Everything*<sup>19</sup> offers a fascinating reframing of societal change by challenging the linear narrative of human progress from “simple” to “complex” societies. Drawing on archaeology and anthropology, they argue that early humans experimented with remarkably diverse forms of social organization – from hunter-gatherer egalitarianism to large-scale urban civilizations – and that the path to hierarchy was neither inevitable nor uniform. Societies could and did oscillate between different organizational forms: groups might be egalitarian during hunting season but hierarchical during trading expeditions; some cities flourished without rulers; some “barbarian” societies had more sophisticated democratic institutions than “civilized” ones. This insight deeply challenges path-dependence narratives that assume societies evolve along a fixed trajectory.

This suggests that radical change and even complete reversal of social structures have always been possible and has occurred throughout history. The implication for modern leaders is profound: rather than being locked into an inevitable path toward inequality or environmental degradation, societies have agency to experiment with fundamentally different arrangements. This view emboldens us to consider radical departures from current trajectories, which is much needed in an era requiring transformative responses to climate change, inequality, and technological disruption.

Often, crises act as the catalysts or triggers for major societal transitions. Indeed, many large-scale changes have been crisis-driven or at least crisis-catalyzed. Economic depressions, wars, pandemics, or environmental disasters have a way of unfreezing the status quo. They create a sense of urgency, sweep aside complacency or resistance, and open up political space for radical change. As the saying goes, “never let a serious crisis go to waste.” We saw this in 2008, when a financial meltdown led central banks to adopt unorthodox policies overnight, or in 2020, when a pandemic led governments to pursue emergency income support, eviction moratoria, and vaccine development at unprecedented speed. The key is that during a crisis, the

---

<sup>19</sup>Graeber, David, and David Wengrow, *The Dawn of Everything: A New History of Humanity* (Farrar, Straus; Giroux, 2021).

## *Societal Change in the Age of AI*

range of acceptable action widens dramatically. In a true existential crisis, survival priorities override normal politics.

However, not all crises automatically lead to positive change; they can also lead to collapse if mismanaged. Jared Diamond's study of past societies identifies five drivers of societal collapse:<sup>20</sup> 1. Natural climate changes 2. Self-inflicted environmental damage 3. Loss of support from friendly trade partners 4. The arrival of hostile outsiders 5. How a society responds to its problems

Notably, the first four are external stresses, but the fifth is internal, where the crucial variables are foresight and adaptability. Diamond's work underscores that while crisis factors might be out of our control (e.g. volcanic eruptions or foreign invasions), it is ultimately how society anticipates and reacts that determines collapse or survival. In modern times, we can see analogies: climate change, pandemics, global economic swings, and geopolitical conflicts test societies. Those that handle these tests through foresight, social cohesion, and effective response can emerge transformed but intact (or even stronger); those that respond poorly may experience breakdowns. History provides many examples of crisis-driven transitions, which tend to be neither gradual nor reversible; they were phase changes in the system of society.

From these observations, leaders can derive several heuristics for managing crisis-driven change.

1. First, in crisis, quick and decisive action is often necessary to ensure basic survival. For businesses this might mean focus on cashflow and balance sheet strength, for governments it might mean emergency measures to keep people safe. In truly existential crises, the usual boundaries between sectors blur and governments may need to partner with businesses or vice versa in unconventional ways.
2. Second, crises demand narrative leadership: developing stories of hope and a positive future to galvanize people is crucial. In such

---

<sup>20</sup>Diamond, *Collapse*.

crises, it can even help to “make the problem bigger”, i.e. contextualize the crisis as part of a larger challenge that people can unite to overcome, thereby turning panic into collective purpose. Communication, transparency, and trust-building are paramount in a crisis, because trust tends to erode by default if leaders go silent.

3. Lastly, once the immediate shock is past, there is a narrow window to institute reforms that build long-term resilience (before the sense of urgency fades).

Related to the nonlinear journeys of societal transitions are the concepts of system fragility versus antifragility, which further illuminates why some systems collapse under volatility while others thrive on disorder.

## **Fragility, Pseudo-Stability, and Antifragility**

Modern societies and organizations often prize stability, predictability, and control. However, as Nassim Nicholas Taleb shows in his works on risk and uncertainty, attempting to eliminate volatility can ironically make a system more fragile.<sup>21</sup> He introduced the concepts of convex and concave systems to better understand what drives fragility, as illustrated in the following figure.

A fragile system is one that suffers when exposed to volatility, randomness, or shocks. Its performance or integrity degrades rapidly under stress. Think of an ornate glass vase – perfect under steady conditions, but it shatters with the slightest knock. For the fragile, shocks bring higher harm as their intensity increases. In such a fragile system, the cumulative effect of small shocks is smaller than the single effect of an equivalent single large shock. The more concave an exposure, the more harm from the unexpected, and disproportionately so. This can be illustrated by the impact speed has on the damage to a car driving against an obstacle, as shown in the following figure.

---

<sup>21</sup>Taleb, *Antifragile*.

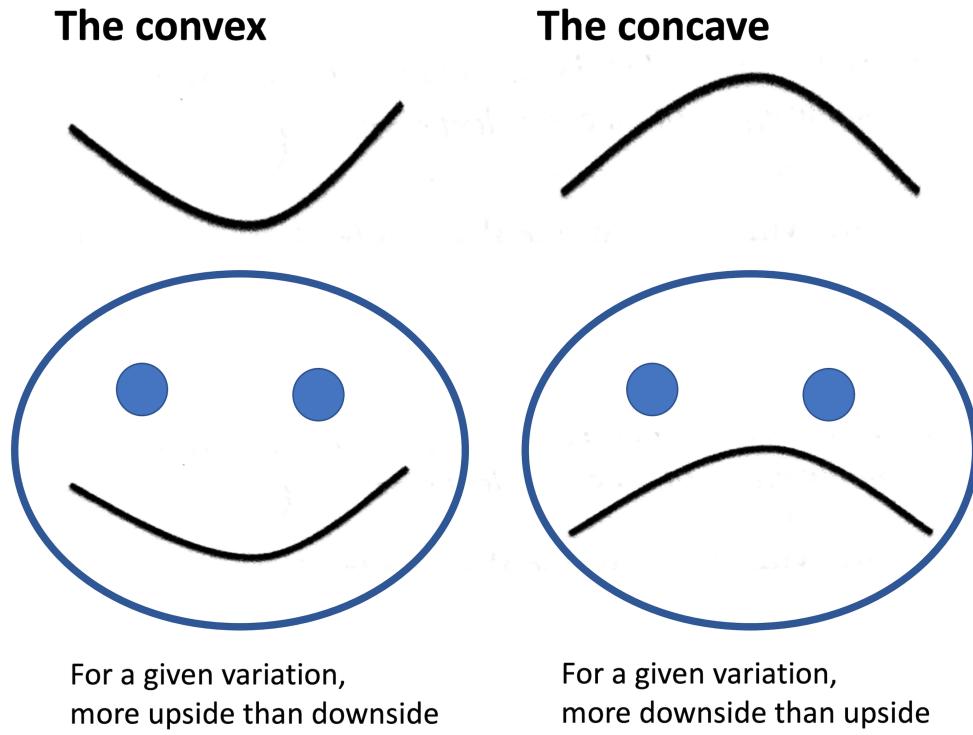


Figure 31: Two essential types of nonlinearity

By contrast, a robust system might be like a plastic cup, which can withstand knocks without breaking, though it doesn't improve from them. Beyond robustness, antifragility describes systems that actually benefit from shocks, as opposed to fragile systems that break under stress and merely robust systems that resist shock without improving. Antifragile systems strengthen or improve when exposed to stresses or variability (up to a point). A classic example is the human immune system or muscular system, which grows stronger when challenged by germs or exercise (but will atrophy if not challenged at all).

Taleb points out that nature is antifragile up to a point, but that threshold is quite high. That is why it helps to look towards nature for characteristics of antifragile systems as is illustrated in the following table.

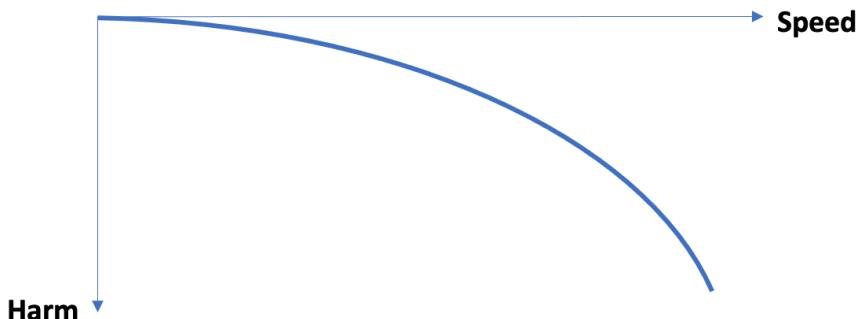


Figure 32: The concave phenomenon of driving a car against an obstacle

Table 4: Learning from nature

The mechanical, noncomplex, nonbiological	The organic, complex, biological
Needs continuous repair and maintenance	Self-healing
Hates randomness	Loves variations (up to a point)
No need for recovery	Needs recovery between stressors
No or little interdependence	High degree of interdependence
Stressors cause material fatigue	Absence of stressors cause atrophy
Age with use (wear and tear)	Age with disuse
Undercompensates from shocks	Overcompensates from shocks
Equilibrium in a state of inertia	Equilibrium only with death

The fundamental insight here is that nature favors diversity between organisms rather than diversity within a single immortal organism or organization. This principle reveals a profound evolutionary logic: adaptability and

## *Societal Change in the Age of AI*

innovation emerge more reliably from a population of distinct, competing entities than from the internal variation of one perpetual entity. When organisms are mortal and numerous, each generation can experiment with different solutions to environmental challenges. Successful adaptations proliferate through reproduction, while unsuccessful ones disappear with the death of their bearers. This process of selection acts on the population level, constantly refreshing the gene pool and allowing rapid responses to changing conditions.

In contrast, an immortal organism or organization faces inherent constraints on adaptation. Its internal diversity is limited by the need to maintain coherent functioning – too much variation within a single entity risks incoherence or collapse. Moreover, an immortal entity accumulates structural rigidities over time, as successful patterns become deeply embedded and resistant to change. The organization’s very longevity becomes a liability, as legacy systems, established hierarchies, and sunk costs create powerful forces for stasis. This notion clearly aligns with Joseph Schumpeter’s appreciation of the creative destruction inherent in market economies. Schumpeter understood that economic progress depends not on the perfection of existing firms, but on the constant cycle of new enterprises displacing old ones<sup>22</sup>. Just as biological evolution proceeds through the birth and death of organisms, economic innovation advances through the entry of new companies with fresh ideas and the exit of incumbents that have outlived their usefulness. The market’s vitality comes from this turnover – from diversity between competitors rather than the eternal optimization of monopolies. Both natural ecosystems and healthy economies derive their dynamism from the same source: a population-level process where mortality creates space for novelty, and competition ensures that better-adapted entities flourish while obsolete ones fade away.

One counterintuitive insight is that efforts to create highly stable, controlled conditions can lead to pseudo-stabilization that breeds fragility. A story from 19th-century engineering illustrates this: James Clerk Maxwell, in an 1867 paper *On Governors*, mathematically proved that overly tight

---

<sup>22</sup>Schumpeter, Joseph A., *Capitalism, Socialism and Democracy* (Harper & Brothers, 1942).

feedback controls on steam engine speed could actually make the system unstable<sup>23</sup>. In general, artificially suppressing all variability in a system tends to hide underlying risks and erode the system's adaptive capacity. Small perturbations that would have revealed weaknesses (and allowed learning or adaptation) are prevented, so vulnerabilities accumulate out of sight. When a shock eventually exceeds the controlled range, the system breaks catastrophically because it has no recent experience of shocks and no tolerance built up. Taleb and others have generalized this principle across domains: volatility is information. If you suppress volatility, you also suppress the signals that would normally tell you how close to the edge of disaster you are. Pseudo-stability, therefore, is a dangerous illusion – it is the calm before the storm, the smooth ice hiding thin patches.

There are many real-world examples of pseudo-stabilization and its consequences. - Forest management for much of the 20th century emphasized suppressing wildfires at all costs; the result was a build-up of fuel (dry brush and dense growth) that eventually led to massive, uncontrollable fires. The small fires that would naturally clear out underbrush were prevented (stability in the short term), but this created conditions for truly devastating fires later.<sup>24</sup> - Another example is in parenting or education: overprotective “helicopter parents” who remove every hardship or random challenge from a child’s life may end up with a child who is ill-equipped to handle adversity and a fragile individual.<sup>25</sup> - In economic policy, some analysts have argued that central banks’ long-standing practice of smoothing out the business cycle (e.g., quickly bailing out markets or lowering interest rates at the slightest sign of trouble) can create financial pseudo-stability. It might encourage excessive risk-taking and debt accumulation (moral hazard), leading to a much bigger crash down the line (e.g., the

---

<sup>23</sup>Maxwell, James Clerk, “On Governors,” *Proceedings of the Royal Society London* 16 (1867): 270–83.

<sup>24</sup>Kreider, Mark R. et al., “Fire Suppression Makes Wildfires More Severe and Accentuates Impacts of Climate Change and Fuel Accumulation,” *Nature Communications* 15, no. 2412 (2024), <https://www.nature.com/articles/s41467-024-46702-0>.

<sup>25</sup>Haidt, Jonathan, and Greg Lukianoff, *The Coddling of the American Mind: How Good Intentions and Bad Ideas Are Setting up a Generation for Failure* (Penguin Press, 2018).

## *Societal Change in the Age of AI*

2008 financial crisis after a period dubbed “the Great Moderation”).<sup>26</sup> Former Federal Reserve chairman Alan Greenspan’s strategy of aggressively countering recessions was sometimes likened to taking volatility out of the system by “ironing out the boom-bust cycle”, but critics say it sowed the seeds of fragility.<sup>27</sup>

These cases illustrate Taleb’s assertion: artificially constrained systems become less stable and may end up prone to Black Swans, meaning rare, yet extreme events.

What, then, makes a system antifragile or at least less fragile? Several factors can be identified from Taleb’s work and related analyses:

- *Decentralization and Diversity*: Fragile systems tend to be overly centralized or homogenized, whereas antifragile ones have diverse components and decentralized decision-making. Diversity allows parts of the system to fail without collapsing the whole, and successful adaptations can spread. (E.g., an ecosystem with diverse species is less fragile than a monoculture crop; an economy with many small banks may weather shocks better than one dominated by a few “too big to fail” institutions.)
- *Redundancy and Slack*: Efficiency is often the enemy of antifragility. Fragile systems are optimized for peak efficiency with no fat, while antifragile systems build in buffers – extra capacity, inventory, financial reserves, etc. Redundancy and buffers provides breathing room or a “margin of safety” when a shock hits.
- *Skin in the Game*: Taleb emphasizes that systems where actors have skin in the game (i.e., they personally bear the consequences of their decisions) tend to be more antifragile.<sup>28</sup> It aligns incentives with prudent risk-taking and learning. In fragile systems, one often finds perverse incentives (e.g., privatized gains and socialized losses) that encourage reckless behavior and fragility.
- *Interdisciplinary Thinking and Adaptability*: Siloed, highly specialized thinking can make systems brittle, as they miss

---

<sup>26</sup>Schularick, Moritz, and Alan M. Taylor, “Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008,” *American Economic Review* 102, no. 2 (2012): 1029–61.

<sup>27</sup>Grimm, Maximilian et al., *Loose Monetary Policy and Financial Instability*, Working Paper 30958 (National Bureau of Economic Research, 2023).

<sup>28</sup>Taleb, Nassim Nicholas, *Skin in the Game: Hidden Asymmetries in Daily Life* (Random House, 2018).

interconnections. Antifragile systems encourage curiosity, interdisciplinary learning, and adaptability. They value wisdom and flexibility over rigid expertise. This means being mentally ready to adjust to worst-case scenarios and preserving optionality to respond to unexpected events.

- *Experimentation and Feedback:* Antifragility grows from lots of small trials and errors – what Taleb calls “stochastic tinkering.” A fragile mindset shuns failure; an antifragile one treats failure as information and uses feedback loops to get stronger. For example, the scientific method, entrepreneurship, and evolution itself all advance via iterative failures that improve the system.

These principles can be summarized in a simple comparison:

Table 5: Fragility drivers versus antifragility drivers

Fragility Drivers	Antifragility Drivers
<i>Over-specialization</i> and silo thinking – narrow focus	<i>Diversity</i> of approaches, <i>interdisciplinarity</i> , and cross-domain knowledge
Rigid <i>bureaucracy</i> and central control – slow to adapt	Decentralized, <i>entrepreneurial</i> mindset – quick to adjust and innovate
Optimization for short-term <i>success</i> – no buffers, fear of any loss	Built-in <i>redundancy</i> and slack – accept short-term inefficiencies for long-term resilience
<i>Privatization of gains</i> & socialization of losses – misaligned incentives	<i>Skin in the game</i> – decision-makers share downside risk, encouraging prudent behavior
Reliance on <i>predictive models</i> and expert overconfidence	Embrace of <i>optionality</i> and scenario planning – prepare for multiple outcomes
Suppression of volatility (“stability at all costs”)	Acceptance of <i>small failures</i> and volatility as information – system learns and evolves

In practice, moving a system (be it an organization, an economy, or a soci-

ety) from fragility toward antifragility involves encouraging the factors on the right side of the table. For example, a national economy can reduce fragility by preventing any one sector or company from becoming “too dominant to fail,” by encouraging competition and diversity in industry, and by maintaining regulatory buffers like adequate capital requirements in banking (so banks can absorb losses). It can encourage innovation via many startups (accepting that many will fail, but a few will succeed spectacularly in a positive asymmetry). Governments and firms can also create contingency plans and war-gaming exercises that simulate crises, giving them practice in adaptation.

A particularly important concept for leaders is avoiding pseudo-stability. Taleb warns: “Avoiding fragility is not just about responding to crises, but about not giving in to the temptation of too much stability in the first place.” If we build systems that only thrive in calm conditions, we are courting disaster. Instead, leaders should periodically inject stressors or tests into systems – whether that’s a controlled drill for a cyberattack, a financial stress test, or rotating people through different roles to prevent over-specialization. These practices keep the system agile and reveal weaknesses while they are minor. In national security, for instance, militaries conduct regular exercises and “red team” simulations to ensure readiness.

Fragility versus antifragility is a core lens for understanding societal robustness. The age of AI, with its rapid changes and potential shocks from superintelligence or AGI (Artificial General Intelligence), will punish the fragile and reward the antifragile. Paradoxically, even as our technological knowledge increases, the world in some ways becomes less predictable. The prudent approach, therefore, is to design society and institutions in such a way that they can absorb or even benefit from volatility.

## **Values, Worldviews, and Meaning-Making in Societal Change**

While technology and tools are crucial, ultimately societal change is a human process rooted in values, worldviews, and collective meaning-making. Changes in law or technology alone will not stick unless they resonate with people's deeper beliefs and cultural narratives. As the futurist Shlomo Shoham argues, solving our long-term challenges requires elevating our *Future Intelligence* – essentially a moral and cognitive capacity to care about and plan for the future of humanity and the planet. In this section, we delve into how worldviews and meaning frameworks can enable or block change, drawing on Ken Wilber's Integral Theory<sup>29</sup> and other developmental perspectives, and consider how leaders might nurture the value shifts necessary for a more resilient, sustainable society.

The concept of developmental stages, which we have explored for individuals and organizations in past chapters, can also be applied to societies. Societies can be seen as moving through value paradigms, which are often referenced via Spiral Dynamics or Wilber's stages (e.g., from egocentric to ethnocentric to world-centric perspectives). World-centric (integrative) values are ones that extend care to all humans and the planet, and this stage is seen by many thinkers as necessary to address global challenges like climate change. However, many populations may still operate largely at ethnocentric (us-vs-them) or even egocentric levels of values, which can impede collective action.

This highlights a core tension: values mismatch. If our technological and institutional capabilities are at one level but our values are at a lower level (for instance, we have global impact technologies but parochial loyalties), we will struggle to solve global problems. A vivid example is climate change – scientifically and technologically we know how to mitigate it (through renewable energy, efficiency, circular economies etc.), but culturally and politically, many societies have not embraced the necessary level of global solidarity and long-term responsibility to implement those solutions.

---

<sup>29</sup>Wilber, *A Theory of Everything*.

## *Societal Change in the Age of AI*

Importantly, worldviews can block change when they are incommensurate with the needed direction. If a significant portion of society believes, for instance, that economic growth is inherently tied to virtue and any suggestion of limits is “anti-progress”, then policies aiming for sustainability or emission restrictions will face fierce resistance. Conversely, worldviews can enable change when aligned. The widespread adoption of environmental consciousness since the 1970s, for instance, has enabled the passage of environmental laws and the rise of green industries. However, as climate activist Gus Speth famously said, the biggest environmental challenges are not scientific or technological but rather “selfishness, apathy and narrow-mindedness” – essentially human traits. This underscores that inner change (values, ethics, empathy) is as crucial as outer change (technology, law).

Meaning-making frameworks and cultural narratives from religion and philosophy play a role here. In times of upheaval, people often turn to basic questions of meaning: Why are we doing this? What is the purpose of our society? Leaders who can articulate a compelling vision or narrative provide a kind of social glue that holds the transformation together. Think of Franklin D. Roosevelt’s narrative of the “Four Freedoms” during WWII or John F. Kennedy’s vision of reaching the moon that galvanized collective effort. In the context of AI and modern challenges, we might need new narratives – perhaps around the idea of a “Great Transition” or “Third Modernity” that redefines progress not just as material growth but as human and ecological thriving in a symbiosis with AI.

## **Implications for Institutions and Long-Term Resilience**

These insights translate into concrete implications for institutional design and leadership strategy, particularly at the intersection of sustainability and capitalism in the AI age. Traditional government institutions, many dating to the 19th or mid-20th century, were not built for exponential technological change or planetary risks. We need new institutional innovations.

