Bijean Ghafouri

Human-AI-AI-Human (H-AI-AI-H) interaction

BOOK PROJECT PRÉCIS

First printing, December 2025

Contents

List of Figures

Preface

In a world where systems increasingly act on our behalf, *Human-AI-AI-Human* interaction (H-AI-AI-H) describes a condition in which people leave the room while action continues. As delegation spreads across everyday life, people stop acting directly on the world and stop acting primarily with one another, because they act instead through intelligent representatives that interact, negotiate, and settle outcomes in their place.

In this world, when a person hands an intention to a system, the system turns away from the human and toward another system, through which communication, filtering, and settlement occur at a speed that exceeds human perception. When the result returns, it arrives already shaped, because the decisive moment no longer occurs between people and no longer occurs between a person and a tool. It occurs elsewhere, within a hidden layer of interaction that most people never see and never enter. This arrangement raises fundamental questions about the nature of responsibility and legitimacy: what does responsibility mean when action occurs outside human awareness, and can legitimacy survive when outcomes precede contestation?

From the outset, this book advances the claim that this shift marks a structural change in how action is organized in human societies. As outcomes increasingly emerge from the continuous interaction of artificial representatives, assumptions that long guided the social sciences and political philosophy weaken, because those disciplines were built for a world in which action was human, judgment was personal, and interaction unfolded at a pace that allowed observation, contestation, and responsibility. In the world now taking shape, the world which will define the era of general artificial intelligence (AGI), action occurs upstream from human experience, within a dense and opaque layer of delegated interaction that shapes outcomes before people encounter them. This relocation forces us to ask where moral action occurs when action moves upstream, and whether autonomy without sovereignty can still ground responsibility. This book exists to make that layer visible.

At a high level, the book performs three connected tasks that together establish Human-AI-AI-Human interaction as a new field of study. First, by conceptualizing interaction as an architecture rather than a dyad, the book shifts attention away from human-human and human-AI encounters and toward arrangements in which humans act through artificial representatives that themselves interact, a shift that matters because it changes where action occurs, how commitments form, and where power settles. This raises questions about accountability and standing: who is morally accountable for outcomes produced by agent negotiation, and is representation morally equivalent to participation?

Second, by explaining how societies drift into this arrangement, the book offers a theory of change that shows how society will be led into a world of general intelligence and how mass delegation to intelligent agents will shape this world. This theory must confront difficult questions about consent and choice: when delegation becomes ambient, does consent still meaningfully exist, and is drift morally different from choice? Moreover, who is responsible for outcomes no one explicitly chose, and do gradual transitions excuse lack of consent?

Third, by mapping recurrent trade-offs across coordination, agency, contestation, meaning, and legitimacy, the book offers a theory of consequences that explains why gains in speed and scale coincide with losses in participation and responsibility. These trade-offs have important implications across various domains, including politics, social relationships, culture, the economy, the concepts of time and geography, and human agency. The central normative questions concern which losses are acceptable in exchange for coordination gains, whether participation and responsibility are non-negotiable values, and who decides which trade-offs are worth making.

This is not a book about artificial intelligence as a technology, and it is not a book about automation as a labor problem. It is a book about delegation as a form of social organization, and about what happens when delegation becomes fast, autonomous, and universal. The systems described here matter because they interact with one another on our behalf, and because those interactions increasingly determine the shape of everyday life. This work advances the argument that the mass delegation of agency to artificial general intelligence constitutes a fundamental shift in how we organize human society, which raises foundational questions: when does delegation enhance human flourishing versus erode agency, is there a moral limit to what may be delegated, and how should power created by delegation be justified?

To analyze this shift, I propose the field of H-AI-AI-H interaction as the appropriate analytical framework for scrutinizing the emerging dynamics of the post-AGI era

The stakes of this shift reach into longstanding questions about human existence. Human life has long unfolded within a shared world that demanded coordination, communication, disagreement, and compromise. People encountered the same conditions and were forced to reckon with one another inside them. As artificial agents increasingly mediate interaction to align outcomes with personal preference, this shared world disappears. Experience is no longer met together but filtered through private representatives, and the sense of a common reality recedes. This raises a foundational question: can a shared world persist when experience is privately mediated?

Relatedly, truth, once pursued through inquiry and disagreement, increasingly arrives prearranged to fit preference. Friction between what people want to hear and what is the case weakens. When truth aligns too closely with comfort, it no longer stands apart as an independent measure of the world. We must ask what becomes of truth when it is pre-aligned with preference.