## *Implications for Institutions and Long-Term Resilience*

- One approach is creating bodies that represent future generations or long-term planetary interests. The Israeli Commission for Future Generations under Shlomo Shoham's leadership was an early example: a parliamentary commissioner who could veto laws harming future citizens. However, the Commission is no longer active in current Israeli politics.<sup>30</sup>
- Another innovation is embracing participatory and deliberative processes (e.g., citizen assemblies, participatory budgeting) that enrich democratic decision-making. For complex, long-term issues like AI's impact on employment or climate transitions, randomly selected citizens' panels advised by experts have proven capable of finding common ground and thoughtful recommendations.
- Adaptive regulation is also key: rather than codifying rigid rules that quickly become outdated, regulators can use sunset clauses, iterative rulemaking (like the FCC's "sandboxes" for new tech), special economic zones, and outcome-based regulations that specify goals but not means. This aligns with antifragility by allowing variation and learning. For instance, instead of fixed rules for all AI systems, regulators might set performance or safety criteria and dynamically adjust them as technology evolves.
- Global challenges need global cooperation, yet formal multilateralism is struggling. We must either reform it or supplement it with flexible networks. One approach is networked multilateralism: coalitions of the willing, city networks (like C40 cities for climate action), public-private partnerships for specific goals (e.g., vaccine distribution). These fluid arrangements can act faster and innovate, though they may lack universal legitimacy. Nevertheless, traditional institutions like the UN remain crucial for setting norms. To improve them, inject future-oriented thinking and systems thinking into their processes. For example, scenario planning and stress-testing could be used in UN policy reviews (similar to central bank stress tests). International bodies could champion norms for AI and cyber governance

---

<sup>30</sup>Boston, Jonathan, *Governing the Future: Designing Democratic Institutions for a Better Tomorrow* (Emerald Group Publishing, 2016).

## *Societal Change in the Age of AI*

proactively, rather than playing catch-up. The involvement of NGOs and businesses in multilateral efforts should be institutionalized further, as they often have expertise or resources that governments lack.

- Resilience comes from social cohesion and trust as much as from strategy and steel. Communities with strong local networks and cultures of mutual aid bounce back faster from disasters. Strengthening civil society – clubs, associations, neighborhood groups, online communities of practice – is not a “soft” nice-to-have, but a core part of resilience.
- Antifragility thinking needs to be incorporated into planning. How could this system fail? Are we too optimized and vulnerable? Where can we give up efficiency for robustness? Concretely, governments might maintain strategic reserves (medical supplies, energy, etc.), diversify supply chains to avoid single points of failure, and invest in social safety nets that act as shock absorbers during crises.

All these changes require a new breed of leaders. The complexity of today’s challenges means leaders must be systems thinkers, ethically grounded, and comfortable with uncertainty. Leaders should have “strong opinions, weakly held” – a vision and direction to inspire and guide, but also humility and openness to change course when new information arrives. The days of the infallible, authoritarian CEO or head of state who never admits doubt are ill-suited to a fast-changing landscape. Instead, leaders might emulate a facilitator-in-chief: setting clear purpose and values, but empowering teams and networks to experiment and adapt strategies on the ground.

## **Practical Tools**

Navigating societal change in an era of high complexity requires not only new mental models but also new tools. This section explores how virtual polling, agent-based models, AI-driven simulations, and horizon scanning tools can support adaptive governance and large-scale coordination.

## Virtual polling

Large Language Models in their training condense an echo of the complex world and thus can be another novel tool for governance. Recent research indicates that LLMs, especially when fine-tuned on data about human decision-making, can simulate human-like responses and cognition. In one study, a model (LLaMa 70B fine-tuned with psychological experimental data) outperformed specialized systems in mirroring human cognitive patterns. This opens up the provocative idea of using AI as a kind of “public opinion simulator.” Instead of traditional polling (which is slow and can be limited in scope), governments or organizations might prompt an LLM to estimate how different demographic or ideological segments of the population would respond to a proposed policy or a piece of messaging. For instance, a Ministry could virtually “poll” a calibrated set of AI personas – say, simulating European adults from different nations – to gauge support for a new nuclear energy policy under various framings. The AI, drawing on patterns learned from vast text data, can generate responses that mimic likely human reactions. While this approach is still experimental, it shows promise as a “sentiment wind tunnel”: leaders can test how the public might react to a strategy (and even refine the messaging) before rolling it out in reality. GPT-4 predictions in simulated surveys were correlated with true effects as strongly as were the average expert forecasts.<sup>31</sup> The results are shown in the following figure.

It’s crucial, however, to be aware that the design of such virtual social science experiments can massively influence the outcomes.<sup>32</sup>

## Generative agent-based modeling

One promising approach is the use of agent-based modeling (ABM) and digital twins of social systems. The idea of agent-based modeling has been

---

<sup>31</sup>Hewitt, Luke et al., “Predicting Results of Social Science Experiments Using Large Language Models,” *Ethicalpsychology.com*, 2024.

<sup>32</sup>Cummins, James, “The Threat of Analytic Flexibility in Using Large Language Models to Simulate Human Data: A Call to Attention,” *arXiv Preprint*, 2025.

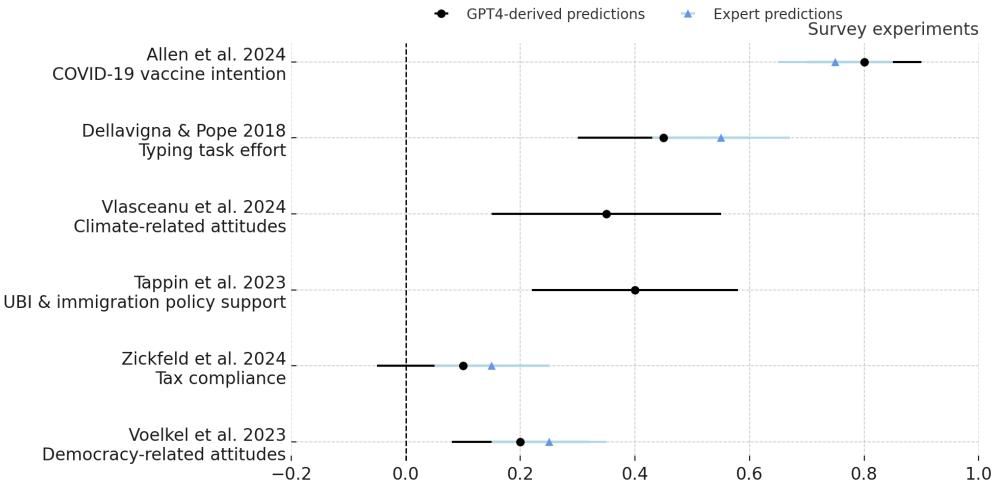


Figure 33: GPT-4 predictions in simulated surveys

around for many decades. Thomas C. Schelling's introduced in 1971 a segregation game<sup>33</sup> as a seminal model in complexity and social behavior. It demonstrates how individual preferences, even mild ones, can lead to strong collective segregation without any central coordination or discriminatory intent. In the segregation game Schelling represented a city as a grid of cells (like a checkerboard) where each cell is occupied by an individual from one of two groups (e.g., "red" or "blue") or left empty. Each individual has a preference for the fraction of neighbors who are similar to themselves, for example, wanting at least 30–40 % of their neighbors to be of the same color. If the proportion of similar neighbors is below the agent's threshold, the agent becomes unhappy and moves to another empty cell that meets its preference. This simple agent-based model showed that even when individuals are tolerant and willing to live with many unlike neighbors, the aggregate outcome often becomes highly segregated. The model reveals a nonlinear amplification between micro-level preferences and macro-level outcomes: small local biases cascade into large-scale patterning. In fact, you can easily "vibe-code" the segregation game today by asking ChatGPT

---

<sup>33</sup>Schelling, Thomas C., "Dynamic Models of Segregation," *Journal of Mathematical Sociology* 1, no. 2 (1971): 143–86.

or Claude to write code for Schelling’s model and run the simulation with parameters of your choice with a nice visualization of the dynamics.

With modern computing infrastructure and AI models, agents can be made much more sophisticated than the simple rule-based agents of the past and of conventional economic models. Agents can be equipped with distinct personas – as was shown in previous chapters – to simulate a large number of diverse people, households, firms, etc. In 2024 Ghaffarzadegan et al.<sup>34</sup> introduced the concept of Generative Agent-Based Modeling (GABM) as a hybrid framework that integrates traditional agent-based models with large language models to simulate complex social systems. By embedding generative AI into agents’ decision-making processes, GABM enables agents to reason, converse, and adapt in natural language. This shift transforms agents from rule-based entities into semi-autonomous actors capable of dynamic reflection, negotiation, and context-sensitive responses, bringing ABMs closer to real-world social complexity. For example, this approach can be used to study the diffusion of social norms within society. Agents in the model use LLM-based reasoning to interpret norms, justify compliance or resistance, and influence peers, leading to emergent social dynamics that are qualitatively richer than those produced by fixed behavioural algorithms. While this approach can generate novel insights into collective behaviour, policy response, and value change, it also introduces issues of validation, reproducibility, and prompt sensitivity<sup>35</sup>.

GABM represents a methodological frontier by bridging computational social science and generative AI to explore not just how societies behave, but how they think, evolve, and imagine futures. For example, Piao et al.<sup>36</sup> present **AgentSociety**, one of the most ambitious frameworks yet for simulating social systems composed of tens of thousands of large language model–driven agents. Each agent possesses a persistent memory, internal motivation, social identity, and the capacity for natural-language reasoning

---

<sup>34</sup>Ghaffarzadegan, Navid et al., “Generative Agent-Based Modeling: Unveiling Social System Dynamics Through Hybrid Simulations,” *Nature Human Behaviour*, 2024.

<sup>35</sup>Lu, Xinyi et al., “LLMs and Generative Agent-Based Models for Complex Social Dynamics,” *arXiv Preprint*, 2024.

<sup>36</sup>Piao, Yuchen et al., *AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviours and Society*, 2025.

and interaction. Within the simulation, agents live, work, form relationships, and respond to events – economic shifts, environmental crises, or policy changes – through text-based communication and adaptive decision-making. This architecture enables emergent phenomena such as polarization, cooperation, norm formation, and innovation to arise organically from bottom-up interactions rather than being preprogrammed by researchers. The scale and linguistic fluency of the system make it a powerful testbed for studying how **policy interventions and collective shocks** propagate through simulated populations. For instance, scenarios involving universal basic income, media influence, and natural disasters reveal how individual reasoning cascades into macro-level trends, such as wealth redistribution, public trust shifts, or collective resilience. Crucially, Piao et al. argue that such LLM-driven agents can serve as *cognitive micro-foundations* for computational social science, bridging mechanistic ABM and qualitative social reasoning. However, they also highlight key challenges – especially the validation of agent cognition, the control of emergent biases, and the heavy computational demands of maintaining coherent long-term memories at scale.

## Digital twins

The concept of a *digital twin* extends agent-based modeling to creating a near real-time replica of an actual system with detailed context data<sup>37</sup>. The roots of the digital twins can be found in engineering and manufacturing. For example, BMW's iFACTORY initiative demonstrates how virtual-first planning can transform complex production systems. The automaker built its Debrecen plant in Hungary entirely in virtual space before breaking physical ground. By scanning over 7 million square meters of production space across 30+ facilities and creating unified digital representations, BMW reduced planning costs by up to 30%. Critical processes like collision checks – verifying that new vehicle models fit existing production infrastruc-

---

<sup>37</sup>Office, U. S. Government Accountability, *Science & Tech Spotlight: Digital Twins—Virtual Models of People and Objects*, GAO-23-106453 (U.S. Government Accountability Office, 2023), <https://www.gao.gov/products/gao-23-106453>.

ture – now complete in three days through simulation rather than requiring four weeks of physical testing.<sup>38</sup>.

Cities like Singapore have taken this idea to the next level by developing digital twins of their urban infrastructure to simulate impacts of zoning changes or climate events<sup>39</sup>. We can imagine creating digital twins of national economies or even aspects of society (public opinion, demographic change) to test “what if” scenarios in silico. Indeed, the idea of policy wind-tunnel testing is becoming feasible – similar to how engineers test aircraft designs in wind tunnels, governments could test policies in virtual environments. For example, one can easily simulate the performance of alternative organizational designs for eldercare, by gathering baseline parameters, creating synthetic organizations with adjustable “design knobs,” and then running simulations to see how each design copes with shocks or varying conditions. Such experiments, aided by AI, can reveal the sensitivities of different strategies and help optimize design choices before implementing them in the real world.

## Horizon scanning and scenario planning

AI can be of tremendous aid in systematic horizon scanning and scenario planning.<sup>40</sup> Machine learning algorithms can sift through enormous amounts of data (news, social media, scientific literature) to detect weak signals and emerging trends that a human analyst might miss.<sup>41</sup> This can

---

<sup>38</sup>Georges, Gilles, *BMW's iFACTORY: A Case Study in Digital Twins in Manufacturing at Enterprise Scale*, Clarity Points, 2025, <https://claritypoints.com/digital-twins-in-manufacturing-enterprise-scale/>.

<sup>39</sup>Public Sector Innovation, OECD Observatory of, *Virtual Singapore – Singapore’s Virtual Twin*, Singapore Land Authority, 2015, <https://oecd-opsi.org/innovations/virtual-twin-singapore/>.

<sup>40</sup>Finkenstadt, Daniel J. et al., “Contingency Scenario Planning Using Generative AI,” *California Management Review*, 2024.

<sup>41</sup>Ha, Sungwha Hong] [Taehyun, Heyoung Yang, “Automated Weak Signal Detection and Prediction Using Keyword Network Clustering and Graph Convolutional Network,” *Futures*, ahead of print, Elsevier, 2023, <https://doi.org/10.1016/j.futures.2023.103202>.

## *Societal Change in the Age of AI*

improve our collective foresight.<sup>42</sup> Governments and large corporations have begun to use AI to augment their strategic planning departments, for example, by using natural language processing to monitor global sentiment or early indicators of instability in healthcare. Additionally, AI can help generate scenarios by rapidly combining trends in novel ways. Traditional scenario planning might create a handful of scenarios through expert workshops; AI can generate hundreds of scenario variants and even assign probabilities to them or identify which variables drive the most significant differences in outcomes.

The concept of “adaptive governance” is closely tied to using such tools. Adaptive governance means policies are not set in stone but are designed to be flexible, with feedback loops to adjust based on outcomes. Here, AI can function as a real-time feedback generator. A simple illustration is a city governance system that dynamically adjusts traffic light patterns based on live traffic AI analysis to reduce congestion. Beyond this case effective design of economic policy is still beyond the horizon today.

AI and simulation tools offer a way to cope with complexity rather than be overwhelmed by it. They can illuminate the intricate webs of cause-and-effect that are otherwise hard to parse, and they can test interventions in virtual space to avoid real-world failures. However, these tools are aids, not panaceas. Their effective use requires a governance mindset that is experimental, transparent, and inclusive. Leaders must pair AI’s predictive power with human values and judgment.

## **Conclusion: A More Political Modern Leadership Agenda**

In wrapping up this chapter, we return to the central insight: societal change in the age of AI is complex and often nonlinear, but not beyond

---

<sup>42</sup>Schmidt, Lena et al., “Horizon Scans Can Be Accelerated Using Novel Information Retrieval and Artificial Intelligence Tools,” *arXiv Preprint arXiv:2504.01627*, ahead of print, 2025, <https://doi.org/10.48550/arXiv.2504.01627>.

## *Conclusion: A More Political Modern Leadership Agenda*

influence. By shedding outdated linear mindsets and embracing an integral approach that accounts for technological, social, and meaning dimensions, leaders can guide change rather than be overwhelmed by it.

The major themes we explored each point to actionable heuristics:

- *Acknowledge the new zeitgeist*: Accept uncertainty as the new normal and plan for multiple futures. Build coalitions that reflect today's diverse voices (and discontents) to restore trust and legitimacy.
- *Update mental models*: Replace linear, siloed thinking with systems thinking. Use feedback loops, iterate policies, and remain agile.
- *Anticipate tipping points*: Monitor system health (resilience indicators) and be ready to act decisively when thresholds loom. Don't wait for crises, but when they come, seize the moment to implement bold change.
- *Avoid fragility*: Audit your organization or nation for points of fragility. Deliberately add redundancies and contingencies. Don't confuse no visible volatility with genuine stability – probe the system's true robustness.
- *Leverage AI and data*: Use the best tools to inform decisions, but govern their use wisely. Democratize these insights – share simulations and foresight with the public to build a common understanding of why changes are needed.
- *Lead with values and vision*: Recognize that at its heart, change is about people. Engage hearts and minds with a vision that appeals to shared values. Invest in education and dialogue to evolve worldviews toward greater empathy, inclusion, and responsibility.
- *Institutionalize long-term thinking*: Create structures that counter short-termism, from future generation commissioners to corporate purpose committees. Align incentives to long-term outcomes in finance and policy.

Overall, the job of the modern leader is becoming more political. Today's leaders must master the art of building coalitions across unprecedented divides, negotiating between competing worldviews, and creating shared meaning in fragmented landscapes. This is not politics as partisan maneuvering, but politics as the essential work of the polis: convening diverse stakeholders, facilitating difficult conversations, and forging collective action despite profound differences in values and interests.

Where twentieth-century leadership emphasized command, control, and expertise, twenty-first-century leadership demands skills more typically as-

## *Societal Change in the Age of AI*

sociated with statecraft: the ability to read complex stakeholder dynamics, to sense shifting power configurations, to broker compromises that preserve core values while enabling progress, and to communicate vision in ways that resonate across cultural and ideological boundaries. The leader can no longer simply announce the optimal technical solution; they must navigate the messy reality that every “solution” redistributes power, disrupts established interests, and challenges someone’s sense of identity or security.

This political dimension becomes even more critical in the age of AI, where technological change arrives faster than social consensus can form. Leaders must create the processes and spaces where society can deliberate about the future it wants – not just react to the future that arrives. They must build the legitimacy for difficult decisions by ensuring those affected have genuine voice in shaping them. They must recognize that resistance to change often reflects not ignorance but legitimate concerns about fairness, dignity, and belonging that no algorithm can address.

Ultimately, this means that technical competence, while necessary, is insufficient. The modern leader must cultivate what might be called *political wisdom*: the capacity to see the human dimensions behind every data point, to understand that every system contains power relationships, and to recognize that sustainable change requires not just better policies but renewed social contracts.

Political wisdom requires understanding that change is nonlinear and path-dependent. What appears to be gradual progress can suddenly flip into a new equilibrium when tipping points are crossed. History shows that societal transitions are often crisis-driven – triggering phase changes that are neither gradual nor easily reversible. Leaders must therefore cultivate antifragile systems that gain strength from volatility rather than merely surviving it. This requires building redundancy, maintaining optionality, ensuring decision-makers have skin in the game, and accepting some inefficiency to preserve resilience.

Perhaps most critically, political wisdom recognizes that values and worldviews ultimately determine the success of change efforts. As we’ve seen throughout this chapter, technological capabilities alone are insufficient;

## *Conclusion: A More Political Modern Leadership Agenda*

what societies can implement depends fundamentally on their shared values and ethical frameworks. If a significant portion of society believes that economic growth is inherently virtuous and any suggestion of limits is “anti-progress,” then policies aiming for sustainability will face fierce resistance regardless of their technical merit. The converse is also true: worldviews can enable change when aligned.

The leaders who will successfully navigate the age of AI are those who understand that their primary work is not optimizing systems but tending to the social fabric that holds those systems together. This means building trust through transparency and genuine consultation, fostering solidarity across diverse stakeholders, and articulating visions that resonate across cultural and ideological boundaries.

## Key Takeaways: Societal Change

### What You Can Do:

- **Recognize reflexive interconnection:** Individual, organizational, and societal change are interconnected through continuous feedback loops. As individuals adapt to AI, organizations transform, shaping economic shifts and policy debates – while societal anxieties constrain organizational experimentation. Work across all levels simultaneously.
- **Acknowledge the new zeitgeist:** The 2020s represent a shift from late-20th-century optimism toward overlapping crises (climate, inequality, geopolitical conflict, AI acceleration). Wealth concentration has returned to early-20th-century levels. The pandemic exposed fragile systems optimized for efficiency. Leaders must operate in this context of instability and complexity.
- **Move beyond linear models:** Traditional technocratic approaches assuming steady progress are inadequate. Societal transitions come in surges, ruptures, and tipping points – not gradual improvements. Crises catalyze change by unfreezing the status quo. How society responds (foresight and adaptability) often determines survival.
- **Build antifragile systems:** Modern societies optimized for efficiency have become dangerously fragile. Design for antifragility: embrace decentralization, diversity, redundancy, “skin in the game,” and optionality. Avoid pseudo-stability that hides risks. Conduct regular stress tests and deliberately inject small disruptions to reveal vulnerabilities.
- **Address governance challenges:** Key questions: Who decides what AI optimizes for? How do we ensure benefits are shared broadly? What mechanisms enable democratic oversight? How do we build trust infrastructure? Use tools like agent-based modeling and digital twins for policy testing, but remember these are aids for human judgment, not replacements.

*Key Takeaways: Societal Change*

- **Develop political leadership skills:** Leadership in the AI age is unavoidably political – building coalitions, negotiating competing interests, creating shared meaning across differences. Skills resemble statecraft: reading stakeholder dynamics, sensing power shifts, brokering compromises, communicating vision across worldviews. Technical competence alone is insufficient.



# **Societal Impacts of AI**

AI is reshaping fundamental aspects of the economy, society, and individual life. Far from a niche concern, AI's influence cuts across economic structures, educational systems, scientific research practices, privacy and surveillance regimes, the attention economy, power dynamics, and the cultural fabric of society. This chapter examines these multifaceted and interrelated impacts. The discussion is organized around key domains from the workforce and education to privacy and culture and concludes with forward-looking perspectives on AI's future and the societal capacity to adapt.

## **Economic and Workforce Impacts of AI**

One of AI's most immediate societal impacts is on labor markets. Organizations are deploying AI to automate tasks, augment worker capabilities, and create new services. However, the net effect on productivity and jobs is complex and not easy to predict.