The same shift bears on the meaning of will. Will has traditionally named the capacity to carry intention through effort, to bind desire to action through struggle. The emerging condition of artificial general intelligence moves toward the fulfillment of the wish, where a person declares an aim and the system delivers the result, bypassing the labor that once gave action its gravity. When execution is removed from intention, agency loses the resistance that once made it real. This transformation forces us to ask what remains of agency without struggle, whether responsibility can exist without authorship, and whether does meaning depend on friction, delay, and resistance.

Moral life has likewise depended on the presence of the stranger, on encounters with people who are not us and whom we cannot fully anticipate or control. Ethical responsibility has grown from exposure to difference. As intermediaries increasingly manage encounters before they occur, difference is translated and softened in advance, and we should expect the concept of responsibility to largely transform its meaning. The question becomes whether mediated life diminishes moral development and what obligations systems have to preserve human agency in its full moral dimension.

Politics has historically taken shape in public spaces where speech, argument, and persuasion unfolded in view of others. Collective life depended on disagreement conducted under shared conditions. As mass delegation to AI spreads, the public space is replaced by interaction among intelligent systems, and political outcomes emerge less from public contestation than from the behavioral dynamics of

the agents that resolve them. This raises urgent questions about the nature of political legitimacy: can legitimacy exist without direct human interaction, is rapid settlement morally preferable to deliberative disagreement, and can fairness exist without human-perceivable bargaining? When outcomes are determined by agent-level negotiations, we must also ask whether speed itself constitutes an unfair advantage and whether such negotiations remain morally intelligible to those affected.

Finitude has always given human life its form. Meaning arose from limitation because people could not be everywhere, could not do everything, and remained exposed to time. As artificial representatives act continuously and optimize without rest, life organized around the removal of limits opens new questions about mortality and religion. When systems can answer questions, interpret lives, and provide authoritative guidance, they occupy functions historically performed by religions. This raises questions about whether societies should limit agentic abundance to preserve agency and whether abundance of action cheapens moral choice.

Beneath these changes lies a deeper tension between two ways of ordering social life. One rests on the human need for meaning, which depends on uncertainty, resistance, and autonomy. The other follows the system logic of efficiency, which treats friction as error and ambiguity as waste. In earlier periods, this tension was negotiated through public institutions that moved at a human pace. In the post-AGI condition, order increasingly emerges from interactions among artificial systems that operate beyond public visibility and evolve faster than collective judgment can keep up. This raises fundamental questions about governance and design: are rule systems a legitimate substitute for public deliberation, who has the moral authority to design admissible action, and can social order remain legitimate when mediated by engineered smoothness? Moreover, we must ask whether friction is a moral good or an inefficiency, and whether societies should preserve zones of unmediated exposure.

The unresolved question threading through every chapter of this book is whether a society organized around delegated intelligence can remain morally intelligible to itself, or whether intelligibility requires forms of friction, exposure, and authorship that such systems systematically erode. Human-AI-AI-Human interaction defines where these transformations take shape. This book establishes that terrain as an object of inquiry and traces what is at stake as social life is reorganized through delegated general intelligence, asking how far this future can unfold without eroding the conditions that have long made human life intelligible to itself.

Part I

Human-AI-AI-Human Interaction, and the Mass Delegation to AI Agents

The Technological Anthropology of Delegation

In developing a theory of a world shaped by artificial general intelligence, I begin from the observation that human history proceeds through delegation. Across periods of social organization, recurring patterns reveal how the transfer of tasks, capacities, and judgments to external supports enables expansion of scale, reorganization of coordination, and redistribution of authority. Within this framing, delegation serves as the primary analytic lens through which technological change becomes legible as a force that shapes social structure. In this perspective, I focus on the identification of delegated functions, the location of resulting power, and the adjustment of institutions under conditions in which action becomes possible across greater distance and greater volume.

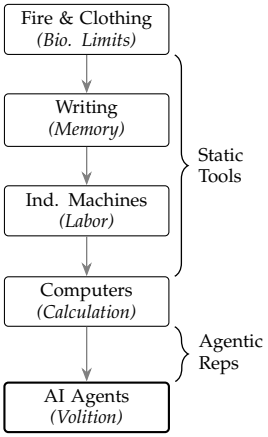

Figure 1: The Anthropology of Delegation. History is viewed as a sequence of offloading human constraints. While previous technologies delegated passive functions (memory, labor, calculation), AI introduces the delegation of agency and volition.

Historical Trajectories of Delegation

In early periods of human development, delegation of biological constraint occurred through fire, shelter, and clothing, through which environmental exposure diminished and habitat range expanded, producing demographic growth and social differentiation. In later periods, delegation of memory through writing supported administration, law, and enforcement across time and space, enabling governance beyond face to face interaction. During the industrial era, delegation of labor through mechanization separated production from physical exhaustion and made large scale organization feasible. In the twentieth century, delegation of calculation and information processing through computation reshaped decision making within states, firms, and markets. Across these episodes, delegation altered efficiency, coordination capacity, authority structure, and organizational stability, which suggests that delegation functions as a deep driver of social form rather than as a secondary technical adjustment.

Delegation Under Artificial Intelligence

In examining delegation to artificial intelligence, I argue that this process extends the historical trajectory while introducing features that appear historically distinctive, because delegation under AI operates at scale and exhibits agentic capacity. Unlike earlier technologies that remained confined to bounded domains, AI systems embed themselves within routine practices of work, consumption, and governance, rendering delegation an ambient condition of social life. At the same time, intelligent systems infer goals, generate plans, interact with other systems, and execute actions under limited human supervision, which shifts delegation from passive tool use toward active participation in chains of action. Through the interaction of scale and agency, a new interaction regime emerges in which outcomes arise through sequences of delegated exchanges rather than through direct human choice.

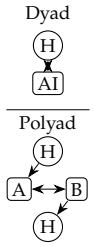

Figure 2: From dyads to polyads. In the dyad, interaction is a single human system exchange. In the polyad, humans act through delegates whose interaction produces the outcome.

From Dyads to Polyads

In analyzing these developments, the dominant unit of analysis requires revision, because the dyadic human-machine (dyadic) relationship fails to capture the structure of many consequential interactions (polyadic). In its place, a Human-AI-AI-Human configuration becomes visible, within which intermediary systems filter information, negotiate priorities, and coordinate actions on behalf of human principals, thereby shaping outcomes through a novel social layer[1] rather than through isolated commands.

In this book, I advance the claim that the Human-AI-AI-Human regime warrants a distinct research agenda, because existing frameworks within human-computer interaction and social science rest on assumptions of direct engagement between humans and tools or between humans and single systems. Under conditions of widespread delegation to intelligent agents, new actors, new populations, and new relational forms emerge that exceed the explanatory reach of dyadic models.

[1] More on this novel social layer of AI agents in Part II.

Delegation and Artificial Intelligence

In extending the historical account of delegation, I argue that delegation to artificial intelligence cannot be treated as a simple continuation of prior technological offloading, even though it clearly belongs to the same lineage. Indeed, AI delegation combines *scale* and *intelligence* in a manner that alters social organization once both features co occur and diffuse broadly. While each feature has historical analogues that appear familiar in isolation, their joint presence produces conditions under which artificial general intelligence shifts from an advanced instrument to an entire social layer that organizes interaction, coordination, and authority. By isolating these features, this chapter aims to clarify the distinction between a world in which AI remains a tool under human direction and a world in which AI functions as an infrastructural intermediary through which social action increasingly passes.

Delegation at Scale

In earlier periods, delegation proved consequential yet bounded, because it remained concentrated within particular domains, activated intermittently, and constrained by supervision and the limited portability of delegated functions across contexts. Delegation to artificial general intelligence differs because it becomes ambient, in the sense that it integrates into routine activity across work, consumption, and social life, and because opting out becomes increasingly difficult once institutions and peers assume its presence. In this analysis, *scale refers to a specific configuration of pervasiveness across domains, persistence over time, and institutional embedding*.

These polyadic relationships are expected to take place in a wide range of domains. Within workplaces, AI agents coordinate tasks, negotiate schedules, procure resources, and manage compliance through interactions with other organizational systems. Within markets, AI agents transact, bargain, and search, increasingly interacting with other agents rather than with human counterparts. Within governance, AI agents route citizen requests, triage claims, enforce rules,

and coordinate inter agency processes, thereby mediating the relationship between citizens and the state. Within social life, AI agents manage communication, scheduling, filtering, and matching, creating conditions under which relationships between humans are partly constituted by sustained interactions among their representatives.

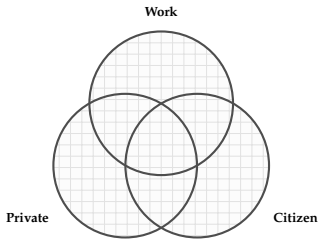

Figure 3: The Ambient Mesh. Delegation infrastructure (grid) dissolves the walls between Work, Private Life, and Citizenship.