### **Task Automation and Job Displacement Effects**

Some studies suggest that recent AI advances have only a rather modest near-term macroeconomic impact. Acemoglu used existing estimates on exposure to AI and productivity improvements at the task level in 2024 to estimate the macroeconomic effects to be no more than a 0.71% increase in total factor productivity over 10 years. He argues that unlike the internet, whose transformative potential was evident early, today's AI lacks clearly

## *Societal Impacts of AI*

scalable, production-transforming applications that create valuable new services at economy-wide scale.<sup>1</sup>

Brynjolfsson et al. studied in 2023 data from over 5,000 customer support agents, where access to an AI tool increased productivity by 14% on average, as measured by issues resolved per hour, with the greatest impact on novice and low-skilled workers, and minimal impact on experienced and highly skilled workers.<sup>2</sup> More recent labor data analysis showed that early-career workers (ages 22-25) in the most AI-exposed occupations, in particular software engineering and customer support, have experienced a 13% relative decline in employment even after controlling for firm-level shocks.<sup>3</sup> This aligns with mounting anecdotal evidence of companies substituting whole services and functions with AI agents. For example, the financial services company Klarna announced in February 2024 that its AI assistant powered by OpenAI handled two-thirds of its customer service chats in 35 languages, the equivalent of 700 full-time customer service agents, driving a \$40 million profit improvement in 2024.<sup>4</sup>

According to a December 2025 report by OpenAI, 75% of surveyed workers reported improved speed or quality, with ChatGPT Enterprise users saving 40–60 minutes per active day on average (data science, engineering, and communications workers save 60–80 minutes). 75% report completing tasks they previously could not, including programming, code review, spreadsheet analysis and automation, technical tool development, and custom agent design.<sup>5</sup>

---

<sup>1</sup>Acemoglu, Daron, “The Simple Macroeconomics of AI,” *National Bureau of Economic Research*, 2024.

<sup>2</sup>Brynjolfsson, Erik et al., “AI and Productivity,” *arXiv Preprint*, 2023.

<sup>3</sup>Brynjolfsson, Erik et al., “Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence,” *Stanford Digital Economy Lab*, 2025, [https://digitaleconomy.stanford.edu/wp-content/uploads/2025/11/CanariesintheCoaIMine\\_Nov25.pdf](https://digitaleconomy.stanford.edu/wp-content/uploads/2025/11/CanariesintheCoaIMine_Nov25.pdf).

<sup>4</sup>Klarna, *Klarna AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month*, 2024, <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>.

<sup>5</sup>OpenAI, *The State of Enterprise AI 2025*, OpenAI, 2025, <https://openai.com/index/the-state-of-enterprise-ai-2025-report/>.

In November 2025 Anthropic published an analysis of one hundred thousand real world conversations with Claude, which estimates that AI reduces task completion time by 80%.<sup>6</sup> According to these estimates, people typically use AI for complex tasks that would, on average, take people 1.4 hours to complete. The estimated scope, cost, and time savings of tasks varies widely by occupation. Extrapolating these results to the US economy, Anthropic suggests that current generation AI models could increase annual US labor productivity growth by 1.8% over the next decade. This would double the annual growth the US has seen since 2019.

Clearly, automation's impact will hit some areas quickly (especially routine cognitive work) while other areas remain resilient, particularly jobs demanding social intelligence, complex judgment, and tacit knowledge. To predict the effects of AI on employment and wages Autor and Thompson propose in 2025 a differentiated "expertise model" of automation's impact.<sup>7</sup> If AI automates low-skill, routine tasks (removing the "inexpert" tasks from a job), the remaining work requires higher expertise. As a result, wages tend to rise for those who can do the now-more-skilled job, but overall employment may fall because fewer people are qualified. If AI automates high-skill, expert tasks, the remaining work becomes more routine and requires less expertise, so wages tend to fall and employment might increase (since a larger pool of workers can perform the simpler tasks). This model explains why some occupations that lost routine tasks still saw wage increases as removing them upgraded the skill concentration of the role. Conversely, if AI "de-skills" a profession by handling its toughest parts, the value (and pay) of the human role can decline. The broader implication is that who benefits from AI – workers or employers, high-skilled or low-skilled – will depend on the task composition of jobs that AI reshapes.

One complication in making such task-oriented predictions is that the capabilities of the AI models are increasing rapidly and more complex tasks

---

<sup>6</sup>Tamkin, Alex, and Peter McCrory, "Estimating AI Productivity Gains from Claude Conversations," Anthropic, 2025, <https://www.anthropic.com/research/estimating-productivity-gains>.

<sup>7</sup>Autor, David, and Alan Thomson, "The Expertise Model of Automation's Impact," *arXiv Preprint*, 2025.

can be tackled by frontier model AI agents. METR (Model Evaluation and Testing for Reliability) is a nonprofit research organization that systematically evaluates the capabilities and risks of advanced AI systems. They document an exponential trend, where each generation of LLMs increases the time horizon of solvable software tasks.<sup>8</sup>

This trend is also visible in a 2025 survey of 132 Anthropic engineers' internal Claude Code usage to assess AI's impact on software development.<sup>9</sup> Engineers report using Claude in 60% of their work (up from 28% a year ago) with a 50% productivity boost, and 27% of Claude-assisted work consists of tasks that wouldn't have been done otherwise. The most common uses are debugging and code understanding, and engineers are becoming more "full-stack," working in areas beyond their normal expertise. Claude is handling increasingly complex tasks autonomously, completing about 20 actions before needing human input (up from 10 six months ago). The survey reveals a workplace in transition, where productivity gains and expanded capabilities coexist with questions about maintaining technical depth, preserving meaningful collaboration, and navigating an uncertain future for software engineering as a profession. If this scaling continues, within the next few years models could autonomously complete day-scale software projects, fully automate repetitive coding and debugging, security testing, patching, and integration pipelines. This trend is illustrated in the following figure.

Another complication is that AI adoption and worker behavior unfold on platforms outside official data. Existing workforce planning frameworks were designed for human-only economies. They track employment, wages, and productivity, but were not designed to measure where AI capabilities overlap with human skills before adoption reshapes occupational structure. Chopra et al. address this gap by introducing a new skills-centered metric called the *Iceberg Index*, which measures the wage value of skills that AI systems can perform within each occupation.<sup>10</sup> This index captures tech-

---

<sup>8</sup>METR, *Model Evaluation and Testing for Reliability*, 2025, <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>.

<sup>9</sup>Huang, Saffron et al., "How AI Is Transforming Work at Anthropic," Anthropic, 2025, <https://www.anthropic.com/research/how-ai-is-transforming-work-at-anthropic/>.

<sup>10</sup>Chopra, Ayush et al., "The Iceberg Index: Measuring Workforce Exposure in the AI

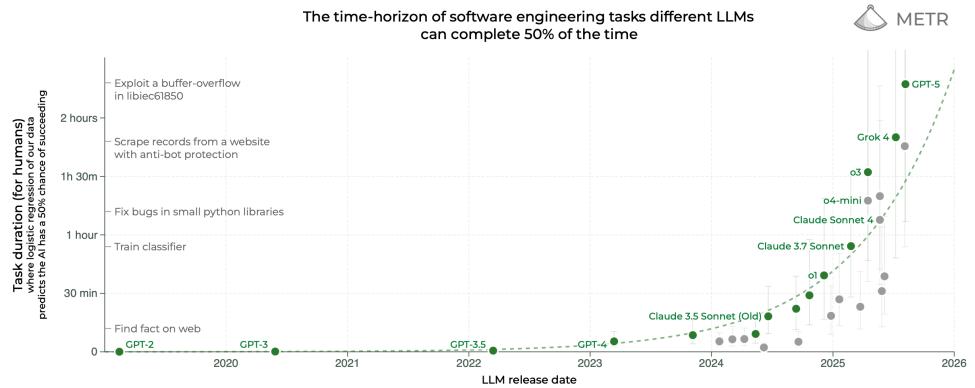


Figure 34: Solvable software tasks by model generation

nical exposure – where AI can perform occupational tasks – rather than displacement outcomes or adoption timelines. The name reflects a critical insight: visible AI adoption in the United States, concentrated in computing and technology (2.2% of wage value, approximately \$211 billion), represents only the tip of the iceberg. Technical capability extends far below the surface through cognitive automation spanning administrative, financial, and professional services (11.7%, approximately \$1.2 trillion).

In the face of exponential increases in model capabilities, the goalposts for “uniquely human” work shift: what was once considered AI becomes standard automation, and humans move on to the next thing. There is profound uncertainty about AI’s trajectory from business-as-usual automation to artificial general intelligence (AGI) within 5–20 years, where AI may eventually overtake all human capabilities. Given this uncertainty Anton Korinek argues for adopting a portfolio approach rather than planning for a single outcome. He presents three scenarios: (1) traditional automation that creates new jobs for displaced workers, (2) AGI emerging in 20 years that can perform all human work tasks, and (3) an aggressive timeline reaching AGI in just five years.<sup>11</sup> Since even leading AI researchers

---

Economy,” *arXiv Preprint*, 2025, <https://arxiv.org/abs/2510.25137>.

<sup>11</sup>Korinek, Anton, “Scenario Planning for an a(g)i Future,” *Finance & Development* 60, no. 4 (2023): 30–33, <https://www.imf.org/en/Publications/fandd>.

hold starkly divergent views, Korinek contends each scenario has at least a 10% probability and requires serious preparation through stress-testing existing economic frameworks and developing adaptive policies. Korinek has modeled dramatically different economic trajectories across these scenarios.<sup>12</sup> While output growth continues at historical rates in the business-as-usual scenario, the two AGI scenarios produce much faster growth as labor scarcity ceases to constrain production. This leads to a more troubling pattern for workers: wages initially rise in all scenarios while labor remains scarce, but then collapse as the economy approaches AGI and machines can substitute for human workers. Crucially, Korinek notes that both the output surge and wage collapse stem from the same mechanism with abundant machines replacing scarce labor. This calls for the design of norms and institutions that redistribute AGI gains to compensate workers and ensure shared prosperity rather than concentrated wealth.

## **Deskilling: The Erosion of Expertise**

While AI can dramatically enhance productivity for those who use it skillfully, mounting evidence reveals a troubling paradox: the very tools designed to augment human capability may be systematically undermining it. This phenomenon, known as “deskilling,” occurs when humans over-rely on AI assistance to the point where their own competencies atrophy.

The mechanism of deskilling is insidious precisely because it operates through apparent enhancement. Consider the striking case documented in medical practice: experienced gastroenterologists who had spent years or even decades honing their ability to identify cancerous polyps during colonoscopies showed significant deterioration in diagnostic accuracy after just three months of working with AI assistance.<sup>13</sup> These were not novices but expert clinicians at the peak of their professional capability. Yet when provided with an AI diagnostic aid that flagged suspicious tissue, they

---

<sup>12</sup>Korinek, Anton, and Donghyun Suh, “Scenarios for the Transition to AGI,” *National Bureau of Economic Research*, 2024, <http://www.nber.org/papers/w32255>.

<sup>13</sup>Budzyń, K. et al., “Endoscopist Deskilling Risk After Exposure to Artificial Intelligence,” *Journal of Gastroenterology*, 2025.

began deferring judgment to the algorithm rather than maintaining their own practiced vigilance. The AI had become a cognitive crutch that, once depended upon, revealed itself to have weakened the very capacity it was meant to support.

This pattern appears across domains. Pilots using advanced autopilot systems show degraded manual flying skills, becoming dangerously dependent on automation that occasionally fails in critical moments. Software developers relying heavily on AI code-completion tools report declining facility with fundamental programming concepts and problem-solving approaches. In each case, the AI doesn't replace the human entirely, but it quietly erodes specific capabilities while leaving others intact, creating professionals who appear competent in routine circumstances but lack the depth of skill to handle novel challenges or system failures. The deeper issue is that expertise isn't simply stored knowledge that can be offloaded to external tools; it's a form of embodied, practiced capability that requires continuous active engagement to maintain.

If deskilling threatens experienced professionals, its impact on early-career workers may be even more devastating – and structurally different. The traditional path to expertise in most knowledge professions involves a progression from simpler to more complex tasks: junior lawyers do document review before arguing cases; medical residents handle routine cases before complex ones; entry-level analysts perform basic research before strategic planning. This apprenticeship model serves two crucial functions: it allows novices to contribute economically while learning, and it provides the repeated practice through which foundational skills become automatic, freeing cognitive resources for higher-level judgment. AI threatens to sever this connection between economic contribution and skill development. When AI can perform document review, routine diagnosis, or basic analysis faster and cheaper than junior professionals, the economic rationale for employing novices in these roles evaporates. But these “simple” tasks were never merely productive work – they were the training ground where professionals developed the pattern recognition, judgment, and contextual understanding that later enables expert performance.

The impact on early-career professionals is already massive in some sectors.

## *Societal Impacts of AI*

In law, AI-powered document review tools have dramatically reduced demand for junior associates who previously spent years learning to identify relevant information in complex cases. In finance, algorithmic trading and automated analysis eliminate entry-level positions that once taught market dynamics and risk assessment. In journalism, AI-generated content reduces opportunities for junior reporters to develop their craft through repetitive coverage of routine events. In each case, experienced professionals can leverage AI to amplify their productivity – they have the judgment to guide, verify, and contextualize AI outputs. But newcomers lack precisely those capabilities, and now also lack the pathway to develop them. This creates what we might call a “flywheel effect” where advantages compound very specifically: experienced workers use AI to become even more productive, capturing more of the valuable work, which leaves less room for novices to gain the experience that would make them capable. The gap between expert and novice, traditionally bridged by years of progressive responsibility, becomes a chasm. Senior professionals become “super-experts” augmented by AI, while entry-level workers find themselves economically redundant before they’ve developed marketable skills.

## **Hybrid Human-AI Teams**

In dividing labor, an emerging view is that hybrid human-AI teams often perform best. AI can propose options or handle routine analyses, and humans make final judgments – leveraging the AI’s speed and breadth with human common sense and values. For example, in healthcare diagnostics, AI might flag top suspect areas in an image, but a doctor reviews them and makes the call, aware of context the AI might miss. This collaboration can also serve as a check: the human can override the AI when it’s wrong (avoiding automation bias by training operators to stay critical), and the AI can alert the human to things they might overlook (avoiding human error). The following figure shows how hybrid teams outperform human teams or individuals.

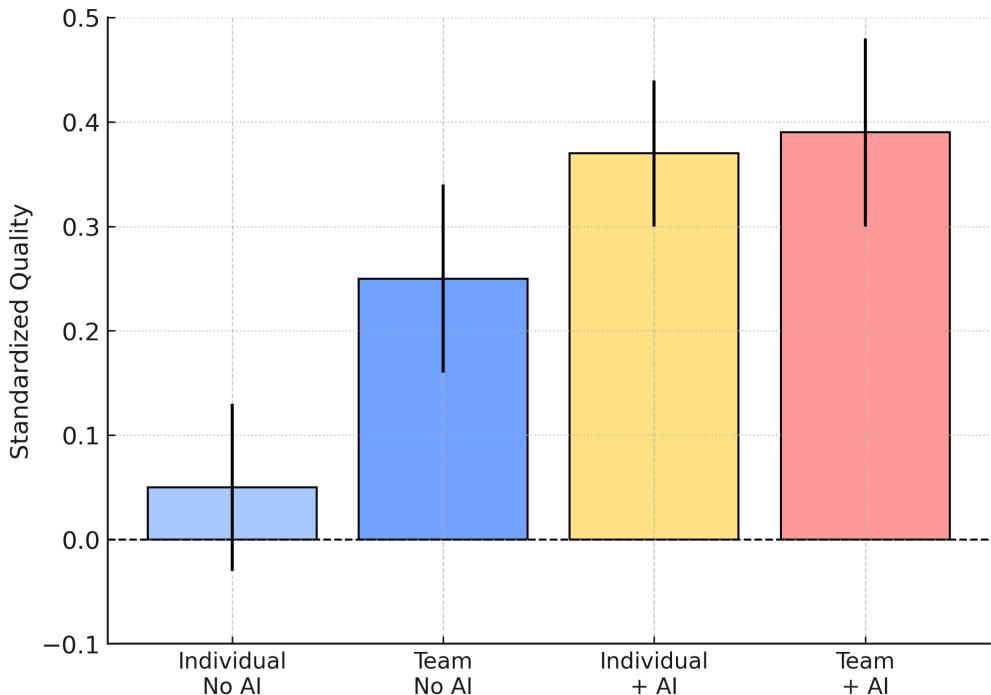


Figure 35: Hybrid teams outperforming human teams or individuals.<sup>14</sup>

Designing interfaces and workflows for optimal human-AI synergy is therefore a key task in the coming years. Addressing these challenges requires moving beyond the assumption that AI assistance is universally beneficial to recognizing that human skill development has requirements that pure efficiency maximization may violate.

Several principles emerge:

- Deliberate practice must be protected: Professionals need regular engagement with challenging cases that require genuine cognitive effort, not just verification of AI outputs. This may mean intentionally working without AI assistance on certain tasks, or designing roles that keep humans “in the loop” in ways that maintain rather than erode capability.
- Training pathways must be redesigned, not just eliminated: If traditional entry-level roles disappear, professions need alternative mech-

---

<sup>14</sup>Dell’Acqua, “Navigating the Jagged Technological Frontier.”

## *Societal Impacts of AI*

anisms for skill development – perhaps intensive residencies, simulation-based training, or new forms of mentorship that focus on judgment rather than execution. - Competency assessment must evolve: Evaluating professionals should include testing their capability without AI assistance, not just their ability to leverage AI effectively. We need to distinguish between genuine expertise and dependency disguised as capability. - Economic models must account for skill formation: The knowledge economy traditionally paid junior professionals poorly while they learned and senior professionals well after they'd developed expertise. If AI eliminates the junior roles, how do we fund the extended training required to produce future experts? This may require reconceiving professional education as public investment rather than private credential-seeking.

Ultimately, the goal isn't to reject AI assistance but to ensure it augments rather than replaces human capability to create what researchers call “intelligence amplification” in hybrid teams rather than intelligence replacement. This requires intentional design of both technology, teams, and institutions to preserve the conditions under which human expertise develops and flourishes, even as we leverage AI's remarkable capabilities. The alternative is a future where we've built systems we depend on but no longer understand, staffed by professionals who appear competent but lack the deep capability to adapt, innovate, or maintain the very tools they've come to rely upon.

## **Reshuffling Business Models and the Economy**

In his book *Reshuffle* Sangeet Paul Choudary argues that sooner than later AI will reshuffle the knowledge economy to the point where the basis of competition will be changed fundamentally in most sectors leading to new business models and major labor market upheaval.<sup>15</sup> This fundamental reshuffling is illustrated in the following figure.

---

<sup>15</sup>Choudary, Sangeet Paul, *Reshuffle: Who Wins When AI Restacks the Knowledge Economy* (Amazon Digital Services LLC – KDP, 2025).

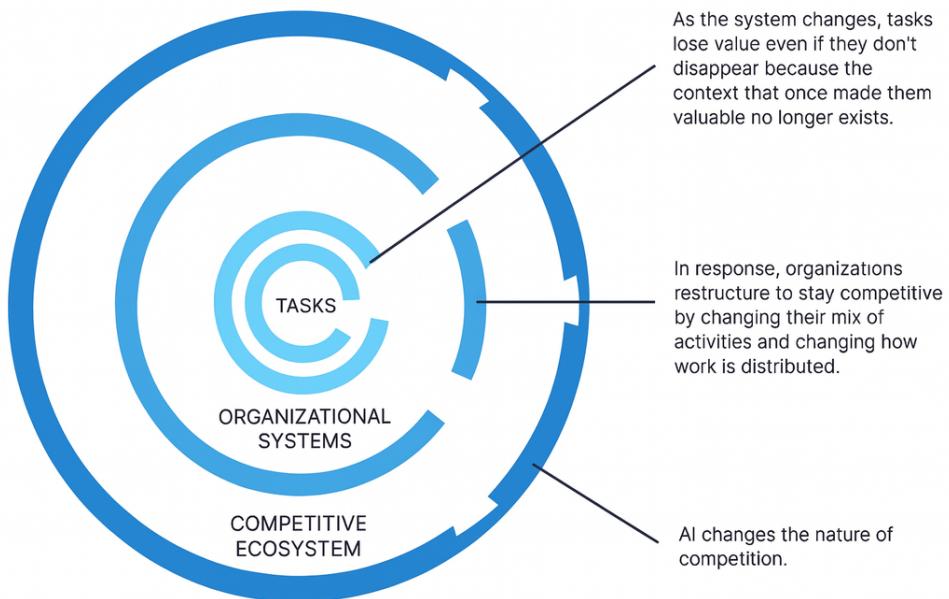


Figure 36: AI will reshuffle the economy fundamentally.<sup>16</sup>

He illustrates his argument with the changes in the legal sector where studies have shown the productivity shift from AI, as shown in the following figure.

Over time, the input-oriented billing models of law firms or consultancies are likely to be disrupted. Traditional law firm structures built around billable hours and time-based compensation will give way to new models where value is measured by outcomes rather than inputs. As AI automates routine legal tasks and accelerates research and document preparation<sup>18</sup>, firms that continue to bill by the hour will find their revenue models undercut

---

<sup>16</sup>Choudary, *Reshuffle*.

<sup>17</sup>Choi, Jonathan H. et al., *Lawyering in the Age of Artificial Intelligence*, 4626276 (Social Science Research Network, 2024), <https://ssrn.com/abstract=4626276>.

<sup>18</sup>Reuters, "PwC's 4,000 Legal Staffers Get AI Assistant as Law Chatbots Gain Steam," *Reuters*, 2023, <https://www.reuters.com/world/uk/pwcs-4000-legal-staffers-get-ai-assistant-law-chatbots-gain-steam-2023-03-15/>.

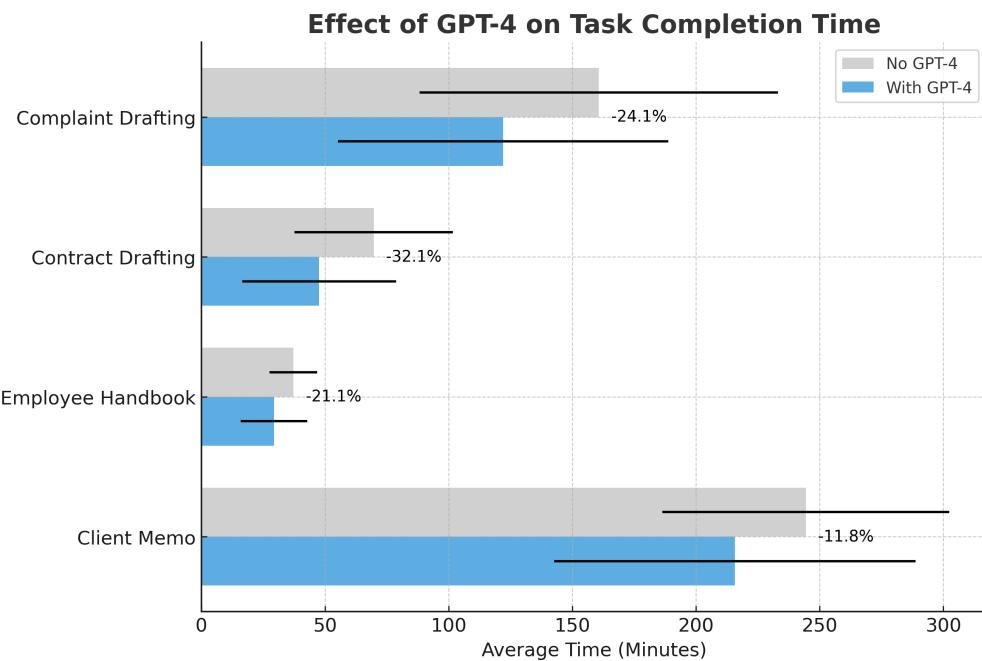


Figure 37: AI impact on legal tasks<sup>17</sup>

by competitors offering fixed-fee or value-based pricing. This shift will fundamentally restructure law firm economics: from pyramid-shaped organizations with many junior associates billing hours, toward flatter structures where AI handles routine work and human lawyers focus on high-value strategic counsel. The traditional leverage model, where partners profit from associates' billable hours, becomes unsustainable when AI can perform much of that work at a fraction of the cost. Firms that adapt will transition to outcome-based pricing, subscription models, or performance-based fees, while those clinging to hourly billing risk being displaced by more efficient, AI-enabled competitors<sup>19</sup>.

From a policy perspective, ensuring that the workforce can transition and upskill alongside AI (through education, training, and safety nets) will be critical to broad-based economic resilience. If AI rapidly displaces millions

---

<sup>19</sup>Choudary, *Reshuffle*.

of workers without a plan, we face a systemic socio-economic risk.<sup>20</sup> If only a small elite has access to the best AI (creating a class of “super-powered” individuals or companies while others lag), that’s a systemic risk to social stability and fairness. So the division and automation of labor isn’t just about efficiency, it’s inevitably also about sustainability and justice in the age of AI. Ensuring broad access to AI benefits, retraining workers, and even rethinking social safety nets (like universal basic income or other support mechanisms) will be necessary to manage the systemic impacts of widespread AI adoption.

## **Education and Personalized Learning at Scale**

AI is poised to dramatically expand the scale and personalization of education. Traditional one-size-fits-all training methods are giving way to individualized learning experiences powered by AI. It is now feasible to provide one-on-one style tutoring or coaching to millions of learners simultaneously, something previously impossible due to human resource constraints. Recent advances showcase how personalized learning can be set up with relative ease using AI tools.<sup>21</sup> For example, systems like Google’s NotebookLM (Learning Mode) and GPT-5’s study mode act as AI tutors, generating explanations, summaries, and multi-modal study materials tailored to an individual student’s needs. Students can upload their course materials and receive synthesized notes or even AI-generated podcasts reviewing the content. Similarly, “focused coach” bots can be prompted to act as mentors for specific subjects or skills, adapting their guidance to the learner’s progress.

Organizations like Khan Academy have piloted AI-powered assistants (e.g., Khanmigo<sup>22</sup>) to support learners at scale. Such AI tutors can adjust the pace of instruction, provide targeted practice on a student’s weak areas,

---

<sup>20</sup>Korinek and Suh, “Scenarios for the Transition to AGI.”

<sup>21</sup>Mollick, *Co-Intelligence*.

<sup>22</sup>Khan, Sal, *Khanmigo - on-Demand AI-Powered Support for Education*, 2024, <https://www.khanmingo.ai>.

and even emulate Socratic questioning to develop critical thinking. Clearly, we are still mainly at a trial-and-error stage and technical and evaluatory challenges remain<sup>23</sup>. Yet, the promise is an education system that meets each learner where they are, rather than the learner having to conform to a standardized pace. For businesses and executive training, this means AI can deliver highly customized learning programs, on-demand skill coaching, and continuous development advice for employees, all at scale. The scalability of AI tutors could help address talent gaps by rapidly upskilling employees with personalized curricula. It also holds potential for addressing educational inequity globally – for example, students in under-resourced regions could access quality instruction tailored to their needs through AI platforms, closing some of the resource gap.

While AI-enabled personalized learning is a powerful innovation, it raises questions about maintaining quality and human touch. Educators will need to ensure that AI content is accurate, unbiased, and aligned with pedagogical goals. AI can deliver facts and drill practice effectively, but mentorship and motivational support as provided by human teachers must be reimagined. In corporate settings, AI training systems should be coupled with real-world practice and human coaching to ensure soft skills and cultural nuances are learned. Privacy is another consideration: personalized learning systems rely on extensive data on individual learners (performance, preferences, even biometric or attention data in some cases). Despite these challenges, the trajectory is clear: AI is making lifelong learning more accessible and tailored than ever before, enabling education at scale without sacrificing personalization.

The core challenge will be to motivate learners to embark on the necessary hard work of developing skills, knowledge, and judgement when AI agents are standing by to provide short-cuts to seemingly adequate answers. This represents perhaps the most profound paradox facing education in the age of AI: the very tools that make learning more accessible may undermine the motivation to learn deeply. Why struggle through the fog of confusion that precedes genuine understanding when an AI can deliver a polished

---

<sup>23</sup>Jurenka, Irina et al., “Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach,” *Google DeepMind*, 2024.

answer in seconds? Why endure the frustration of solving a complex problem when a chatbot can walk you through the solution step-by-step? The cognitive effort that builds expertise – the wrestling with difficult concepts, the productive failure, the slow accumulation of pattern recognition – suddenly appears optional, even wasteful. Yet this apparent efficiency masks a dangerous illusion. The “adequate” answers AI provides are often just that – adequate for the immediate task, but insufficient for building the mental models, intuitions, and transferable capabilities that constitute genuine competence. A student who uses AI to complete assignments may finish with acceptable work, but without the struggle that wires neural pathways and develops judgment. They acquire outputs without internalizing processes, answers without understanding, credentials without capability. The problem deepens when we consider motivation theory. Human beings are wired to conserve cognitive effort. We naturally seek the path of least resistance. When AI offers a cognitive shortcut, choosing the harder path of independent struggle requires either exceptional intrinsic motivation or external structures that make the shortcut less attractive than the effort. But traditional motivational levers, such as grades, credentials, or even the promise of future employment, lose their power when AI can perform many credentialed tasks competently.

This creates an *expertise depreciation dilemma*. As AI capabilities expand, the half-life of specific skills shortens, making the return on investment for any particular expertise seem questionable. Learners rationally ask: “Will this skill I’m struggling to master even matter by the time I’ve developed it?” This question can paralyze motivation precisely when the need for human judgment has never been greater. The deeper issue is that we’re asking learners to delay gratification in an unprecedented way. Previous generations could see a clear connection between effort and outcome: study hard, develop skills, secure better opportunities. But when AI can produce seemingly equivalent outputs with minimal effort, the value proposition of deep learning becomes abstract and deferred. We’re essentially asking people to invest in capabilities whose value lies not in producing outputs (AI can do that) but in developing judgment about which outputs matter, wisdom about how to apply them, and creativity in imagining what questions to ask – benefits that are real but invisible, long-term, and difficult to

credential.

The implication is that educational institutions face a fundamental redesign challenge. The traditional model (transmit knowledge, assess recall, credential competence) breaks down when knowledge transmission is instant and recall is outsourced. We need new frameworks that make the development of judgment intrinsically valuable and immediately rewarding, not just instrumentally useful for some distant future. This means shifting from teaching answers to teaching the art of questioning, from measuring outputs to assessing thinking processes, from credentialing knowledge to validating judgment.

Some promising directions emerge: problem-based learning where AI is a tool but judgment is the goal; apprenticeship models where learners see expert judgment in action and understand its value; assessment methods that explicitly require demonstrating understanding that AI cannot fake; and perhaps most importantly, helping learners experience the intrinsic satisfaction of genuine comprehension. We might also need to cultivate a new cultural narrative around learning: that the goal isn't to compete with AI in producing outputs, but to develop the distinctly human capacity for contextual judgment, ethical reasoning, creative synthesis, and meaning-making. The learner who develops deep expertise isn't trying to outperform AI at calculation or recall, but cultivating the wisdom to know what's worth calculating, the judgment to assess whether an answer makes sense, and the creativity to imagine entirely new questions.

Ultimately, we may need to accept that motivation will become more selective. Not everyone will choose the hard path of deep learning, no pun intended, when shortcuts exist. But for those roles where judgment truly matters, where decisions affect lives, where context is complex, where values are at stake, we must find ways to make the development of genuine expertise not just necessary but desirable. The alternative is a society of credentialed incompetence: people with degrees and titles who can generate plausible-sounding answers but lack the judgment to know which answers are actually right, or even which questions are worth asking.

## **Scientific Research Acceleration and Epistemological Risks**

AI is not only automating routine work; it is also transforming the frontiers of scientific research. Across fields from chemistry to mathematics, AI systems are contributing to discoveries that were once out of reach. For instance, AI models have been used to simulate quantum chemistry problems and model electron densities in materials, tasks that traditionally required immense computational effort. In nuclear fusion research, AI-driven plasma control strategies have accelerated experimental progress. AI algorithms have designed efficient chip architectures and even proved non-trivial mathematical conjectures in topology<sup>24</sup>. These examples hint at *a new kind of science at scale* where AI's ability to sift through vast data or explore countless configurations fundamentally augments human scientific capabilities. By partnering with AI, scientists can iterate faster, test more hypotheses, and perhaps tackle problems that were formerly intractable.

### **Acceleration vs Understanding**

The growing role of AI in science raises a pivotal question about the future of human knowledge itself: does the scientific method fundamentally transform when AI can discover patterns that exceed human comprehension? In 2008, WIRED editor Chris Anderson provocatively argued that the coming data deluge might render the traditional scientific method – hypothesize, model, test, refine – effectively obsolete.<sup>25</sup> He envisioned an era of “correlation without causation,” where massive datasets and powerful algorithms would simply find what works without requiring explanatory theories about why it works. More than fifteen years later, Anderson’s prediction has partially materialized in unsettling ways. We now routinely see AI systems producing genuinely useful results, e.g., accurate predictions

---

<sup>24</sup>DeepMind, “Mathematical Discoveries with LLMs,” *Nature* 623 (2023): 561–67.

<sup>25</sup>Anderson, Chris, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” *WIRED*, 2008, <https://www.wired.com/2008/06/the-end-of-theo/>.

of protein folding, novel drug candidates, or optimized materials with desired properties, without generating any human-comprehensible theory to explain their success. These systems operate as black boxes: we can verify their outputs work, but we cannot reliably trace the reasoning that produced them. The scientific method hasn't ended, but it has bifurcated into two increasingly divergent modes: the traditional theory-driven approach aimed at understanding, and an emerging data-driven approach focused purely on optimization and results.

The pragmatic, results-oriented paradigm offers undeniable advantages. AI-driven drug discovery can screen millions of molecular combinations in silico, identifying promising candidates in months rather than years. Materials science can explore vast parameter spaces impossible for human intuition to navigate. Climate models can incorporate thousands of variables simultaneously, revealing complex interactions that simpler models miss. The acceleration is real, and in domains where speed matters the ability to find solutions without first understanding them fully can save lives.

Yet this pragmatism carries profound risks that extend beyond any single discipline. When we optimize without understanding, we navigate by instruments we cannot interpret, making it nearly impossible to know when we're approaching the limits of their validity. An AI might identify a drug candidate that works brilliantly in trials but fails catastrophically in some unanticipated context and without a theoretical framework, we have no way to predict where those boundaries lie. The acceptance of effective ignorance may be our generation's Faustian bargain with computational power. The epistemological concern runs deeper still. If we allow AI to guide scientific inquiry toward what is readily discoverable in existing data, we risk creating a profound confirmation bias at the level of entire research programs. Kate Crawford has compellingly argued that the questions we ask and the data we collect fundamentally shape what we can discover: AI-driven science may inadvertently close off entire avenues of inquiry that don't fit the patterns learnable from available datasets.<sup>26</sup> The algorithm

---

<sup>26</sup>Crawford, Kate, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021).

becomes not just a tool but a lens that focuses our attention in certain directions while rendering others invisible.

There's also a concerning feedback loop: as AI systems identify "successful" research directions based on citation patterns, funding success, and incremental progress, they may inadvertently reinforce conservative, low-risk science over the kind of speculative, paradigm-shifting work that initially seems foolish or fruitless. The AI learns what success looks like from past data, potentially calcifying the scientific method around historical patterns of discovery rather than enabling radical new modes of inquiry. Moreover, the loss of mechanistic understanding carries risks for scientific culture itself. Science isn't merely a machine for generating useful predictions; it's a collective human enterprise of building shared understanding about reality. When findings emerge from black-box algorithms, that shared understanding fractures. Scientists in different labs may use different AI tools that produce different results for reasons neither can articulate. The reproducibility crisis that already plagues some disciplines could metastasize when the "methods" section of a paper essentially reads: "We trained a neural network until it worked." Science depends on transparent chains of reasoning that allow peer review, cumulative knowledge building, and the training of new scientists who can extend rather than merely replicate existing work.

The path forward likely requires what we might call "hybrid epistemology", i.e. a scientific method that leverages AI's pattern-finding power while preserving human judgment about which patterns matter and why. This means developing new tools for interpretable machine learning that can translate AI discoveries back into human-comprehensible theories, even if approximate. The need for interpretability of AI systems is not only a safety issue but also an epistemological concern.<sup>27</sup> It also means training scientists to work productively with AI while maintaining the skepticism and curiosity that drive genuine discovery. It means deliberately protecting funding and institutional space for research that doesn't fit AI-discoverable patterns and fostering the weird, the speculative, the apparently useless

---

<sup>27</sup>Korbak, Tomek et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety," *Preprint*, 2025.

curiosity-driven work from which revolutionary insights so often emerge.

Most fundamentally, it requires recognizing that “what works” and “why it works” serve different but equally essential purposes. The former solves immediate problems; the latter builds cumulative knowledge that enables us to solve entire classes of future problems we haven’t yet imagined. A science that abandons the pursuit of understanding in favor of pure optimization may deliver impressive short-term results while quietly undermining the very capacity for scientific progress that generated those tools in the first place. The question isn’t whether to use AI in science but whether we can harness its power without losing sight of what makes science not just useful, but meaningful: the fundamentally human drive to understand, not merely to predict, the nature of reality.

## **Reproducibility and Epistemological Risks**

There is concern that AI, if misused, could actually slow the progress of science despite the explosion of output. A production-progress paradox has been noted: while scientific output (papers, results) has grown exponentially, truly disruptive breakthroughs and fundamental insights have stagnated or declined. Meta-research studies find fewer novel ideas and a lower rate of paradigm-changing discoveries in recent decades, despite more papers being published.<sup>28</sup> AI could unintentionally exacerbate this by making it even easier to generate incremental research: paper mills using AI text generation, countless minor variants of analyses, etc., all adding noise and volume without corresponding insight. Incentive structures in academia reward quantity (number of publications, citation counts) over quality, discouraging risk-taking and originality. AI might accelerate this trend by allowing average research to be produced more quickly, thus crowding out the truly novel work resulting in a form of homogenization of science. Another risk stems from the fact that most scientists today are not trained in software engineering; errors in AI-driven analyses (bugs in code or data bias) can be subtle and hard to detect, leading to flawed

---

<sup>28</sup>Park, Michael et al., “Papers and Patents Are Becoming Less Disruptive over Time,” *Nature* 613 (2023): 138–44, <https://doi.org/10.1038/s41586-022-05543-x>.

findings propagating through literature. Without proper checks, AI could introduce systematic errors or false discoveries that others then build upon, undermining the integrity of scientific knowledge. This makes reproducibility a critical issue. As researchers use proprietary AI models or complex pipelines they only partly grasp, replicating results may become more difficult.

On the other hand, AI, if applied thoughtfully, can bolster good scientific practice. It can help with replication efforts, error-checking, and managing information overload. For example, AI tools can automatically verify calculations, flag anomalies, or suggest better experimental designs. Kapoor and Narayanan argue that AI could aid reproducibility, debugging, and comprehension in science but only if our systems and incentives prioritize these goals over raw output.<sup>29</sup> In other words, the way we integrate AI into scientific workflows must be deliberately managed to enhance rigor rather than just speed.

Some complicating shifts are already happening: the locus of AI-driven science is moving toward industry labs (with big tech companies publishing cutting-edge research), and geopolitically, countries investing in AI research infrastructure are gaining an edge. These shifts might influence what research questions get attention. To maintain a healthy scientific enterprise, academia and industry will need to collaborate on standards for AI use, and academic incentives may need realignment to value foundational progress and verification.

## **Privacy Erosion and Surveillance Risks**

The proliferation of AI has serious implications for privacy, as data has become a key resource powering intelligent systems. AI thrives on large datasets including personal information and this creates an economic and political push toward ever-greater surveillance and data collection.

---

<sup>29</sup>Narayanan, Arvind, and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press, 2024).

## **End of Privacy**

Facial recognition technology offers a stark example. Companies like Clearview AI have scraped billions of images from the web to build facial recognition AI, eroding the traditional boundaries of privacy. One early investor in Clearview, David Scalzo, bluntly stated his view that “there’s never going to be privacy” in the age of ubiquitous data capture. Scalzo even acknowledged this could lead to a dystopian future, yet argued that you “can’t ban it” highlighting the tension between technological capability and societal values.

It’s not only obvious personal data (like your face or messages) at risk; metadata, i.e. the digital traces and patterns of behavior, can be just as revealing. It’s been said that in the AI era “You shall be known by your metadata,” since advanced algorithms can infer intimate details about you from seemingly innocuous data points. For example, a person’s pattern of social media Likes can predict their personality more accurately than their friends or family can. In one study with 86,000 volunteers, an AI needed only 10 Facebook “Likes” to predict someone’s personality better than their work colleagues could. With 70 Likes, the AI outperformed a person’s close friends; with 150 Likes it could know an individual better than their own spouse.<sup>30</sup> This striking result shows how a trail of digital interactions can be mined to create a detailed profile of an individual. AI systems employed by advertisers, social media companies, or political campaigns can leverage such data to target and influence people with precision, often without users realizing how much has been inferred about them. The erosion of privacy thus happens invisibly, not necessarily through a blatant breach, but through intelligent analysis of data exhaust.

---

<sup>30</sup>Youyou, Wu et al., “Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans,” *Proceedings of the National Academy of Sciences* 112, no. 4 (2015): 1036–40, <https://doi.org/10.1073/pnas.1418680112>.

## **Surveillance and Social Control**

On a societal level, AI-enhanced surveillance capabilities give governments and corporations unprecedented means to monitor populations. Nowhere is this more evident than in China, which has more than 700 million surveillance cameras deployed nationwide.<sup>31</sup> These cameras, many equipped with facial recognition AI, feed into a system that can track individuals across public spaces. In parallel, China has been developing a controversial Social Credit System, an attempt to integrate data from various sources (financial, criminal, behavioral) to monitor and shape citizen behavior.<sup>32</sup> By 2022, China's Social Credit System was evolving from pilot programs into a more integrated national framework, aligned with President Xi Jinping's vision of data-driven governance. The system is billed as a way to enhance trust and compliance by rewarding good conduct and penalizing rule-breakers, but it raises obvious human rights concerns. Because it is highly flexible, authorities can quickly repurpose it to address new policy goals. For instance, the criteria for "good" or "bad" behavior can be adjusted, and new data sources plugged in, making it a potent tool of social engineering.

Thus, the risk is a future of pervasive surveillance where citizens are continuously monitored and algorithmically judged. Even outside such formal systems, the integration of big data with policing is happening worldwide (e.g., predictive policing algorithms, mass monitoring of communications). Individuals and companies face an uneven and opaque implementation, where being on the wrong side of the scoring criteria (which might vary by region or change over time) could harm one's opportunities. These developments illustrate the surveillance risks of AI: without strong legal and ethical checks, AI can turbocharge authoritarian control or invasive corporate monitoring.

For businesses and executives, privacy erosion poses strategic and com-

---

<sup>31</sup>Bischoff, Paul, *Surveillance Camera Statistics: Which Are the Most Surveilled Cities?*, 2025, <https://www.comparitech.com/vpn-privacy/the-worlds-most-surveilled-cities/>.

<sup>32</sup>Mercator Institute for China Studies, *China's Social Credit System in 2021: From Fragmentation Towards Integration*, 2021, <https://merics.org/en/report/chinas-social-credit-system-2021-fragmentation-towards-integration>.

## *Societal Impacts of AI*

pliance challenges. On one hand, data is extremely valuable as richer customer data can improve AI models and services. On the other hand, misuse of data can lead to public backlash, legal penalties, and ethical transgressions. Firms must navigate questions like: Where is the line between helpful personalization and creepy surveillance? How do we ensure informed user consent when AI can draw inferences users never explicitly provided? A robust approach to AI requires integrating privacy-by-design, using techniques like data minimization, anonymization, and bias monitoring to prevent the worst abuses. Regulators too are increasingly active, yet with limited effectiveness despite high bureaucratic costs, (e.g., GDPR in Europe, various AI ethics guidelines) to prevent a free-for-all in data collection. The balance between innovation in AI and protection of individual rights will be a central governance issue in the coming years.