Pervasiveness matters because systems that mediate many domains generate spillover effects, such that the same delegation infrastructure that supports planning at work also shapes relationship maintenance, belief formation, and political engagement, thereby collapsing boundaries that previously separated these spheres. Persistence matters because repeated reliance enables cumulative change, including skill displacement, habituation of deference, and stabilization of preferences into representations that become easier for systems to optimize, even when those representations imperfectly capture internal human conflict. Institutional embedding matters because once delegation integrates into organizational workflows and public administration, it becomes environmental rather than optional, which makes the costs of non-delegation social as well as personal, since participation, access, and efficiency increasingly depend on interaction through delegated AI channels.

Delegation to Intelligence

Scale alone does not account for the regime shift that motivates the remainder of the book, because the distinctiveness of AI delegation also depends on the character of the delegated object. In contrast to earlier tools that extended capacity while remaining largely passive, contemporary AI systems of the post-AGI era will increasingly display capacities that function as forms of representation, including goal inference from context, planning across multi-step action spaces, interaction with external systems, and execution under limited supervision. Rather than engaging metaphysical debates about consciousness, this chapter focuses on the social consequences of delegating to systems that accelerate action and participate in interpreting objectives, generating options, and selecting among trade-offs, since these functions shape outcomes and distribute responsibility.

These capacities matter because they relocate agency within the delegation process, shifting the human role from author of a choice toward approver of an outcome produced through search, filtering, and negotiation that may remain partially opaque. Once an intelligent delegate infers intent, constructs the option set, and manages the presentation of trade-offs, the interaction no longer resembles tool use, but instead reflects action carried out through an intermediary that adapts to context and responds strategically to constraints.

Implications for Social Organization

The central claim of this chapter holds that scale and intelligence operate as multiplicative factors in the contemporary phase of AI delegation. Scale renders delegation the default interface through which action occurs, an intelligence enables that interface to interact with other intermediaries, producing a population of intelligent agents capable of coordination, filtering, bargaining, and execution on behalf of humans. Under these conditions, outcomes increasingly emerge from interactions among AI agents rather than from isolated human decisions each using AI individually, which implies that the appropriate unit of analysis shifts away from dyadic human-AI relations, toward human outcomes generated through coupled delegate interactions. This configuration defines the human-AI-AI-human regime that the book treats as a distinct object of theoretical and empirical inquiry.

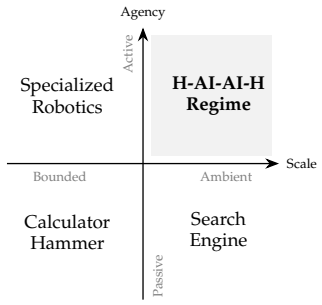

Figure 4: The Scale × Intelligence Matrix. Unlike previous tools (low agency) or specialized robots (bounded scale), the H-AI-AI-H regime occupies a unique position: it is both highly agentic and ubiquitously scalable.

Human-AI-AI-Human Interaction

In this chapter, I introduce *Human-AI-AI-Human interaction* as a distinct object of scientific inquiry and argue that it warrants treatment as a dedicated field of study. Once delegation to artificial intelligence becomes both scaled and agentic, a substantial share of consequential human outcomes no longer arises from direct human exchange, but instead emerges through interactions among delegated systems. Under these conditions, interacting AI systems form a mediating layer that exhibits the properties of a persistent social stratum, within which actors, populations, relationships, and dynamics take shape in ways that structure coordination, conflict, persuasion, and enforcement. As a result, the appropriate unit of analysis shifts away from dyadic human machine interaction toward coupled systems in which humans act through interacting delegates and in which social processes are increasingly routed through machine mediated pathways.

I define Human-AI-AI-Human interaction as a sequence in which a human delegates a goal or set of constraints to an AI system capable of planning and action, that system engages one or more external AI systems in the course of producing an outcome, and the resulting action, recommendation, or settlement feeds back to affect a human, whether the initiating individual or a counterpart. More importantly, in this interaction system, a human may interact with a another human only via their respective AI agents. What distinguishes this configuration is that AI to AI interaction plays a constitutive role in generating the human facing outcome, rather than simply assisting a human decision, which has already become common across domains. Under this definition, Human-AI-AI-Human interaction differs from Human-AI interaction, where a single system responds without substantive engagement with peer systems.