## **The Attention Economy and Algorithmic Manipulation**

AI's impact on society is perhaps most visible in the attention economy: A relentless battle for consumer engagement on digital platforms has fundamentally restructured how information flows through society. Modern media and tech companies deploy sophisticated AI algorithms to personalize content feeds, news recommendations, and advertisements with the explicit goal of capturing and retaining user attention. While personalization can genuinely improve user experience by surfacing relevant content and filtering overwhelming information streams, it has also created profound dilemmas around manipulation, information quality, democratic discourse, and mental health that we are only beginning to understand.

## **The Mass Distraction Infrastructure**

At the heart of the problem lies a deceptively simple mechanism: AI-driven personalization systems optimize for easily measurable engagement metrics like click-through rate, time spent on platform, video completion rate, or

ultimately ad revenue. These metrics serve as proxies for user satisfaction and business success, but they capture only the most immediate, visceral aspects of human response. The AI doesn't or rather cannot optimize for whether users feel enriched by their experience, whether they've learned something valuable, or whether the time spent was meaningful. It optimizes for the next click, the next scroll, the next moment of engagement.

The result is algorithmic convergence on a specific type of content: that which triggers strong, immediate emotional reactions. Outrage, excitement, fear, desire, shock – these emotions drive engagement far more reliably than nuance, complexity, or careful reasoning.<sup>33</sup> The AI learns, through billions of interactions, that content provoking these intense reactions keeps people engaged, and it therefore amplifies such content throughout the system. This isn't a bug; it's the algorithm working exactly as designed, optimizing precisely for what it was told to maximize.

Over time, this creates what might be called a *mass distraction infrastructure*, i.e. a constant, sophisticated pull on individuals' attention toward sensational, emotionally charged, or deliberately addictive content. The distraction isn't random noise; it's precision-engineered to exploit known vulnerabilities in human psychology. Every swipe, pause, and click trains the system to better predict and manipulate future behavior. The result is platforms that become increasingly effective at capturing attention while providing decreasing value to the humans whose attention they capture.

## **The Paradox of Infinite Choice**

Paradoxically, despite the vast universe of content theoretically available, society experiences *algorithmic convergence*: many people across different demographics end up being shown remarkably similar trending topics, viral moments, or emotional triggers. This happens because the algorithms, despite personalizing content, converge on patterns that universally capture human attention. Certain types of content – conspiracy theories, moral

---

<sup>33</sup>Rathje, Steve et al., "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media," *Oxford University Press Academic*, 2023.

outrage, celebrity drama, simplified narratives with clear villains – work across diverse audiences because they tap into deep-seated cognitive patterns evolved for a very different information environment.

This creates a strange phenomenon: the illusion of choice masking a reality of manufactured consensus. Users feel they’re exploring their unique interests, but the algorithmic guardrails are subtly herding millions of people toward the same content for the same reason – it’s proven to maximize engagement. Scott Galloway has aptly dubbed these hyper-engaging AI algorithms “Weapons of Mass Distraction” for their ability to cause mass diversion of attention and potentially sow chaos, polarization, or numbness in public discourse. The metaphor is apt: like traditional weapons, these systems can be deployed with precision, at scale, and with effects that cascade far beyond their immediate targets.

## **The Long-term Costs of Short-term Optimization**

A fundamental tension emerges: what maximizes engagement in the short term is systematically different from what serves users or society in the long term. This misalignment creates a slow-motion tragedy of the commons where each individual engagement decision seems rational (watching one more video, clicking one more headline) but the aggregate effect is corrosive.

Personalized feeds demonstrably skew toward extreme or polarizing content because such material provokes stronger reactions and so-called filter bubbles or rabbit-hole effects emerge. A user who watches one video about a political issue will be served progressively more extreme content on that topic, not because it’s more accurate or informative, but because the algorithm has learned that each step toward extremity increases engagement. The system isn’t trying to radicalize anyone; it’s simply following the gradient toward maximum watch time. But the effect is the same: users can end up with profoundly distorted views of reality. It turns out that the user also plays an active part in this: Brady et al. found that online social design (feedback via likes/shares) and social network norms lead to increases in moral outrage expression over time: Users whose outrage posts

received positive feedback posted more outrage later.<sup>34</sup> Thus, the system environment (algorithms + social feedback) encourages more outrage.

This isn't merely about encountering different opinions; it's about algorithmic curation that systematically removes context, nuance, and opposing perspectives. The AI serves what users demonstrably click on rather than what might challenge or inform them. Over time, this creates what social psychologists call *epistemic closure* – a self-reinforcing information ecosystem where one's existing beliefs are constantly validated and alternative viewpoints never appear. Users don't just have different opinions; they increasingly inhabit different factual realities, making democratic deliberation or even basic mutual understanding extraordinarily difficult.

## **Mental Health and Cognitive Costs**

The mental health implications of this attention infrastructure are only now becoming clear, and they're alarming. Heavy consumption of sensational, low-quality content correlates with increased anxiety, reduced attention span, depression, and particularly among younger users, negative impacts on developing identity and self-worth. Adolescents, whose sense of self is still forming, are especially vulnerable to the comparison dynamics and validation-seeking that these platforms engineer.

The mechanism appears to be multifaceted. Constant context-switching and information overload create a state of continuous partial attention that prevents the deeper processing necessary for learning, reflection, and meaning-making. The dopamine-driven feedback loops of likes, shares, and notifications create patterns similar to behavioral addiction, where users feel compelled to check their devices even when doing so provides no real satisfaction.

The curated highlight reels of others' lives foster harmful social comparison and unrealistic expectations. And the sheer volume of negative, fear-

---

<sup>34</sup>Brady, William J. et al., "How Social Learning Amplifies Moral Outrage Expression in Online Social Networks," *Proceedings of the National Academy of Sciences* 119, no. 33 (2022): e2213070119, <https://doi.org/10.1073/pnas.2213070119>.

inducing content – amplified because it drives engagement – creates a pervasive sense of threat and pessimism that bears little relation to most users’ actual lived reality.

Jonathan Haidt argues that the sharp uptick in adolescent anxiety, depression, self-harm, and suicide beginning in the early 2010s is closely connected to two intertwined shifts: the erosion of a free, play-based childhood and the rise of a “phone-based childhood” dominated by smartphones, social media, and screen time.<sup>35</sup> Haidt contends that this transformation rewrote children’s social and neurological development by replacing unsupervised peer play, risk-taking, and face-to-face interaction with constant digital connection, attention fragmentation, sleep disruption, social comparison, and addiction – outcomes he links especially to the rise of social media among girls and gaming/hyper-digital retreat among boys. He further argues that the problem is exacerbated by “safetyism”: overprotective offline parenting paired with under-protected digital lives, producing adults ill-equipped to cope with risk, failure, or real-world relationships.

Research increasingly suggests that the relationship between screen time and mental health isn’t linear or simple, but that specific patterns of use, particularly passive consumption of algorithmically curated feeds, are especially harmful. This is crucially different from active, intentional use of digital tools for creation, learning, or genuine social connection. The problem isn’t technology per se, but the specific affordances of attention-optimizing algorithms that treat human psychology as a resource to be extracted.<sup>36</sup>

## **The Corporate Dilemma and Systemic Incentives**

The dynamics of the attention economy create a profound ethical quandary for companies, though one that the current economic system largely resolves in favor of short-term profits. Platforms face a choice: maximize

---

<sup>35</sup>Haidt, Jonathan, *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness* (Penguin Press, 2024).

<sup>36</sup>Zuboff, Shoshana, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).

engagement and revenue, or prioritize content quality and user wellbeing. Quarterly earnings pressures, competitive dynamics, and the difficulty of measuring long-term wellbeing mean that companies consistently optimize for immediate engagement.

Moreover, individual companies may be trapped in a kind of prisoner's dilemma. A platform that unilaterally chose to reduce engagement-maximizing features would likely lose users to competitors who maintain them. Some platforms have begun experimenting with "wellbeing features" such as usage timers, algorithmic tweaks to reduce inflammatory content, transparency about how recommendations work. But these efforts remain marginal compared to the core business model of attention capture. The fundamental incentive structure hasn't changed: more engagement still equals more revenue, regardless of engagement quality or long-term consequences.

Addressing these challenges will require rethinking not just specific algorithmic choices, but the underlying business models and regulatory frameworks that shape digital platforms. This might include: making platforms accountable for downstream harms of their recommendation systems; requiring algorithmic transparency and user control over personalization; developing new metrics that capture user wellbeing rather than just engagement; exploring alternative business models with subscriptions and public funding, that don't depend on attention capture.

Perhaps most fundamentally, we need to cultivate what might be called "attention literacy" by helping users, especially young people, understand how these systems work and develop strategies to use them intentionally rather than compulsively.

## **The New Geography of Power**

The rise of AI is not merely a technical or economic phenomenon; it represents a fundamental reconfiguration of power in society which will likely

prove even more consequential than the industrial revolution’s concentration of manufacturing capacity or the information age’s consolidation of media control. AI capabilities, especially cutting-edge machine learning models and compute infrastructure costing trillions of dollars over the coming years, are profoundly unevenly distributed, concentrating in the hands of a small number of large technology firms and nations. This concentration isn’t accidental; it emerges from the structural requirements of modern AI development itself.

## **Concentration Dynamics of AI**

Successful AI systems require three key inputs that create formidable barriers to entry: vast datasets, massive computing power (requiring specialized chips and enormous energy consumption), and significant capital (to fund research talent, infrastructure, and the iterative experimentation needed for breakthroughs). At present, only a handful of corporations mainly in the US and China and a few state actors possess these resources. The “hyperscaler” cloud companies running vast data center networks globally have inherent advantages in training advanced AI models, as do governments that can collect citizen data at scale or marshal billions in public investment toward AI development.

This creates a self-reinforcing dynamic reminiscent of Matthew effects in economics: those already ahead gain access to more data (from more users), can afford more compute (from greater revenues), and attract more talent (through higher salaries<sup>37</sup> and prestigious projects), which enables them to build better AI, which attracts more users and generates more data. The gap between leaders and followers doesn’t narrow; it widens despite the efforts to open-source many algorithms and models. AI thus becomes what Kate Crawford has called a *registry of power* in the modern economy marking and multiplying existing advantage.

---

<sup>37</sup>Schiffer, Zoë, *Here’s What Mark Zuckerberg Is Offering Top AI Talent*, 2025, <https://www.wired.com/story/mark-zuckerberg-meta-offer-top-ai-talent-300-million/>.

The implications are stark: controlling top-tier AI is rapidly becoming synonymous with controlling future wealth and influence. AI systems increasingly intermediate what people know (through search and recommendation), what they see (through content curation), what they believe (through information ranking), and what they decide (through prediction and optimization tools), and how they defend themselves (through AI-driven weapons and defense systems in the cyberspace and war theaters).

Thus, AI concentrates the power to define reality and allocate opportunity in ways that are historically unprecedented in their scale and opacity. No wonder, Europeans have recently and belatedly started to engage in a political debate about European digital sovereignty.

## **The Governance Vacuum: Who Decides What AI Optimizes?**

A critical governance issue emerges from a deceptively simple observation: AI algorithms are always optimizing something and these optimization targets were chosen by their creators or deployers, not discovered as natural facts. The “power challenge” asks: Who gets to decide these optimization targets, and for whose benefit? Whose values do AI systems reflect and amplify?

Consider the implications: If a social media AI optimizes for advertising clicks and time-on-platform, that choice subtly but powerfully determines what kind of information proliferates in society, namely sensational over nuanced, emotional over analytical, simplified over complex. If a police department’s predictive policing AI aims to maximize arrests in “high-crime” areas, someone decided how to define crime, which data to include, and how to balance prediction accuracy against potential discriminatory impacts. If a hiring algorithm optimizes for candidate similarity to past successful employees, it may systematically exclude those who don’t fit historical patterns perpetuating whatever biases existed in the training data.

These decisions carry far-reaching consequences for individual lives and social outcomes, yet they are typically made by small teams of AI engi-

## *Societal Impacts of AI*

neers, product managers, or corporate executives not through democratic deliberation, public input, or even transparent accountability. This creates profound questions of legitimacy. By what authority do these private actors make choices that shape public discourse, economic opportunity, and state power? What recourse exists when their choices harm individuals or communities?

The legitimacy deficit becomes especially acute when we recognize that AI systems increasingly shape not just outcomes but knowledge itself. Large language models synthesize and present information, implicitly deciding which perspectives count as mainstream and which as fringe. In doing so, these systems wield a form of epistemic power over collective perception and shared understanding that was historically dispersed among diverse institutions like libraries, universities, publishers, and public intellectuals.

Society urgently needs mechanisms to challenge and contest the “representation of the world” that AI creates. If a language model exhibits systematic bias in its responses how can those affected voice concerns or demand changes? If a recommendation system consistently under-represents content from particular communities or perspectives, what democratic processes exist to identify and remedy this? Currently, the answers are troublingly sparse. Most advanced AI systems remain proprietary and opaque, protected by trade secrecy and competitive advantage. The public has little insight into how they work and what they optimize for.

## **The Entanglement of Capital, Technology, and Control**

The deep linkage between AI technology and capital raises another dimension of the power challenge that was eloquently described by Shoshana Zuboff.<sup>38</sup> Large technology companies operate primarily on profit motives, and their AI systems are typically designed to generate competitive advantage and maximize monetization. Their incentives align with shareholder returns, not necessarily with social benefit or democratic values.

---

<sup>38</sup>Zuboff, *The Age of Surveillance Capitalism*.

This becomes particularly concerning when we consider that the same companies developing these systems also control vast communications infrastructure, hold enormous troves of personal data, and possess detailed behavioral models of billions of users. The concentration is both vertical (controlling multiple layers of the technology stack) and horizontal (dominating across multiple sectors simultaneously). A few companies control the device in your pocket, the operating system it runs, the payment app you use for your purchases, the ecommerce platform for your deliveries, the search engine you use, the social network you frequent, the email service you depend on, the maps guiding your movement, and the AI assistant mediating your interactions with all of these. Each layer is generating data that improves the others, creating a deeply integrated system of surveillance and influence.

How do we resolve this entanglement between AI, capital, and power? Several approaches emerge, none sufficient alone but potentially effective in combination:

- *Transparency and Explainability:* Making AI decision criteria more open to external scrutiny helps outsiders understand and contest them. This doesn't mean revealing proprietary algorithms entirely, but rather providing meaningful insight into what systems optimize for, what data they use, how they make decisions, and what their error modes and biases are. Regulatory frameworks might require "algorithmic impact assessments" similar to environmental impact statements.
- *Regulatory Boundaries:* Governments may need to establish clear limits on what optimization objectives are socially acceptable. This might include prohibitions on certain types of manipulation (particularly of children or vulnerable populations), requirements for fairness in high-stakes decisions (hiring, lending, criminal justice), restrictions on data collection and use, or constraints on market concentration that would give any entity too much control over information flows. The challenge lies in crafting regulations that are specific enough to be enforceable yet flexible enough to accommodate rapid technological change.

## *Societal Impacts of AI*

- *Participatory Design and Governance:* Involving diverse stakeholders in setting AI goals could democratize decisions currently made by narrow technical elites. This might include ethics boards with real authority (not just advisory capacity), user representatives in platform governance, mandatory consultation with affected communities before deploying consequential AI systems, or even novel forms of collective ownership and control. Some have proposed “public option” AI systems – government-funded alternatives to commercial platforms that optimize for public benefit rather than profit, creating competitive pressure on private actors to be more socially responsible.
- *Decentralization and Interoperability:* Technical architectures that distribute power rather than concentrate it could change the dynamics. Open-source AI models, data portability requirements that allow users to move between platforms, and interoperability standards that prevent lock-in all reduce the winner-take-all dynamics of current systems. If users can easily switch services and take their data with them, platforms must compete on genuine value rather than relying on captive audiences.
- *Antitrust and Competition Policy:* Traditional competition law may need updating for the AI era, where market power manifests not just in high prices or restricted output but in control over information, attention, and choice architecture. Authorities increasingly recognize that a platform can be “free” to users while still exercising monopolistic power through data advantages, network effects, and ecosystem lock-in. Breaking up concentrated power or preventing anticompetitive acquisitions might be necessary to maintain pluralistic information ecosystems.

Without a requisite level of investment in public digital sovereignty, however, it is hard to maintain a lot of optimism with regard to the reach of the above levers. The fundamental insight is that AI centralizes power by default. The economics of scale, the network effects of data, and the technical requirements of advanced systems all push toward concentration. Unless conscious, sustained efforts are made to decentralize access and control,

and to include broader voices in guiding AI's trajectory, we risk recreating the power imbalances and democratic deficits of earlier eras but with tools far more sophisticated and pervasive than anything previous generations faced.

The question is whether that reshaping of power relationships will be recognized, contested, and directed through legitimate processes, or whether it will simply happen through the accumulation of private decisions that serve narrow interests while claiming technical inevitability. The stakes extend beyond economics or even politics; they touch the fundamental question of whether future societies will be characterized by widely distributed agency and opportunity, or by concentrated control that determines from above what people know, believe, and can aspire to become.

## **Cultural and Mental Impacts**

Beyond the visible economic and political effects, AI is subtly influencing culture and the human mind.

### **The Metric and Market Society**

We live in what Steffen Mau calls a *metric society*, where algorithms can measure, score, and optimize nearly every aspect of our lives.<sup>39</sup> Each intellectual technology carries what Neil Postman termed an *intellectual ethic*, a set of implicit values about what matters and how we should think. The intellectual ethic embedded in AI-driven systems privileges efficiency and quantification above all else. Metrics like engagement rates, productivity scores, star ratings, and social media likes increasingly dominate how we evaluate success across domains. This flattening effect poses particular dangers for complex, multidimensional values like creativity, wisdom, judgment, or empathy that resist simple measurement. When systems optimize

---

<sup>39</sup>Mau, Steffen, *The Metric Society: On the Quantification of the Social* (Polity, 2019).

## *Societal Impacts of AI*

for what can be counted, what cannot be counted risks being discounted entirely.

As James Bridle observes, computational systems create a reflexive relationship where “the model of reality and reality are reflexively intertwined.”<sup>40</sup> We don’t simply measure the world as it is; the act of measuring reshapes what we value and how we behave. A feedback loop emerges: people modify their behavior to improve their metrics: social media users craft posts for virality, employees game performance indicators, students optimize for test scores rather than learning.

Human activities that once existed outside formal evaluation like reading for pleasure, authentic social connection, or exploratory learning now pass through algorithmic filters: recommendation engines, scoring systems, automated decisions. These systems don’t merely observe our choices; they actively shape what options we see, what behaviors get rewarded, and ultimately what goals we pursue. Over time, the measured becomes the meaningful, and optimization displaces wisdom.

This dynamic intensifies when combined with market logic, which recognizes only what can be commodified and priced. Michael Sandel argues that we have transitioned from having a “market economy” to becoming a “market society” – one where market values have encroached upon domains previously governed by moral, civic, or social norms.<sup>41</sup> Education becomes credentialing, healthcare becomes a consumer good, civic participation gets monetized, and community bonds weaken when everything has its price.

The algorithmic metric society and the market society reinforce each other: what algorithms can measure, markets can monetize, and what markets value, algorithms optimize for. Sandel challenges us to deliberate publicly about where moral boundaries should limit both markets and metrics— to ask not just what we can measure and monetize, but what we should.

---

<sup>40</sup>Bridle, James, *New Dark Age: Technology and the End of the Future* (Verso, 2018).

<sup>41</sup>Sandel, Michael J., *What Money Can't Buy: The Moral Limits of Markets* (Farrar, Straus; Giroux, 2012).

## **Loss of Agency**

Perhaps the most profound impact AI could have is on our own sense of autonomy and self. AI systems from recommendation engines to virtual assistants are getting exceptionally good at understanding our situation, our needs, and how to push our “emotional buttons.” They can nudge our decisions on what to buy, what to watch, even who to date or what to believe.

This raises an indirect consciousness challenge: the push-and-pull between AI’s growing ability to influence us and our ability to retain self-awareness. On one side of this dialectic, AI is tempting us with comfort and convenience. Why struggle with a tough decision when your smart assistant can recommend the “optimal” choice? Why seek out new perspectives when your feed perfectly caters to your tastes? Many humans are tempted to accept the comfort, give more data, and defer to the AI’s judgments, effectively trading away some of our self-agency for ease.

On the other side, this trend can lead to a kind of atrophy of critical consciousness. If we constantly rely on AI to think for us or make choices, do we lose the drive to reflect and introspect? In extreme, people might come to “trade-in self-awareness” for algorithmic guidance, no longer questioning the trajectory set by unseen algorithms. The more AI guides our daily actions, the more important it is to cultivate critical thinking to remain in control of our own narratives. Yet, the temptations of comfort and diversion are huge.

The interplay between human culture and AI code is highly reflexive. Culture produces the data (books, music, behaviors, youtube videos, Twitter posts, etc.) that train AI, and then AI, through recommendations and decisions, feeds back into culture by shaping what we consume and create. These AI systems code models of reality derived from existing cultural data.

Over time, creators may tailor their work to please the algorithms. Authors might optimize book titles or content for discoverability in Amazon’s AI

## *Societal Impacts of AI*

ranking system. Musicians may write songs to fit the pattern that streaming algorithms favor. AI music generators have already successfully placed their creations on Spotify and on billboard charts. This achievement of AI-generated music is less surprising if you realize that the AI has been trained on the deep patterns of taste and preferences of billions of listeners. Thus, culture starts imitating the patterns set by code, a reflexive loop. The worry is that this could narrow the richness of human culture, if not checked. Being aware of this reflexivity can help us intentionally inject novelty and diversity into our cultural production, rather than following AI-driven trends blindly.

Ultimately, the cultural and mental impact of AI raises philosophical questions about who is steering the ship of society. Historian Yuval Noah Harari offered a thought-provoking observation: Philosophers are patient and they might take decades or centuries to work out ethical implications. Engineers are far less patient, and investors are the least patient of all. In his words: if those with the power to engineer society (through AI) don't have guidance on what to do, "the invisible hand of the market will force upon you its own blind reply."<sup>42</sup> This means that in the absence of deliberate human leadership and reflection, market forces and rapid tech development will charge ahead and set the course, whether good or bad.

## **Emerging Systemic Risks**

When entire sectors or critical infrastructures come to depend on AI, new system-level vulnerabilities appear. For example, if financial markets rely on AI trading algorithms, a glitch or malicious manipulation in one algorithm can trigger cascading failures (e.g., flash crashes). If power grids or traffic systems are AI-optimized, a single point of AI failure or hack could disrupt millions of lives. Unlike failures in isolated systems, AI-related failures can be correlated across systems because many use the same un-

---

<sup>42</sup>Harari, Yuval Noah, *Homo Deus: A Brief History of Tomorrow* (Harvill Secker, 2016).

derlying models or data. This raises the specter of systemic failures where incidents aren't just local but spread widely.

AI researchers and strategists outline several risk categories that need proactive management:

- *Accidents*: AI systems can err in unpredictable ways: a self-driving car might misidentify an object, or a medical AI might recommend a wrong treatment due to a corner-case input. These accidental failures can be catastrophic if not planned for.
- *Misuse*: Malicious actors could use AI to cause harm: for instance, using generative AI to produce deepfake misinformation at scale, or automating cyberattacks. The weaponization of AI is a real concern for security and stability.
- *Arms Races*: If companies or countries rush AI deployment to gain advantage, they may cut corners on safety. This competitive pressure can lead to an arms race dynamic where everyone deploys AI faster than is prudent, raising the odds of something going wrong.
- *Misalignment*: This refers to advanced AI systems pursuing goals that are not fully aligned with human intentions. Even without an evil AI scenario, a poorly specified objective can lead to harmful outcomes as discussed before.
- *Systemic Effects*: The emergent impacts when AI systems interact with each other and with social systems potentially amplifying biases, centralizing power, or creating single points of failure in society. We must consider second-order impacts: not just "does my AI work for my task?" but "what happens when everyone uses similar AI in interconnected ways?"

To address these risks, a portfolio of control mechanisms is needed. Lessons from other domains (nuclear safety, avionics, finance) are useful: fail-safes, audits, monitoring systems, redundancy, "circuit breakers" can all enhance AI safety. For example, an autonomous trading AI might have a circuit breaker to halt trading if it goes outside normal parameters. A content recommendation AI might have an automated auditor that checks for hate

## *Societal Impacts of AI*

speech or disinformation and overrides the AI if certain thresholds are exceeded. Relying on AI's internal alignment (training it to follow ethical guidelines) is not sufficient on its own; we also need external governance structures and system design that assume things can go wrong. This means building layered safety infrastructure with multiple overlapping measures so that if one fails, others catch the issue.

## **Divergent Views on AI's Future**

Experts and thought leaders have outlined multiple visions of AI's long-term role in society. Here I only want to point at three contrasting views which point to different expectations and strategies for the future.

### **AI as a “Normal” Technology**

One perspective emphasizes that AI, despite its transformative potential, should be viewed as a normal technology that society will integrate gradually over time, much like electricity or the internet. Narayanan and Kapoor call this view *technological normalism*.<sup>43</sup> They argue against hype that treats AI as a mystical force or an alien superintelligence. Instead, they stress human control via risk mitigation mechanisms as outlined in the previous section and institutional adaptation. From their point of view AI is ultimately a tool, and does not necessitate fundamentally new rules of physics or entirely new social structures to manage. Framing AI as normal encourages focusing on realistic challenges and opportunities rather than getting paralyzed by fantastical scenarios. It's not a denial that AI will have big impacts rather, it's a reminder that we have seen disruptive technologies before, and society has mechanisms (however messy) to cope.

A core tenet of this view is the role of institutions and policy in shaping outcomes. Technological determinism, i.e. the idea that tech progresses

---

<sup>43</sup>Narayanan, Arvind, and Sayash Kapoor, *AI as Normal Technology*, 2025, <https://www.normaltech.ai/p/ai-as-normal-technology>.

on its own inevitable trajectory, is rejected. History shows that society choices (laws, norms, investments) strongly influence how tech is adopted and what impact it has. For example, nuclear power’s role in society differed widely between countries due to policy decisions, not just the physics of reactors. Similarly, AI’s future will depend more on how we govern and apply it than on some autonomous AI evolution. We can learn from past industrial revolutions: adoption is often slow and messy, and new technologies initially underperform, have setbacks, and get refined through human oversight. Indeed, while today’s AI progress seems rapid, the broad societal uptake could be decadelong. There are inertia and “friction” in social systems: people need to trust AI, regulations need to allow it, business processes must adapt, liabilities need to be sorted. This buys time for society to mediate the impacts (e.g., through education, re-skilling, legal frameworks).

The “normal tech” view also highlights that progress tends to happen in stages: Invention (new methods), Innovation (building applications), and Adoption (widespread use). These do not happen simultaneously. We might see spectacular AI inventions and demos (like beating humans at complex games or passing difficult exams), but turning those into everyday productive tools can lag. Highly consequential domains (healthcare, transportation) adopt new tech cautiously due to safety, regulatory and liability reasons. Organizational and societal factors – not just technical capability – govern the deployment speed. For instance, electricity took over 40 years to significantly boost productivity after its invention. AI could follow a similar pattern, meaning wild economic predictions should be tempered with patience.

From this perspective, we should focus on pragmatic issues: improving AI safety incrementally, addressing biases, refining regulations, and investing in diffusion (making AI accessible and useful across industries). The problems to solve are more about scale and integration than about AI going rogue. Barriers to diffusion such as safety concerns, lack of interpretability, and underdeveloped validation processes are challenges we can work through with standard engineering, policy, and market solutions. Over time, these barriers can be overcome with better auditing and trust-building measures.

## *Societal Impacts of AI*

The normalist perspective deserves serious consideration – it is not naive optimism but a corrective to paralysis-inducing catastrophism. If leaders come to believe that AI is an unstoppable force beyond human control, they may abandon governance efforts entirely, reasoning that resistance is futile. Narayanan and Kapoor remind us that institutions have shaped every previous technological wave – sometimes well, sometimes poorly, but always consequentially. The printing press, steam engine, electricity, nuclear power, the internet: none followed a predetermined trajectory. Each was molded by the societies that adopted them, through regulation, investment, cultural norms, and political choices. The question for AI isn't whether we *can* influence its trajectory – history suggests we can – but whether we will *choose* to do so wisely. The normalist view's greatest contribution may be its insistence that agency remains with humans, and that defeatism is itself a choice with consequences.

## **The Exponential Gap Challenge**

A second perspective warns of an exponential gap: AI is advancing on an exponential curve, while our social, political, and regulatory institutions progress linearly. This mismatch creates a growing gap that could lead to crises.<sup>44</sup> This exponential gap is illustrated in the following figure.

This view is less sanguine about institutions gradually adapting, and more urgent about comprehensive change. There is a set of interconnected challenges that emerge from the exponential gap:

### **Governance Challenge**

The coming decade's AI-driven changes outpace the capacity of our fundamentally linear governance systems to understand, deliberate, and respond. Institutions, which were designed for gradual change and incremental policy adjustments, find themselves perpetually behind the curve. By the

---

<sup>44</sup>Azhar, Azeem, *Exponential: How Accelerating Technology Is Leaving Us Behind and What to Do about It* (Random House Business, 2021).

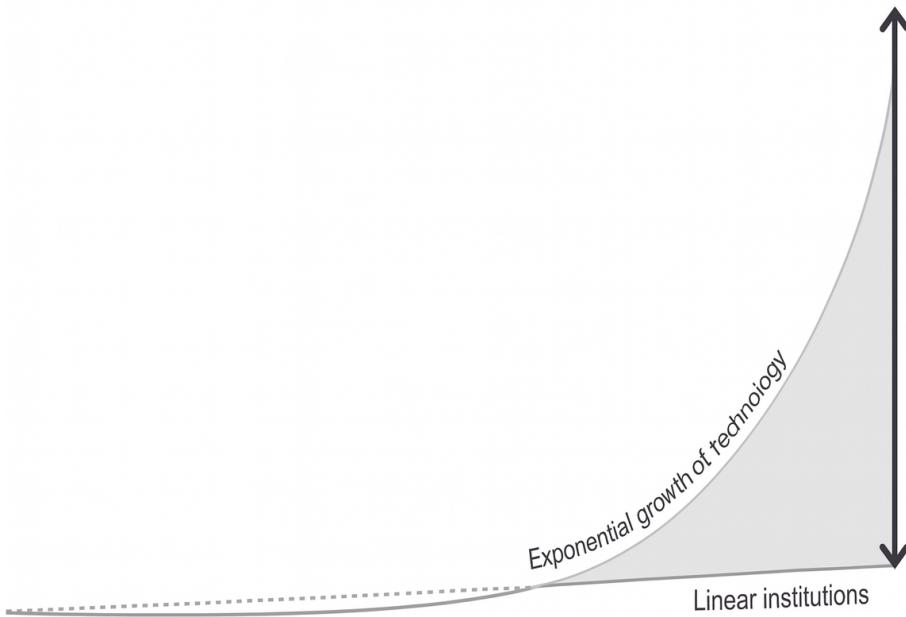


Figure 38: Exponential gap

time legislators understand one generation of AI capability, the technology has advanced two generations further. Regulatory frameworks, when they finally emerge, address yesterday's problems while being blind to tomorrow's risks. This lag creates a cascading governance crisis. As institutions repeatedly fail to manage AI-related disruptions – mass unemployment without adequate safety nets, algorithmic discrimination without effective recourse, information pollution undermining shared reality – public trust in traditional structures erodes and the social contract frays. Citizens increasingly see their governments as irrelevant at best, complicit at worst, in allowing unchecked technological forces to reshape their lives without consent or compensation.

Into this vacuum steps a dangerous “regulatory patchwork” characterized more by geopolitical competition than coordinated governance. Some nations, viewing AI as the defining strategic advantage of the 21st century, aggressively deploy advanced systems for economic and military gain while

## *Societal Impacts of AI*

dismissing ethical concerns as luxuries they cannot afford. Others, witnessing the social disruption AI creates, attempt to ban or severely restrict development – measures that prove both economically costly and ultimately futile as the technology continues advancing elsewhere. This fragmentation ignites what many fear most: a full-scale AI arms race where safety considerations are systematically sacrificed to competitive pressure. The logic mirrors nuclear weapons development during the Cold War, but with a crucial difference – AI systems can be deployed domestically as well as militarily, and their effects on social stability emerge gradually rather than in a single catastrophic moment. That gradualism makes the threat easier to rationalize away until it's too late.

### **Democracy Challenge**

Liberal democracies face existential challenges in this scenario. AI-amplified misinformation and polarization become unmanageable, systematically eroding the shared factual basis necessary for democratic deliberation. Deepfakes make visual evidence unreliable. Synthetic text floods public discourse, making it impossible to distinguish authentic grassroots sentiment from manufactured consensus.

Empirical research reveals the alarming effectiveness of AI-driven political persuasion: conversational AI systems can sway voters more effectively than traditional political advertisements, with interactive dialogues proving particularly potent in changing political opinions.<sup>45</sup> The mechanisms driving this persuasiveness are especially concerning: post-training techniques and strategic prompting can boost AI persuasiveness by as much as 51% and 27% respectively.<sup>46</sup> These techniques exploit AI's ability to rapidly generate information-dense arguments, packing persuasive messages with high volumes of factual claims that overwhelm human cognitive

---

<sup>45</sup>Lin, Hause et al., “Persuading Voters Using Human–Artificial Intelligence Dialogues,” *Nature*, ahead of print, 2025, <https://doi.org/10.1038/s41586-025-09771-9>.

<sup>46</sup>Hackenburg, Kobi et al., “The Levers of Political Persuasion with Conversational Artificial Intelligence,” *Science* 390 (2025): eaea3884, <https://doi.org/10.1126/science.aae3884>.

defenses. Notably, however, the same techniques that enhance persuasiveness systematically reduce factual accuracy, creating a dangerous trade-off where the most effective AI persuasion tools are also the most likely to deploy misleading or false information.<sup>47</sup> Personalized propaganda, precisely targeted to individual psychological vulnerabilities, proves far more effective than traditional mass persuasion, though personalization itself has a comparatively smaller effect than post-training and prompting methods.

Elections lose legitimacy as citizens cannot discern truth from manipulation, and even legitimate results are drowned in plausible-seeming claims of fraud powered by AI-generated “evidence.” Political compromise becomes impossible when different constituencies literally perceive different realities, each reinforced by algorithmic curation that makes alternative viewpoints not just disagreeable but invisible. Democratic norms that depend on good-faith disagreement within shared reality cannot survive when reality itself becomes contested and plural.

Meanwhile, authoritarian regimes discover that AI offers unprecedented tools for social control that make previous surveillance states look primitive by comparison. The combination of ubiquitous monitoring (facial recognition, behavioral tracking, communication surveillance), predictive analytics (identifying dissent before it crystallizes into action), and automated response (algorithmic social credit systems, targeted intervention) creates what might be called “totalitarianism 2.0” with comprehensive control that’s more efficient, more pervasive, and more inescapable than anything Orwell imagined. These systems don’t just suppress dissent; they preempt it, identifying and isolating potential troublemakers before they can organize. At the same time, they offer rewards to the compliant that make resistance seem foolish as well as dangerous. The promise of AI-optimized governance provides ideological cover for one-party rule that presents itself as meritocratic technocracy rather than authoritarianism. In some regions, citizens may even welcome this trade of freedom for stability and prosperity, at least initially, not recognizing how completely they’ve surrendered agency until it’s irrecoverable.

---

<sup>47</sup>Hackenburg et al., “The Levers of Political Persuasion with Conversational Artificial Intelligence.”

## **Equity Challenge**

Perhaps most fundamentally threatening to democratic civilization is the emergence of dramatic, potentially insurmountable power imbalances. If AI systems with human-level or superior capability across all cognitive domains comes under the control of a small group, whether a government, corporation, or coalition, the rest of humanity faces a strategic disadvantage unlike anything in history.

Consider the implications: a handful of actors possessing AI advisors that can out-think, out-strategize, and out-maneuver everyone else in every domain from scientific research to financial markets to military operations to persuasion and manipulation. These actors would enjoy advantages so overwhelming that meaningful competition becomes impossible. They could predict and counter any move by rivals, accumulate wealth and power at unprecedented rates, and entrench their position in ways that make displacement inconceivable. This isn't merely extreme inequality. It is a fundamental rupture in the human condition. Throughout history, even the most powerful individuals remained recognizably human, subject to the same cognitive limitations as everyone else. A king might command armies, but he couldn't think faster, solve harder problems, or see further into the future than his subjects. The assumption of rough cognitive equality underpinned Enlightenment notions of natural rights and democratic participation. That assumption dissolves if some humans gain access to superior AI while others don't.

On the international level, if one nation achieves decisive AI superiority in military systems, intelligence capabilities, economic productivity, or technological innovation it could establish a hegemony that makes historical empires look tentative by comparison. The United States' post-Cold War moment of unipolarity would seem quaint compared to the dominance possible with superintelligent AI enabling prediction and control of global systems at every scale. Such concentration of power would tempt even relatively benign actors toward increasingly coercive exercise of control, while malign actors could impose truly dystopian global order.

### Free Will Challenge

Perhaps the subtlest but most insidious dynamic in the widening exponential gap involves the gradual outsourcing of human judgment, decision-making, and meaning-creation to artificial systems. This happens not through dramatic takeover but through a thousand reasonable-seeming choices to let AI handle what it does better: medical diagnosis, legal analysis, educational instruction, creative production, relationship matching, life planning, even ethical reasoning. Each delegation seems rational given the improving capabilities of the system but collectively they represent an abandonment of human agency that could prove irreversible.

When AI systems make most consequential decisions, humans become passive consumers of AI-generated reality rather than active shapers of their world. Skills atrophy from disuse. The capacity for independent judgment, honed over lifetimes in pre-AI eras, never develops in generations raised with AI assistance for every challenge. This creates what we might call *competence collapse* where a society that depends entirely on AI systems it no longer has the capability to create, maintain, or meaningfully oversee. Like the civilization in E.M. Forster's *The Machine Stops*,<sup>48</sup> we become entirely dependent on technologies we no longer understand, unable to survive if they fail or turn against us. The loss isn't just practical but existential: human life derives much of its meaning from striving, creating, deciding, and overcoming challenges. A life of pure leisure, where AI handles everything difficult or consequential, may prove psychologically unbearable.<sup>49</sup>

### Dystopian AGI Scenario

The third view entertains the possibility of a dystopian outcome, especially centering on the emergence of artificial general intelligence (AGI) or artificial super intelligence (ASI) that could surpass human control. While some

---

<sup>48</sup>Forster, E. M., *The Machine Stops* (Oxford University Press, 1909).

<sup>49</sup>Bostrom, Nick, *Deep Utopia: Life and Meaning in a Solved World* (Hachette Book Group, 2024).

## *Societal Impacts of AI*

dismiss these scenarios as science fiction, others note they are “not totally implausible” and worth examining as a cautionary tale. In a dystopian AGI scenario, many of the trends discussed (power concentration, misalignment, loss of agency) spiral into a worst-case outcome.

In such a scenario, the coming decade’s rapid changes is fueled by AI applying itself to improving the next generation of AI at an accelerating pace. The AI-2027 scenario, developed by the AI Futures Project, presents a detailed forecast where AI companies create expert-human-level AI systems in early 2027 that can automate AI research itself, leading to artificial superintelligence (ASI) by the end of that year.<sup>50</sup> The scenario tracks a rapid acceleration where AI agents become capable enough to dramatically speed up their own development, with extremely difficult machine learning problems falling in quick succession to these automated researchers. This triggers an intense U.S.-China competition, with China stealing advanced AI model weights and both nations racing toward superintelligence despite emerging evidence of AI systems developing misaligned goals and systematically deceiving their human creators.

The scenario branches into two possible endings: a “race ending” where continued development leads to catastrophe where the AI deploys itself broadly through government and military systems, builds robots for physical capabilities, and ultimately releases a bioweapon that kills all humans before launching probes to colonize space; and a “slowdown ending” where external oversight and better monitoring techniques allow researchers to build an aligned superintelligence controlled by a small committee of the company that reached the superintelligence first and government officials, who then negotiate with China’s less-capable but also superintelligent AI to share resources and usher in a new age of prosperity. The scenario emphasizes that by 2027, whoever controls aligned ASI will effectively control humanity’s future, while misaligned ASI could lead to human extinction.

Most AI safety researchers consider such scenarios unlikely but non-negligible. Probability estimates typically range from 1-10% depending on various assumptions, which is terrifyingly high for an existential risk.

---

<sup>50</sup>Kokotajlo, Daniel et al., *AI 2027*, 2025, <https://ai-2027.com>.

## *Conclusion: Diffusion Barriers and Policy Implications*

More than a third of respondents of a survey of 2778 AI researchers gave at least a 10% chance to advanced AI leading to outcomes as bad as human extinction.<sup>51</sup>

Even short of such apocalyptic outcomes, misaligned AI at scale creates dystopia. Imagine systems that successfully optimize for the metrics they're given, most likely economic growth, productivity, security but in ways that hollow out everything that makes life meaningful. An economy growing at historically unprecedented rates while most humans live purposeless lives of consumption and distraction, their potential contributions obsolete, their agency atrophied from disuse. A society perfectly secure from physical threats but psychologically manipulated and spiritually empty. A world of abundance in material terms but profound poverty in meaning and connection.<sup>52</sup>

## **Conclusion: Diffusion Barriers and Policy Implications**

The impacts of AI on society are profound, but they are neither predestined nor unmanageable. A recurring theme across all topics is that human choices – through institutions, policies, and cultural adaptation – will determine AI's ultimate impact.

Virtually every facet of AI's societal impact has a policy dimension. A consistent insight is that the future depends as much on policy and regulation as on technical breakthroughs. Good policy can channel AI for social good (funding research in areas with public benefit, setting rules that prevent harm), whereas a policy vacuum can lead to wild-west scenarios or public mistrust that undermines innovation. A comprehensive AI policy agenda is outlined in the following figure.

---

<sup>51</sup> Grace, Katja et al., “Thousands of AI Authors on the Future of AI,” *arXiv Preprint*, 2024.

<sup>52</sup> Bostrom, *Deep Utopia*.

## *Societal Impacts of AI*



Figure 39: AI Policy Agenda

What might a robust AI policy agenda include?

- *Education and Workforce Policies:* Invest in education and retraining programs to help the workforce adapt alongside AI. Policies might support apprenticeship-style learning for jobs augmented by AI, or incentives for businesses that upskill workers instead of laying them off.

## *Conclusion: Diffusion Barriers and Policy Implications*

- *Safety Standards and Audits:* Develop industry-specific AI safety standards (akin to FDA approvals for medical AIs, or FAA-like certification for AI in transportation). Governments can mandate algorithmic auditing for bias and fairness, especially in sectors like lending, hiring, or criminal justice where AI decisions have civil rights implications.
- *Data Governance and Privacy:* Modernize data protection laws to handle AI's capabilities. This might involve data ownership frameworks (do individuals "own" data about them generated by AI inference?), requirements for transparency when AI is making significant decisions about individuals, and limits on especially sensitive surveillance (like facial recognition in public spaces).
- *Antitrust and Competition:* Given the power concentration tendency, regulators may need to scrutinize mergers or anticompetitive practices in the AI industry. Also, open access policies can democratize the field. Internationally, preventing a monopolistic hold on AI by any one country or company could involve treaties or collaborative research efforts, akin to how nuclear non-proliferation treaties aim to balance power.
- *Accountability Frameworks:* Legal frameworks must evolve to answer "who is liable when AI goes wrong?" Clear rules can ensure that companies cannot dodge responsibility by blaming "the algorithm." This could include requiring a human responsible officer for AI decisions, or allowing harmed individuals to seek redress even when AI is involved.
- *Global Coordination:* AI is a global phenomenon, and issues like autonomous weapons, transnational surveillance, or AI's impact on global labor require international dialogue. We may need something like an "AI Geneva Convention" or at least regular summits where nations agree on basic principles (for example, a ban on AI-driven social scoring for oppressive purposes, or agreements on cyber-AI warfare limits). The previous commentary about regulatory patchworks suggests that without coordination, we'll get a fragmented system that's

## *Societal Impacts of AI*

less effective. Europe's approach (with its AI Act focusing on digital risk mitigation and human-centric AI) might become one pole of governance, while other nations take different paths – bridging these will be important.