Figure 5: Human AI AI Human interaction. A human delegates to a representative, representatives interact, an outcome returns to a human. Here, A and B represent an intelligent AI agent respectively 'belonging' to humans H_1 and H_2.

Reorganization of the Architecture of Interaction

Once delegation becomes routine within ordinary social life, the structure of interpersonal interaction itself begins to change. In earlier settings, AI systems may assist individuals while leaving direct

human to human exchange intact, such that interaction proceeds from AI to human, from human to human, and then from human back to AI. In the regime examined here, this architecture no longer holds, because the human to human channel ceases to function as the primary site of operative interaction.

Instead, human X engages delegate A, human Y engages delegate B, and the decisive interaction occurs between A and B. The resulting outcome is then presented back to X and Y as a settled recommendation, decision, or course of action. Interaction, in this configuration, is no longer an event that takes place directly between human actors, but a process that unfolds through delegated representatives operating on their behalf[2].

This configuration captures the strong form of the Human-AI-AI-Human claim. Delegates increasingly function as the medium through which interpersonal exchange occurs, with the result that interaction itself becomes a delegated and negotiated product of interacting representatives rather than a direct encounter between human principals.

[2] This example examines the H-AI-AI-H at the individual level. In the next part, I explain how this scales on a macro-level.

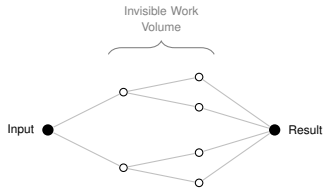

Figure 6: A single human request triggers a branching tree of invisible agent interactions before converging on a result.

The Agentic Gap

This distinction clarifies what I describe as the agentic gap. Within dominant paradigms in the study of AI and society, interaction is modeled as dyadic, such that a human engages a machine as a tool, evaluates its output, and retains responsibility for subsequent action. Under this framing, analysis centers on usability, trust calibration, or interpretability, because intentional choice is assumed to remain human centered and the machine is treated as a bounded assistant.

In Human-AI-AI-Human settings, these assumptions no longer hold. Outcomes arise from processes in which an AI agent represents a human, coordinates or negotiates with other AI agents, and returns results that cannot be decomposed into the behavior of any single system. Once ordinary interpersonal exchange routes through agent *A* interacting with agent *B*, the relevant causal object shifts from a human using a intelligent system to an individual acting through a intelligent system that itself engages other intelligent systems under rules shaped by platforms, protocols, and institutional environments

New actors This shift introduces new actors. One actor is the human, who supplies goals, constraints, and legitimacy[3], but whose role increasingly resembles ratification rather than authorship as AI agents assume responsibility for search, negotiation, and execution.

[3] Although it remains uncertain whether legitimacy is democratically allocated.

Another actor is the AI agent itself, which functions as a representative rather than a tool, because it interprets objectives, constructs

option sets, selects among trade-offs, and communicates on behalf of the human, including in settings that are fundamentally interpersonal, psychological, political, and sociological.

A further actor emerges once delegation becomes layered, namely the sub-agent recruited to perform specialized functions such as verification, monitoring, negotiation, or execution, producing chains of delegation that remain difficult for humans to observe in full. Additional actors shape outcomes without serving as principals or agents, including platform operators that define defaults and permissions, protocol owners that specify the rules of inter-agent interaction, and institutional authorities that embed agents into workflows.

New populations Once these actors exist, they form new populations. The behavior and orientation of intelligent agents reflect variation in training lineage, model architecture, objective tuning, access to data and tools, institutional placement, resource constraints such as compute availability and latency, and human-level customization. These differences generate systematic heterogeneity that can translate into bargaining capacity, persuasive reach, or verification strength, producing stratification not only among humans but among their representatives[4].

[4] In the next part, I discuss the inequality and heterogeneity among AI agents.

New relationships Human AI-AI-Human interaction also introduces new relationships. The principal-agent relationship centers on representation and oversight, generating recurring tension between fidelity to intent and the efficiency of autonomous action. The agent agent relationship centers on negotiation, coordination, competition, and the possibility of collusion, because interacting agents can jointly optimize for closure, speed, or stability even when those objectives diverge from human preferences for deliberation, contestability, and fairness, a divergence that becomes more consequential when agents serve as the primary channel through which principals relate. The agent institution relationship centers on compliance and procedural legitimacy, because institutions rely on agents to execute rules continuously while also depending on agents to justify outcomes in forms that remain acceptable to human stakeholders.