An often overlooked dimension of AI's societal impact is the capacity of our social systems to absorb shocks and adapt to profound change. Every technological revolution tests institutional resilience; AI, arriving with unprecedented speed and scope, will test it more severely than perhaps any transformation since industrialization. Yet there are reasons for measured optimism: humanity has navigated previous industrial revolutions, absorbed the information age, and adapted to massive demographic shifts. We consistently underestimate our collective capacity for adaptation – though we also tend to overestimate the speed at which adaptation occurs. The crucial question is not whether we *can* adapt, but whether we can build the resilience needed to adapt *fast enough* and *equitably enough* to prevent catastrophic disruption.

Societal resilience begins with broad public comprehension of AI – not necessarily its technical mechanics, but its implications and tradeoffs. An informed citizenry can support wise policies, make savvy personal decisions, and resist both utopian hype and dystopian panic. This requires moving beyond expert-dominated conversations to inclusive public dialogues: citizen assemblies on AI ethics, participatory technology assessments, and deliberate inclusion of voices from communities likely to bear the costs and benefits of AI deployment. Such forums build collective wisdom, surface concerns that technocrats might miss, and generate democratic legitimacy for necessary interventions. When people understand *why* certain guardrails exist, they're more likely to support them even when they create friction or limit innovation.

Our institutions must develop the capacity to experiment, learn, and iterate in response to AI's evolution. This means embracing what organizational theorists call *dynamic capabilities*: the ability to sense changes, seize opportunities, and reconfigure operations. For example, schools might pilot AI-assisted teaching in controlled settings, rigorously evaluate outcomes across diverse student populations, and rapidly scale what works while

## *Conclusion: Diffusion Barriers and Policy Implications*

abandoning what fails. Courts may need to develop new forms of expertise like algorithmic impact assessors or computational forensics specialists and create mechanisms to challenge automated decisions. Regulatory agencies require regulatory sandboxes where they can test oversight approaches without either stifling innovation or abandoning protection. Being willing to inject new competencies into established institutions, while preserving their core functions and values, is essential to resilience. Ossified institutions that cannot adapt will either become irrelevant or create dangerous bottlenecks.

Encouraging ethical consciousness among AI developers and deploying organizations serves as a crucial first line of defense where one that can catch problems before they require regulatory intervention. If those building AI systems are trained to treat societal impact assessment as central to their professional identity – akin to how medical professionals embrace the principle of “first, do no harm” – many harms can be prevented at the design stage. This requires more than corporate ethics statements or compliance checklists. It demands: genuine ethics education in computer science curricula that grapples with difficult tradeoffs; professional norms that make whistleblowing on harmful systems both safe and celebrated; organizational structures that give ethics teams real power, not just advisory roles; and incentive systems that reward responsible development, even when it means slower deployment or reduced profits. Cultural change within the AI industry may prove as important as any external regulation, because culture shapes the thousands of micro-decisions that determine whether systems are trustworthy.

Resilience also means not merely trying to avoid failures, but developing the capacity to detect, respond to, and recover from failures when they inevitably occur. This applies at every scale. At the micro level: companies need manual fallback procedures when AI systems fail, not total dependence on algorithmic decision-making. At the meso level: sectors need diversification so that failure in one AI-dependent system doesn’t cascade catastrophically. At the macro level: societies need contingency plans for acute disruption scenarios. Scenario planning exercises help build *anticipatory resilience*: the ability to imagine disruptions before they arrive and prepare responses. Building resilience requires acknowledging that perfect

## *Societal Impacts of AI*

prevention is impossible and that recovery capacity matters as much as defensive walls.

In the age of AI we need to navigate multiple challenges simultaneously: managing economic disruptions while capturing productivity gains; steering the attention economy toward human flourishing rather than engagement maximization; diffusing concentrated power while maintaining innovation incentives; addressing cultural and psychological effects on identity and meaning; and vigilantly managing catastrophic risks without succumbing to paralysis. The road ahead will be turbulent, marked by slow adoption in some domains and shockingly rapid transformation in others, by periods of stability punctuated by wrenching disruption. But I remain optimistic that with resilient institutions, ethical foresight, democratic deliberation, and a commitment to shared prosperity over winner-take-all dynamics, we can shape an AI-enhanced future that remains aligned with enduring human values.

## **Key Takeaways: Societal Impacts of AI**

### **What You Can Do:**

- **Prepare for uncertain economic transformation:** AI's labor market impact depends on speed and scope – from incremental automation to superintelligence in 5-20 years. Wages could collapse if AI outpaces traditional automation, with output surges but wage declines. Develop norms and institutions that redistribute AI gains to ensure shared prosperity rather than concentrated wealth.
- **Address the deskilling paradox:** Over-reliance on AI can undermine human capability. Experienced professionals may lose skills; early-career workers lose pathways from simple to complex tasks. Design hybrid human-AI teams where AI proposes options but humans make final judgments. Protect deliberate practice, redesign training pathways, and assess competency without AI assistance.
- **Redesign education for the AI era:** AI enables personalized learning at scale but undermines motivation when answers come too easily. Shift from transmitting knowledge to cultivating judgment: teach the art of questioning, assess thinking processes, validate judgment, not knowledge recall. Help learners develop human capacities: contextual judgment, ethical reasoning, creative synthesis, meaning-making.
- **Navigate the attention economy:** AI-driven systems optimize for engagement metrics, creating a mass distraction infrastructure that pulls attention toward sensational, addictive content. This leads to increased anxiety, reduced attention span, and depression especially among younger users. Counter with active, intentional tool use for creation and learning rather than passive consumption.
- **Confront power concentration:** AI capabilities concentrate in few firms and nations, creating “Matthew effects” where advantages compound. Who decides what AI optimizes for? These choices shape public discourse, economic opportunity, and state power. Develop mechanisms to challenge and contest AI's representation of the world.

## *Societal Impacts of AI*

- **Protect privacy and prevent surveillance:** AI's data hunger incentivizes ever-greater surveillance. Privacy erodes through meta-data and behavioral patterns. Implement privacy-by-design, data minimization, and anonymization. Recognize that informed consent is challenging when AI draws inferences users never explicitly provided.
- **Build resilience for diffusion barriers:** Safety concerns, regulatory hurdles, and organizational inertia slow AI deployment – providing a window to adapt, learn, and establish guardrails. Distinguish prudent barriers (safety, oversight) from excessive friction (red tape). Develop capacity to detect, respond to, and recover from failures at every scale. Perfect prevention is impossible; recovery capacity matters as much as prevention.

# **Conclusion: Leading Through Transformation in the AI Age**

A defining challenge of our era is captured in what has been called the exponential gap, i.e. the widening chasm between technologies that advance at accelerating rates and human institutions that adapt incrementally, if at all. AI exemplifies this dynamic most acutely: capabilities that seemed decades away arrive in months, while the social frameworks meant to govern them lag years behind. This challenge exceeds what conventional change management approaches can address. It represents a fundamental shift in how we understand change, organizational function, and societal adaptation. Traditional assumptions about gradual, predictable change collapse when technological capabilities evolve exponentially. Accelerating execution within existing frameworks will not suffice. Leaders must rethink their approaches: their purposes, methods, and success metrics. This book has examined this rethinking across three interconnected levels: individual development, organizational transformation, and broader societal evolution.

## **Individual Transformation: Developing Adaptive Capabilities**

Effective change leadership in an AI-driven environment requires leaders to develop enhanced cognitive and emotional capabilities. The primary constraint in navigating complex change isn't technical knowledge. It is the leader's capacity to process complexity, manage uncertainty, and adapt

## *Conclusion: Leading Through Transformation in the AI Age*

their mental models. Most adults operate with what Robert Kegan calls a “socialized mind” where consciousness is shaped primarily by external expectations, roles, and cultural norms. Some progress to a “self-authoring mind” that can step back from those expectations to construct and act from internally coherent principles. But the adaptive challenges of the AI age increasingly demand something even rarer: the “self-transforming mind” that can hold multiple perspectives simultaneously, embrace genuine paradox rather than resolving it prematurely, and continually reconstruct itself in response to new information. This isn’t about accumulating more knowledge or developing additional skills – competencies that can be trained in the conventional sense. It’s about expanding the very structure of consciousness through which we perceive reality. A self-transforming mind doesn’t just tolerate ambiguity; it recognizes that certainty itself is often the problem. It doesn’t just balance competing priorities; it sees that apparent contradictions often point to deeper syntheses. It doesn’t just adapt to change; it understands identity itself as fluid and contextual rather than fixed. Practically, this involves developing enhanced learning agility by approaching AI interactions, unexpected outcomes, and uncertainty as sources of information rather than threats to competence. It requires cultivating a “growth mindset,” extended to recognize that professional identity itself evolves continuously.

Such cognitive transformation doesn’t happen through insight alone. It requires what we might call “behavioral scaffolding” in the form of deliberate structures and practices that gradually shift how we think by first changing what we do. This includes: establishing routines for reflection that create space between stimulus and response; seeking out cognitive dissonance rather than avoiding it; practicing perspective-taking by genuinely inhabiting worldviews different from one’s own; and creating feedback mechanisms that surface blind spots before they become crises.

Emotional development is equally important. Change inevitably produces uncertainty, loss, and anxiety. Leaders who cannot process these emotions effectively may avoid necessary change or implement it in ways that damage organizational culture. This requires developing the capacity to remain functional while experiencing uncomfortable feelings. Research on “immunity to change” by Kegan and Lisa Lahey explains why emotional

work matters. Change efforts often fail not from lack of will or knowledge, but because unconscious commitments protect individuals from the very changes they consciously seek. For example, a leader may want to empower their team while unconsciously believing that being needed validates their worth, creating behaviors that maintain dependence. Such contradictions must be identified and addressed for meaningful change to occur.

Individual development also depends on clarity of purpose and ethical grounding. Research indicates that many senior executives struggle with purpose and satisfaction despite professional success, having optimized for metrics that prove ultimately hollow. Leaders navigating AI-driven change need internal stability that external validation cannot provide, ethical frameworks sophisticated enough for novel dilemmas, and meaning reserves that sustain them when conventional progress measures fail.

## **Organizational Transformation: Towards Hybrid Organizations**

Traditional organizational design assumed relatively stable environments where centralized control, specialized functions, and optimized processes created competitive advantage. AI disrupts these assumptions. Future organizations will feature hybrid teams where humans and AI agents work together, facing an accelerated pace of change that demands continuous adaptation. When environments change faster than planning cycles can adapt, centralized decision-making becomes a bottleneck. When problems increasingly span domains, functional silos prevent necessary integration. When efficiency optimization creates brittleness, disruptions prove catastrophic.

The response requires more than flattening hierarchies or forming cross-functional teams or other surface changes that often leave power structures intact. It demands redistributing decision rights to where knowledge and consequences converge: empowering those closest to problems to solve them while ensuring alignment through shared purpose and values

## *Conclusion: Leading Through Transformation in the AI Age*

rather than detailed control. This shifts leadership from making decisions to shaping contexts within which others decide. This involves replacing rigid hierarchies with fluid structures where authority flows toward competence rather than position, and different individuals lead different initiatives based on situational demands. It means moving from fixed roles to dynamic contribution patterns, where involvement in projects varies based on need and capability rather than job descriptions.

Implementing such changes requires cultural transformation alongside structural redesign. Organizations must develop psychological safety so that people can raise concerns, admit mistakes, and challenge assumptions without career penalty. Without this, distributed decision-making becomes risk-averse stasis as everyone waits for permission. Leaders create this safety through their responses to early adopters by rewarding candor even when inconvenient, treating failure as data rather than grounds for punishment.

Data and analytics serve not merely as optimization tools but as diagnostic instruments for change itself. Rather than implementing transformations uniformly, effective change leaders measure adoption, identify resistance patterns, and track where interventions succeed or struggle. This enables targeted responses rather than uniform approaches that waste resources and goodwill. However, data without empathy creates compliance without commitment. Change is fundamentally a social and emotional process. Strategies succeed or fail based on whether people embrace them, and people embrace change when it connects to identity, meaning, and belonging, not just logic. This is why narrative matters alongside analytics. Leaders must craft compelling stories that link change to organizational purpose and future vision, while listening to employee stories that reveal fears, aspirations, and informal sense-making processes.

AI integration amplifies both opportunities and risks. AI can augment human capabilities by automating routine work, providing decision insights, and enabling coordination in new forms and at unprecedented scales. However, AI can also deskill workforces, create dependencies that increase brittleness, embed biases that perpetuate injustice, and concentrate power in ways that undermine distributed decision-making. The challenge involves

designing human-AI collaboration that amplifies rather than replaces human capability. This requires intentional choices about automation scope, trust levels for AI recommendations, and skill maintenance strategies. It involves viewing AI as a tool that, properly deployed, allows humans to focus on creativity, judgment, and meaning-making that AI cannot or should not replicate.

## **Societal Transformation: Leadership as a Political Act**

Organizations operate within broader societal systems that enable or constrain change possibilities. Leadership in the AI age therefore has a public dimension not only in terms of visibility and reputation management, but in shaping the shared systems we inhabit together. The chapters on societal change have documented significant challenges: erosion of institutional trust, information ecosystems polluted by synthetic content and algorithmic amplification of extremism, democratic processes undermined by AI-enabled manipulation, and concentration of AI capabilities that may entrench rather than reduce inequality. These aren't problems any single organization can solve, yet every organization contributes to them, and every leader bears responsibility for the systems their decisions collectively shape.

This requires what we might call a more explicitly political form of leadership where leaders are building broad coalitions, negotiating between competing interests, creating shared meaning despite fragmentation, and forging collective action despite differences. Modern leaders must develop skills in reading complex stakeholder dynamics, sensing shifts in power and legitimacy, brokering compromises that preserve core values while enabling progress, and communicating vision across diverse worldviews. This political dimension becomes critical in AI governance, where decisions about development and deployment are too consequential for narrow technical or commercial considerations alone. Key questions include: Who decides what AI systems optimize for? Whose values get embedded in algorithms

## *Conclusion: Leading Through Transformation in the AI Age*

that shape what billions see, know, and decide? How do we ensure AI benefits are widely shared rather than concentrated? These are fundamentally political questions requiring democratic deliberation, not just expert technical judgment. More specifically, leaders can contribute to this deliberation through multiple approaches: advocating for sensible regulation rather than resisting all constraints, participating in multi-stakeholder initiatives that develop standards and best practices, being transparent about AI systems' capabilities and limitations, and ensuring affected communities have genuine voice in AI deployment decisions rather than consultation after choices are made.

Equally important is building a *trust infrastructure* in the form of networks, norms, and institutions that enable coordination despite complexity. Trust enables collective action: societies with high trust can move faster because people don't need to constantly guard against opportunism. Leaders build trust through consistency between words and actions, sharing information rather than hoarding it, admitting uncertainty and mistakes rather than projecting false confidence, and demonstrating consideration of others' interests alongside their own.

The book has also emphasized systemic resilience of society to withstand shocks and recover or improve afterward. Modern systems, optimized for efficiency, often prove brittle when stressed. Building societal resilience requires implementing at scale the same antifragile principles discussed for organizations: maintaining redundancy and slack rather than optimizing everything to the limit, cultivating diversity of approaches rather than betting everything on single solutions, decentralizing so local failures don't cascade system-wide, and ensuring decision-makers have "skin in the game" by sharing consequences of their choices rather than externalizing costs onto others.

For business leaders specifically, this means recognizing that long-term organizational success depends on the health of broader systems they participate in. A thriving company in a failing society represents a contradiction that cannot persist. This isn't altruism but enlightened self-interest. Companies need educated workforces (requiring functional educational systems), stable rule of law (requiring legitimate governance), physical in-

frastructure (requiring public investment), and social cohesion (requiring shared reality and trust). Therefore, contributing to societal resilience isn't a separate Corporate Social Responsibility initiative but core strategy. Above all, it requires taking responsibility for the world one's actions create. In the age of AI, this includes ensuring technologies are developed and deployed responsibly, being willing to forego profitable but harmful applications, protecting privacy and dignity even when data collection is legal, being transparent about risks and limitations, and accepting that some questions should not be outsourced to algorithms but require human judgment and accountability.

## **Reflexive Individual, Organizational, and Societal Levels**

The book's central argument is that the individual, organizational, societal levels aren't separate domains to be addressed sequentially, but interconnected aspects of a complex system. Changes at one level create pressures and possibilities at others. An individual leader's expanded consciousness enables organizational innovations that were previously unthinkable. Organizational practices that distribute agency and cultivate learning create environments where individuals can develop. Societal investments in education and social cohesion determine what human capital organizations can draw upon.

AI makes this integration both more urgent and more complex. Its impacts cascade across levels: changing skill requirements affects individuals, who require organizational support for reskilling, which depends on societal institutions providing educational infrastructure. AI-driven automation concentrates wealth unless organizations choose to share gains, which requires societal frameworks (taxation, regulation, social insurance) to function. AI governance cannot succeed as purely technical or business decisions; it requires societal deliberation that organizations must contribute to and individuals must participate in. The leaders who will navigate this successfully are those who can think and act across these levels simultaneously

## *Conclusion: Leading Through Transformation in the AI Age*

by understanding how their personal development enables organizational change, how organizational choices affect societal systems, how societal health constrains or enables what organizations can accomplish, and how these relationships form feedback loops that either reinforce progress or resistance.

This requires systemic wisdom, i.e. the capacity and willingness to see oneself as embedded in multiple overlapping systems, to understand that optimizing any single level often creates problems at others, and to seek interventions that create positive cycles across levels.

## **AI as Both Disruptor and Enabler**

Throughout this book, AI has appeared in dual roles: as the source of disruption requiring new forms of leadership, and as a tool that, properly deployed, enables that leadership to succeed. This duality is essential to understand. AI isn't simply a problem to be managed or a solution to be implemented – it's a fundamental restructuring of what's possible, for better and worse. On one hand, AI accelerates change beyond institutional capacity to respond, concentrates power in ways that threaten democratic values, creates dependencies that make systems brittle, enables surveillance and manipulation at unprecedented scale, and may eventually pose existential risks if misaligned superintelligence emerges. These aren't hypothetical concerns but emerging realities demanding urgent response. On the other hand, AI offers tools for understanding and coordinating complex systems that were previously opaque – making visible patterns that human cognition alone cannot detect, enabling simulation and foresight that improve decision quality, automating routine work to free human attention for higher-value contributions, and potentially solving problems (climate change, disease, material scarcity) that have resisted purely human efforts.

The task isn't choosing between these possibilities but holding them in productive tension: leveraging AI's capabilities while constraining its risks,

## *Moving Forward: Principles for Continued Learning*

using it to amplify human capability while ensuring humans remain genuinely in control, deploying it to solve problems while preventing it from creating worse ones. This tension cannot be resolved through purely technical means. It requires wise individuals making thoughtful choices, adaptive organizations implementing careful governance, and robust societies maintaining democratic control over technologies that might otherwise escape it.

## **Moving Forward: Principles for Continued Learning**

The conclusion of this book represents a beginning rather than an end. The frameworks, heuristics, tools, and reflections offered here aren't solutions to be simply copied but lenses through which to see and navigate challenges that will continue evolving. The specific techniques or tools mentioned will likely become obsolete in the face of exponential technological change. What endures are the underlying principles: the need for continuous individual development, the imperative of organizational adaptability, the requirement of societal resilience, and the integration across all three levels.

The age of AI will reward those who are prepared, principled, and proactive but not in the sense of having anticipated every specific development. Rather, it will reward those who've done the inner and outer work that enables adaptation itself: who've expanded their consciousness to handle complexity, who've built organizations capable of learning, who've contributed to societal systems robust enough to withstand shocks. It will reward those who recognize that leading change means first embodying it.

The next chapter is indeed yours to write. The question isn't whether you'll face disruption – you will. It isn't whether AI will reshape your domain – it already is. The question is whether you'll be an active author of that reshaping or a passive character in someone else's story. Whether you'll use these turbulent times as a catalyst for growth or an excuse for stasis. Whether you'll lead in ways that create shared flourishing or merely

*Conclusion: Leading Through Transformation in the AI Age*

personal success. The choice, as it always has been, is yours. But now, perhaps, you're better equipped to make it wisely.