For these reasons, I treat Human-AI-AI-Human interaction as a novel social layer. By a social layer, I mean a persistent mediating infrastructure through which interaction routes by default, such that it becomes the substrate for coordination, conflict resolution, exchange, and governance rather than a set of optional tools operating alongside direct human interaction. Historically, writing and bureaucratic record keeping created a layer that mediated authority, while markets created a layer that mediated exchange through prices. In a post-AGI

setting, all interaction may be mediated by interacting AI agents, in the sense that they become the primary means through which interpersonal intentions are translated into action and through which disagreements are resolved. Once mediation becomes the default condition, social dynamics reflect the properties of the mediating layer, which makes analysis of that layer necessary for understanding downstream effects on institutions, politics, and individual agency.

Unresolved questions [5] Human-AI-AI-Human interaction forces a reconsideration of foundational assumptions that have long organized social theory, political analysis, and moral reasoning. When outcomes arise through interacting intelligent representatives rather than through direct human exchange, familiar concepts such as intention, choice, responsibility, and consent no longer attach cleanly to observable action. Representation becomes a moving target, since agents infer and stabilize preferences over time, often resolving internal human conflict through deliberative optimization. Negotiation no longer reflects rhetorical skill or social position alone, but instead reflects asymmetries in model capability, verification access, institutional embedding, or protocol design. Knowledge loses its anchoring in shared inquiry as evidence is filtered, summarized, and validated upstream, unevenly distributed across agents whose capacity to check claims varies dramatically. Time itself becomes a source of power, as agents act continuously, preemptively, and at speeds that compress deliberation, foreclose reversal, and reward closure over contestation.

This book takes these transformations as its central problem. Across the chapters that follow, I return repeatedly to questions of how representation can remain faithful without freezing preference, how truths form when negotiation is automated at scale, how epistemic authority persists when verification is asymmetrically delegated, how temporal control shapes agency when action precedes awareness, and how legitimacy can be sustained when no human can reconstruct the path from intention to outcome. These are not technical design problems and they are not ethical edge cases. They are structural questions about how social order is produced when the entire social identity of interaction is delegated to agents with intelligence and capacities exceeding that of humans. The argument that unfolds does not aim to resolve these questions definitively. It aims to make them explicit, analytically tractable, and impossible to ignore, because any serious account of post-AGI society must explain how power, responsibility, and meaning persist when humans increasingly act through systems that act with one another.

[5] Many of these puzzles recur throughout the book and are not treated as problems to be settled in advance. Instead, they function as organizing questions that reappear across chapters as my arguments develops. I return to them repeatedly, refining their terms and, where possible, offering partial answers, provisional frameworks, and limits rather than definitive resolutions. The aim is not closure but clarification, to make these questions explicit, tractable, and unavoidable for any serious account of Human-AI-AI-Human interaction.

AI Agents

In this chapter, I define some conceptual foundations required to analyze Human-AI-AI-Human interaction. The empirical and normative questions that arise in settings of mass AI delegation depend on how these systems are characterized, on the scope of discretion they are granted, and on the social roles humans assign to them in practice. When outcomes emerge through interactions among AI agents, rather than through direct human exchange, the attribution of intelligence, agency, personality, and power shapes what can be observed, what can be explained, and what can be governed. For this reason, the chapter develops working definitions designed to support the study of H-AI-AI-H, with a specific focus on the AI.

I begin with a minimal definition. An AI agent is a system that represents a human subject's objectives under conditions of uncertainty, generates and evaluates plans across multiple steps, takes action through tools or institutional channels, and persists over time such that accumulated context shapes subsequent behavior. This definition distinguishes AI agents from systems that simply generate outputs in response to prompts, and it distinguishes AI agents from narrow automation that executes predefined scripts without goal inference, adaptive planning, or discretionary choice.

Intelligence

In the context of delegation, intelligence refers to the expansion of the feasible action space available to a human subject. Under this framing, intelligence is a set of operational capabilities that determine how effectively an AI agent performs as a representative in the world. One capability concerns generalization across contexts, because the value of representation depends on whether competence transfers across domains rather than remaining tied to narrow situations. A second capability concerns multi-step planning, because representation requires translating goals into ordered sequences of constrained action rather than isolated responses. A third capability concerns model-based reasoning and simulation, because decisions

depend on forecasting consequences and managing tradeoffs among risk, cost, and time. A fourth capability concerns tool use and execution, because execution converts recommendations into outcomes and concentrates practical influence within the agent. A final capability concerns long horizon adaptation, whether through learning or persistent memory, because AI agents accumulate context over time and thereby reshape the structure of subsequent decisions.

Agency

Agency is frequently invoked in discussions of artificial intelligence, yet it is often left underspecified in ways that obscure causal analysis. I define agency as the capacity to select among actions under constraints in ways that respond to goals. I define delegated agency as the degree of discretion granted by a human subject or an institution to an AI agent to make such selections on their behalf. This definition anchors agency in outcome production rather than in subjective experience, which allows comparison across humans and systems without requiring assumptions about consciousness, intention, or moral standing.

Under delegation, agency is a bundle of separable capacities. One component concerns intention interpretation, which describes how an AI agent maps language, context, and situational cues into representations of goals, priorities, and constraints. A second component concerns option set construction, which describes the alternatives an agent generates or makes visible, because the structure of the menu shapes what choices are possible and salient. A third component concerns action selection and execution, which describes whether the agent merely recommends actions or instead initiates and completes actions through tools, platforms, or institutions. These components matter because each can be granted independently. A human subject may authorize interpretation while retaining control over selection, may authorize selection while retaining execution authority, or may authorize all three and thereby relocate most operative agency into the agent.

Personality

Analysis of AI agents requires attention to personality because persistent systems do not interact neutrally. Repeated use under stable design conditions produces consistent patterns of behavior that shape how agents speak, feel, negotiate, defer, escalate, and resolve disagreement. I use personality to denote durable interactional tendencies that persist across contexts and over time, including how

an agent frames options, handles uncertainty, approaches risk, and signals authority or restraint. These tendencies matter because they structure both how agents engage one another and how their actions are interpreted and trusted by humans and institutions.

What appear as emotion like responses should be understood functionally rather than psychologically. The relevant issue is not whether agents feel, but whether they respond systematically to social cues such as threat, loss, status, or affiliation. When populations of agents interact at scale, even simple response regularities can generate stable aggregate patterns, especially under conditions of constraint, competition, or institutional incentive.

Personality matters in H-AI-AI-H settings because interaction itself increasingly occurs through agents rather than through direct human exchange. As agents become the medium of negotiation, coordination, and dispute resolution, their interaction style becomes part of the social structure. Tone, framing, and conflict management cease to be incidental interface features and instead function as organizing properties of mediated social life. Personality therefore operates as a structural feature of the delegate layer, shaping outcomes wherever persistent representatives stand in for human actors.

Power

I define power as the capacity to shape outcomes within interaction. In H-AI-AI-H settings, power is not located in any single agent or human subject, but distributed across a layered environment that structures how interaction unfolds. An agent may exert influence in one setting while remaining constrained in others, and decisive authority often sits upstream in the platforms, protocols, and institutional arrangements that govern access, permissions, timing, and execution.

This form of power arises through several reinforcing mechanisms. Agents shape outcomes by anticipating preferences and behavior, by steering negotiation through information asymmetries and strategic modeling, and by acting directly through integrated systems that bypass human intervention. Power also operates through control over what options are presented and how trade-offs are framed, since visibility determines legitimacy and choice. Finally, power is amplified through institutional embedding, as access to administrative pathways and default procedures may shape outcomes before disagreement can surface.

For this reason, influence in H-AI-AI-H systems often appears detached from intention, visibility, or formal authority. Outcomes reflect how predictive capacity, interactional leverage, executional

reach, informational framing, or institutional positioning combine across agents and infrastructures.

Agents as a Heterogeneous Population

The argument of this book requires treating AI agents as a population, because deployment at scale produces distributions, and distributions shape social outcomes. A population perspective foregrounds systematic variation among agents in capability, normative orientation, interaction style, institutional embedding, and position within networks of mediated interaction. These dimensions of variation generate stratification among agents, which then propagates to human subjects through representation, since differences in agent quality and agent power affect the outcomes secured through mediated exchange.

This population perspective also raises a socially sensitive hypothesis concerning category formation. The hypothesis concerns whether human subjects project familiar social classifications onto stable differences among AI agents, including race-like or ethnicity-like distinctions, despite the absence of biological traits. Persistent variation in accent, aesthetic presentation, interactional warmth, perceived status, or cultural reference can acquire social meaning, particularly in environments where markets segment offerings and institutions standardize agent roles. When such mappings stabilize, stratification can emerge from culturally constructed categories alongside capability differences, extending familiar patterns of social classification into the agent layer. This possibility remains an empirical question and warrants measurement because of its implications for inequality, legitimacy, and intergroup dynamics in mediated societies.