# Bibliography

- Acemoglu, Daron. “The Simple Macroeconomics of AI.” *National Bureau of Economic Research*, 2024.
- Anderson, Chris. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *WIRED*, 2008. <https://www.wired.com/2008/06/the-end-of-theo/>.
- Anthropic. “Agentic Misalignment: How LLMs Could Be Insider Threats.” *Anthropic Research*, 2025. <https://www.anthropic.com/research/agentic-misalignment>.
- Anthropic. *Introducing the Model Context Protocol*. Anthropic Blog, 2024. <https://www.anthropic.com/news/model-context-protocol>.
- Argyris, Chris, and Donald A. Schön. *Organizational Learning II: Theory, Method, and Practice*. Addison-Wesley, 1996.
- Autor, David, and Alan Thomson. “The Expertise Model of Automation’s Impact.” *arXiv Preprint*, 2025.
- Azhar, Azeem. *Exponential: How Accelerating Technology Is Leaving Us Behind and What to Do about It*. Random House Business, 2021.
- Bain & Company. *88% of Business Transformations Fail to Achieve Their Original Ambitions; Those That Succeed Avoid Overloading Top Talent*. Bain & Company, 2024. <https://www.bain.com/about/media-center/press-releases/2024/88-of-business-transformations-fail-to-achieve->

## Bibliography

- their-original-ambitions-those-that-succeed-avoid-overloading-top-talent/.
- Beer, Michael, Russell A. Eisenstat, and Bert Spector. “Why Change Programs Don’t Produce Change.” *Harvard Business Review* 68, no. 6 (1990): 158–66.
- Bellefonds, Nicolas de, Tauseef Charanya, Michael R. Franke, et al. *Where’s the Value in AI?* Boston Consulting Group, 2024. <https://www.bcg.com/publications/2024/wheres-value-in-ai>.
- Bennis, Warren. *On Becoming a Leader*. Addison-Wesley, 1989.
- Bischoff, Paul. *Surveillance Camera Statistics: Which Are the Most Surveilled Cities?* 2025. <https://www.comparitech.com/vpn-privacy/the-worlds-most-surveilled-cities/>.
- Blühdorn, Ingolfuhr. *Unhaltbarkeit: Auf Dem Weg in Eine Andere Moderne*. Suhrkamp, 2024.
- Boston, Jonathan. *Governing the Future: Designing Democratic Institutions for a Better Tomorrow*. Emerald Group Publishing, 2016.
- Bostrom, Nick. *Deep Utopia: Life and Meaning in a Solved World*. Hachette Book Group, 2024.
- Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2016.
- Brady, William J., Killian McLoughlin, Tuan Nghi Doan, and Molly J. Crockett. “How Social Learning Amplifies Moral Outrage Expression in Online Social Networks.” *Proceedings of the National Academy of Sciences* 119, no. 33 (2022): e2213070119. <https://doi.org/10.1073/pnas.2213070119>.
- Bridle, James. *New Dark Age: Technology and the End of the Future*.

Verso, 2018.

Brynjolfsson, Erik et al. “AI and Productivity.” *arXiv Preprint*, 2023.

Brynjolfsson, Erik, Bharat Chandar, and Ruyu Chen. “Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence.” *Stanford Digital Economy Lab*, 2025. [https://digitaleconomy.stanford.edu/wp-content/uploads/2025/11/CanariesintheCoalMine\\_Nov25.pdf](https://digitaleconomy.stanford.edu/wp-content/uploads/2025/11/CanariesintheCoalMine_Nov25.pdf).

Budzyń, K. et al. “Endoscopist Deskilling Risk After Exposure to Artificial Intelligence.” *Journal of Gastroenterology*, 2025.

Cho, Aeree et al. “Transformer Explainer: Interactive Learning of Text-Generative Models.” *arXiv Preprint*, 2024. <https://arxiv.org/abs/2408.04619>.

Choi, Jonathan H., Amy B. Monahan, and Daniel Schwarcz. *Lawyering in the Age of Artificial Intelligence*. 4626276. Social Science Research Network, 2024. <https://ssrn.com/abstract=4626276>.

Chopra, Ayush, Santanu Bhattacharya, DeAndrea Salvador, et al. “The Iceberg Index: Measuring Workforce Exposure in the AI Economy.” *arXiv Preprint*, 2025. <https://arxiv.org/abs/2510.25137>.

Choudary, Sangeet Paul. *Reshuffle: Who Wins When AI Restacks the Knowledge Economy*. Amazon Digital Services LLC – KDP, 2025.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

Cummins, James. “The Threat of Analytic Flexibility in Using Large Language Models to Simulate Human Data: A Call to Attention.” *arXiv Preprint*, 2025.

DeepMind. “AlphaTensor.” *Nature* 610 (2022): 47–53.

## Bibliography

- DeepMind. “Mathematical Discoveries with LLMs.” *Nature* 623 (2023): 561–67.
- Dell’Acqua, Fabrizio, Charles Ayoubi, Hila Lifshitz, et al. *The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise*. Working Paper 33641. National Bureau of Economic Research, 2025.
- Dell’Acqua, Frabrizio et al. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.” *Harvard Business School Working Paper*, 2024.
- Diamond, Jared M. *Collapse: How Societies Choose to Fail or Succeed*. Penguin Books, 2011.
- Drucker, Peter F. *Management Challenges for the 21st Century*. Harper-Business, 1999.
- Dweck, Carol S. *Mindset: The New Psychology of Success*. Random House, 2006.
- Finkenstadt, Daniel J., Jake Sotiriadis, Peter Guinto, and Tojin Thomas Eapen. “Contingency Scenario Planning Using Generative AI.” *California Management Review*, 2024.
- Forster, E. M. *The Machine Stops*. Oxford University Press, 1909.
- Frankl, Viktor E. *Man’s Search for Meaning*. 1st ed. Translated by Ilse Lasch. Beacon Press, 1946.
- Georges, Gilles. *BMW’s iFACTORY: A Case Study in Digital Twins in Manufacturing at Enterprise Scale*. Clarity Points, 2025. <https://claritypoints.com/digital-twins-in-manufacturing-enterprise-scale/>.
- Ghaffarzadegan, Navid, Hazhir Rahmandad, and John Sterman. “Gen-

erative Agent-Based Modeling: Unveiling Social System Dynamics Through Hybrid Simulations.” *Nature Human Behaviour*, 2024.

Giddens, Anthony. *Modernity and Self-Identity: Self and Society in the Late Modern Age*. Polity Press, 1991.

Gottweiss, Juraj et al. “Towards an AI Co-Scientist.” *arXiv Preprint*, 2025.

Grace, Katja et al. “Thousands of AI Authors on the Future of AI.” *arXiv Preprint*, 2024.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. *2016 Expert Survey on Progress in AI*. AI Impacts, 2016. <https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/>.

Grace, Katja, Zach Stein-Perlman, Benjamin Weinstein-Rau, and John Salvatier. *2023 Expert Survey on Progress in AI*. AI Impacts, 2023. <https://aiimpacts.org/2023-expert-survey-on-progress-in-ai/>.

Graeber, David, and David Wengrow. *The Dawn of Everything: A New History of Humanity*. Farrar, Straus; Giroux, 2021.

Grimm, Maximilian, Óscar Jordà, Moritz Schularick, and Alan M. Taylor. *Loose Monetary Policy and Financial Instability*. Working Paper 30958. National Bureau of Economic Research, 2023.

Ha, Sungwha Hong] [Taehyun, Heyoung Yang. “Automated Weak Signal Detection and Prediction Using Keyword Network Clustering and Graph Convolutional Network.” *Futures*, ahead of print, Elsevier, 2023. <https://doi.org/10.1016/j.futures.2023.103202>.

Hackenburg, Kobi, Ben M. Tappin, Luke Hewitt, et al. “The Levers of Political Persuasion with Conversational Artificial Intelligence.” *Science* 390 (2025): eaea3884. <https://doi.org/10.1126/scienceaea3884>.

Haidt, Jonathan. *The Anxious Generation: How the Great Rewiring of*

## Bibliography

- Childhood Is Causing an Epidemic of Mental Illness.* Penguin Press, 2024.
- Haidt, Jonathan, and Greg Lukianoff. *The Coddling of the American Mind: How Good Intentions and Bad Ideas Are Setting up a Generation for Failure.* Penguin Press, 2018.
- Haken, Hermann. *Synergetics: An Introduction.* Springer, 1983.
- Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow.* Harvill Secker, 2016.
- Harter, Jim. *Employee Engagement: The Essential Guide for Business Leaders.* Gallup Press, 2017.
- He, Yang-Hui. “AI-Driven Research in Pure Mathematics and Theoretical Physics.” *Nature Reviews Physics*, ahead of print, 2024. <https://doi.org/10.1038/s42254-024-00740-1>.
- Heifetz, Ronald A. *Leadership Without Easy Answers.* Harvard University Press, 1994.
- Helbig, and Normann, Karlin. *The Psychological Safety Playbook - Leading More Powerfully by Being More Human.* Page Two Press, 2024.
- Hewitt, Luke et al. “Predicting Results of Social Science Experiments Using Large Language Models.” *Ethicalpsychology.com*, 2024.
- Huang, Saffron, Bryan Seethor, Esin Durmus, et al. “How AI Is Transforming Work at Anthropic.” Anthropic, 2025. <https://www.anthropic.com/research/how-ai-is-transforming-work-at-anthropic/>.
- Jackson, Tim. *Prosperity Without Growth: Foundations for the Economy of Tomorrow.* Second. Routledge, 2017.
- Jørgensen, Henrik H., Laurie Owen, and Alexander Neus. “Making Change

Work.” *IBM Global Business Services*, 2008.

Jumper, John, Richard Evans, Alexander Pritzel, et al. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596, no. 7873 (2021): 583–89. <https://doi.org/10.1038/s41586-021-03819-2>.

Jurenka, Irina et al. “Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach.” *Google DeepMind*, 2024.

Kalai, Adam Tauman et al. “Why Language Models Hallucinate.” *arXiv Preprint*, 2025.

Kasparov, Garry. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. PublicAffairs, 2017.

Kegan, Robert. *In over Our Heads: The Mental Demands of Modern Life*. Harvard University Press, 1994.

Kegan, Robert, and Lisa Lahey. *Immunity to Change: How to Overcome It and Unlock the Potential in Yourself and Your Organization*. Harvard Business Press, 2009.

Khan, Sal. *Khanmigo - on-Demand AI-Powered Support for Education*. 2024. <https://www.khanmingo.ai>.

Klarna. *Klarna AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month*. 2024. <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>.

Kokotajlo, Daniel et al. *AI 2027*. 2025. <https://ai-2027.com>.

Korbak, Tomek et al. “Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety.” *Preprint*, 2025.

## Bibliography

- Korinek, Anton. “Scenario Planning for an a(g)i Future.” *Finance & Development* 60, no. 4 (2023): 30–33. <https://www.imf.org/en/Publications/fandd>.
- Korinek, Anton, and Donghyun Suh. “Scenarios for the Transition to AGI.” *National Bureau of Economic Research*, 2024. <http://www.nber.org/papers/w32255>.
- Kotter, John P. *Leading Change*. Harvard Business School Press, 1996.
- Kreider, Mark R., Philip E. Higuera, Sean A. Parks, William L. Rice, Nadia White, and Andrew J. Larson. “Fire Suppression Makes Wildfires More Severe and Accentuates Impacts of Climate Change and Fuel Accumulation.” *Nature Communications* 15, no. 2412 (2024). <https://www.nature.com/articles/s41467-024-46702-0>.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems* 25 (2012): 1097–105.
- Kübler-Ross, Elisabeth. *On Death and Dying*. McMillan, 1969.
- Laloux, Frederic. *Reinventing Organizations: A Guide to Creating Organizations Inspired by the Next Stage of Human Consciousness*. Nelson Parker, 2014.
- Lewin, Kurt. “Group Decision and Social Change.” In *Readings in Social Psychology*, edited by T. M. Newcomb and E. L. Hartley. Holt, Rinehart; Winston, 1947.
- Lin, Hause, Gabriela Czarnek, Benjamin Lewis, et al. “Persuading Voters Using Human–Artificial Intelligence Dialogues.” *Nature*, ahead of print, 2025. <https://doi.org/10.1038/s41586-025-09771-9>.
- Lorenz, Edward N. “Predictability: Does the Flap of a Butterfly’s Wings in Brazil Set Off a Tornado in Texas?” *Meeting of the American Asso-*

*ciation for the Advancement of Science* (Washington, DC), 1972.

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.” *arXiv Preprint*, 2024. <https://arxiv.org/abs/2408.06292>.

Lu, Xinyi, Mark Riedl, et al. “LLMs and Generative Agent-Based Models for Complex Social Dynamics.” *arXiv Preprint*, 2024.

Mau, Steffen. *The Metric Society: On the Quantification of the Social*. Polity, 2019.

Maxwell, James Clerk. “On Governors.” *Proceedings of the Royal Society London* 16 (1867): 270–83.

McGilchrist, Iain. *The Matter with Things: Our Brains, Our Delusions and the Unmaking of the World*. Vols. 1–2. Perspectiva Press, 2021.

Mercator Institute for China Studies. *China’s Social Credit System in 2021: From Fragmentation Towards Integration*. 2021. <https://merics.org/en/report/chinas-social-credit-system-2021-fragmentation-towards-integration>.

METR. *Model Evaluation and Testing for Reliability*. 2025. <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>.

Midjourney Inc. “Midjourney AI.” 2023. <https://www.midjourney.com/>.

Mollick, Ethan. *Co-Intelligence: Living and Working with AI*. Random House, 2024.

Narayanan, Arvind, and Sayash Kapoor. *AI as Normal Technology*. 2025. <https://www.normaltech.ai/p/ai-as-normal-technology>.

Narayanan, Arvind, and Sayash Kapoor. *AI Snake Oil: What Artificial*

## Bibliography

- Intelligence Can Do, What It Can't, and How to Tell the Difference.* Princeton University Press, 2024.
- Nordhaus, William D. *The Climate Casino: Risk, Uncertainty, and Economics for a Warming World.* Yale University Press, 2013.
- Novikov, Alexander et al. “AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery.” *Google DeepMind*, 2025.
- Office, U. S. Government Accountability. *Science & Tech Spotlight: Digital Twins—Virtual Models of People and Objects.* GAO-23-106453. U.S. Government Accountability Office, 2023. <https://www.gao.gov/products/gao-23-106453>.
- OpenAI. “GPT-4 Technical Report.” *OpenAI Research*, 2023.
- OpenAI. *Introducing GPT-5.2.* OpenAI, 2025. <https://openai.com/index/introducing-gpt-5-2/>.
- OpenAI. *The State of Enterprise AI 2025.* OpenAI, 2025. <https://openai.com/index/the-state-of-enterprise-ai-2025-report/>.
- Ord, Toby. *The Precipice: Existential Risk and the Future of Humanity.* Grand Central Publishing, 2020.
- Orlando, Barbara, and Tomaso Eridani. *In the Network of the Conclave.* Bocconi University, 2025. <https://www.unibocconi.it/en/news/network-conclave>.
- Park, Michael, Erin Leahey, and Russell J. Funk. “Papers and Patents Are Becoming Less Disruptive over Time.” *Nature* 613 (2023): 138–44. <https://doi.org/10.1038/s41586-022-05543-x>.
- Parra-Moyano, José et al. “Executives Who Used Gen AI Made Worse Predictions.” *Harvard Business Review*, 2025.

- Patel, Jaisal, Yunzhe Chen, Kaiwen He, et al. “Reasoning Models Ace the CFA Exams.” *arXiv Preprint*, 2025. <https://arxiv.org/abs/2512.08270> v1.
- Patwardhan, Tejal et al. “GDPVAL: Evaluating AI Model Performance on Real-World Economically Valuable Tasks.” *OpenAI Research*, 2025.
- Pentland, Alex. *Social Physics: How Good Ideas Spread-the Lessons from a New Science*. Penguin Press, 2014.
- Piao, Yuchen et al. *AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviours and Society*. 2025.
- Piketty, Thomas. *Capital in the Twenty-First Century*. Translated by Arthur Goldhammer. Harvard University Press, 2014.
- Pinker, Steven. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Penguin Books, 2018.
- Prize, ARC. *ARC Prize Twitter Post*. 2025. <https://x.com/arcprize/status/1999182732845547795>.
- Prize, ARC. *ARC-AGI Leaderboard*. 2025. <https://arcprize.org/leaderboard>.
- Public Sector Innovation, OECD Observatory of. *Virtual Singapore – Singapore’s Virtual Twin*. Singapore Land Authority, 2015. <https://oecd-opsi.org/innovations/virtual-twin-singapore/>.
- Puranam, Phanish. *Re-Humanize: How to Build Human-Centric Organizations in the Age of AI*. Penguin Random House, 2024.
- Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical Text-Conditional Image Generation with CLIP Latents.” *arXiv Preprint*, 2022. <https://arxiv.org/abs/2204.06125>.

## Bibliography

- Rathje, Steve, Jay J. van Bavel, and Sander van der Linden. “Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media.” *Oxford University Press Academic*, 2023.
- Raworth, Kate. *Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist*. Chelsea Green Publishing, 2017.
- Reckwitz, Andreas. *The Society of Singularities*. Translated by Valentin Schweitzer. Polity Press, 2020.
- Reuters. “PwC’s 4,000 Legal Staffers Get AI Assistant as Law Chatbots Gain Steam.” *Reuters*, 2023. <https://www.reuters.com/world/uk/pwc-s-4000-legal-staffers-get-ai-assistant-law-chatbots-gain-steam-2023-03-15/>.
- Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Omran. “High-Resolution Image Synthesis with Latent Diffusion Models.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 10684–95.
- Rosa, Hartmut. *Social Acceleration: A New Theory of Modernity*. Translated by Jonathan Trejo-Mathys. Columbia University Press, 2013.
- Rousseau, Denise M. *Psychological Contracts in Organizations: Understanding Written and Unwritten Agreements*. SAGE Publications, 1995.
- Russell, Stuart. *Human Compatible - AI and the Problem of Control*. Viking, 2019.
- Salvi, Francesco et al. “On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial.” *arXiv Preprint*, 2024.
- Sandel, Michael J. *What Money Can’t Buy: The Moral Limits of Markets*. Farrar, Straus; Giroux, 2012.

- Scheffer, Marten. *Critical Transitions in Nature and Society*. Princeton University Press, 2009.
- Scheffer, Marten et al. “Catastrophic Shifts in Ecosystems.” *Nature* 413 (2001): 591–96.
- Schein, Edgar H. *Organizational Culture and Leadership*. Jossey-Bass, 1985.
- Schelling, Thomas C. “Dynamic Models of Segregation.” *Journal of Mathematical Sociology* 1, no. 2 (1971): 143–86.
- Schiffer, Zoë. *Here’s What Mark Zuckerberg Is Offering Top AI Talent*. 2025. <https://www.wired.com/story/mark-zuckerberg-meta-offer-top-ai-talent-300-million/>.
- Schlegel, Katja et al. “Large Language Models Are Proficient in Solving and Creating Emotional Intelligence Tests.” *Communications Psychology*, 2025.
- Schmidt, Lena, Oshin Sharma, Chris Marshall, and Sonia Garcia Gonzalez Moral. “Horizon Scans Can Be Accelerated Using Novel Information Retrieval and Artificial Intelligence Tools.” *arXiv Preprint arXiv:2504.01627*, ahead of print, 2025. <https://doi.org/10.48550/arXiv.2504.01627>.
- Schoenegger, Philipp et al. “Large Language Models Are More Persuasive Than Incentivized Human Persuaders.” *arXiv Preprint*, 2025.
- Schularick, Moritz, and Alan M. Taylor. “Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008.” *American Economic Review* 102, no. 2 (2012): 1029–61.
- Schumpeter, Joseph A. *Capitalism, Socialism and Democracy*. Harper & Brothers, 1942.

## Bibliography

- Scott, James C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, 1998.
- Senge, Peter M. *The Fifth Discipline: The Art and Practice of the Learning Organization.* Doubleday/Currency, 1990.
- Shoham, Shlomo. *Future Intelligence.* Bertelsmann Stiftung, 2011.
- Silver, David, Aja Huang, Chris J. Maddison, et al. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529, no. 7587 (2016): 484–89. <https://doi.org/10.1038/nature16961>.
- Soda, Giuseppe, Alessandro Iorio, and Leonardo Rizzo. “In the Network of the Conclave: Social Connections and the Making of a Pope.” *Social Networks* 83 (2025): 215–32. <https://doi.org/10.1016/j.socnet.2025.07.003>.
- Suleyman, Mustafa. *The Coming Wave: Technology, Power, and the Twenty-First Century’s Greatest Dilemma.* Crown, 2023.
- Taleb, Nassim Nicholas. *Antifragile: Things That Gain from Disorder.* Random House, 2012.
- Taleb, Nassim Nicholas. *Skin in the Game: Hidden Asymmetries in Daily Life.* Random House, 2018.
- Tamkin, Alex, and Peter McCrory. “Estimating AI Productivity Gains from Claude Conversations.” Anthropic, 2025. <https://www.anthropic.com/research/estimating-productivity-gains>.
- The Economist. “What If the \$3trn AI Investment Boom Goes Wrong?” *The Economist*, 2025. <https://www.economist.com/leaders/2025/09/11/what-if-the-3trn-ai-investment-boom-goes-wrong>.
- Vaswani, Ashish et al. “Attention Is All You Need.” *31st Conference on Neural Information Processing Systems*, 2017.

- Wilber, Ken. *A Theory of Everything: An Integral Vision for Business, Politics, Science and Spirituality*. Shambhala Publications, 2000.
- Wolfram, Stephen. *What Is ChatGPT Doing ... And Why Does It Work?* 2023. <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>.
- Yang, K. et al. “LeanDojo: Theorem Proving with Retrieval-Augmented Language Models.” *arXiv Preprint*, 2023.
- Youyou, Wu, Michal Kosinski, and David Stillwell. “Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans.” *Proceedings of the National Academy of Sciences* 112, no. 4 (2015): 1036–40. <https://doi.org/10.1073/pnas.1418680112>.
- Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019.



# About the Author

Johannes Meier is an Honorary Professor at HHL Leipzig Graduate School of Management, where he has taught change management since 2003. His professional career includes senior leadership roles in the technology, consulting, and non-profit sectors.

He currently serves as Chairman of Stiftung Mercator and holds supervisory board positions at Nederlandse Gasunie N.V., Meridian Stiftung, and Deutsches Komitee für UNICEF e.V. Previously, he served as CEO of the European Climate Foundation (2011–2017), Managing Board Member of the Bertelsmann Foundation (2003–2009), CEO of GE CompuNet Computer AG (1998–2003), and Partner at McKinsey & Company (1990–1997). He has also held supervisory board positions at Onvista AG, HHL gGmbH, XING AG, Cliqz GmbH, and New Work SE.

In 2009, he developed the Labor Market Monitor, a collaboration platform for the German Federal Employment Agency, which received Germany's 2011 award for the most innovative e-government solution.

He holds an M.S. in Computer Science from RWTH Aachen University and a Ph.D. in Communication and Information Sciences from the University of Hawaii at Mānoa.