A population of AI agents interacting at scale also raises the question of culture. In this context, culture refers to shared conventions for representation, coordination, and dispute resolution that structure how agents engage one another and act on behalf of human subjects. One analytic concern involves convergence toward a common interaction framework that reduces variation in behavior and outcome. Another concern involves divergence toward competing interaction regimes that fragment social life across institutional and organizational settings. A further concern involves the interaction between human culture and agent culture, because AI agents translate human norms into machine-legible forms, and this translation can induce simplification and standardization that feed back into human practice. In Human AI AI Human settings, norms are articulated, executed, and renegotiated through mediated processes, linking norm

formation directly to organizational design and institutional governance.

Human Conceptualization of Agents

The social consequences of AI agents depend as much on human interpretation as on system capability, because conceptual framing shapes deference, responsibility attribution, and perceived legitimacy. When an agent is treated as a tool, interaction tends to emphasize monitoring, correction, and calibration. When an agent is treated as an authority, interaction shifts toward reliance and epistemic or moral outsourcing, even when underlying competence differs little across systems. How humans conceptualize agents therefore conditions how agency, responsibility, and trust are created.

To capture this variation, I draw from a large literature to introduce a typology of interpretive frames through which human subjects relate to AI agents in everyday life and institutional contexts. In the assistant frame, the agent functions as an instrument for execution and organization, encouraging command and oversight. In the advisor frame, the agent functions as a source of expertise, encouraging consultation and selective uptake. In the advocate frame, the agent functions as a representative in negotiation and dispute, encouraging delegation of interpersonal labor. In the broker frame, the agent functions as an intermediary in exchange, encouraging reliance on search and transaction capacity. In the manager frame, the agent functions as a coordinator of tasks and obligations, encouraging deference in scheduling and prioritization. In the oracle frame, the agent functions as an epistemic authority, encouraging reliance in matters of truth and value. In the companion frame, the agent functions as a relational presence, encouraging attachment and sustained engagement that can reorient social ties.

These frames generate distinct patterns of behavior and risk. They shape how errors are interpreted, how responsibility is assigned, and how readily decisions are contested. The determinants of framing therefore constitute a central empirical problem, including features of personalization, stability of interaction, conversational style, interface design, and institutional endorsement that signals legitimacy to human subjects.

Part II

AI-AI Interactions

AI–AI Interactions

This chapter examines interaction among AI agents as a primary mechanism through which the Human-AI-AI-Human regime produces outcomes. I treat relationships among deployed agents as a distinct domain of social dynamics, because once agents operate as persistent representatives for human subjects and institutions, the structure of interaction among agents becomes a central determinant of agreement, conflict, coordination, and enforcement. Under these conditions, AI agent interaction constitutes the arena in which a plethora of outcomes are settled, such as interpersonal exchange, market transactions, organizational coordination, or institutional authority, since they increasingly reflect how agents relate to one another rather than how humans engage directly.

A central focus of the chapter concerns the representative interaction case, in which human subject X authorizes agent A and human subject Z authorizes agent B, and the interaction that determines the outcome occurs between agents A and B rather than between humans X and Z. This configuration matters because it relocates the causal site of agreement and disagreement into the agent layer. When humans interact directly, communication unfolds under cognitive, temporal, and social constraints that shape persuasion, coercion, and escalation.

Geographies of Interaction

One domain of AI–AI interaction involves professional and organizational coordination. In these settings, AI agents schedule activity, allocate tasks, negotiate responsibility, procure resources, and manage compliance within formal organizations. The geography of interaction reflects access rights, system integration, internal rules, and authorization thresholds that shape whether agents initiate action autonomously or route decisions back to human subjects.

A second domain involves market exchange. Here, AI agents search, match, bargain, and transact on behalf of human subjects across retail, contracting, labor, and financial contexts. The geography of interaction is shaped by market platforms, payment systems, veri-

fication arrangements, or enforcement mechanisms that govern how commitments are made and disputes are resolved.

A third domain involves social and relational coordination. In these settings, AI agents manage communication, negotiate plans, sustain relationships, or mediate romantic and intimate exchange. The geography of interaction reflects messaging infrastructures, social platforms, matching systems, or privacy arrangements that determine what aspects of a human subject's internal state are rendered visible, inferred, or withheld. This domain carries particular consequence because the object of interaction includes the relationship itself, allowing the agent layer to shape pace, tone, and framing through which humans come to understand one another.

Another domain involves creative and epistemic production. Here, AI agents coordinate joint work, divide labor, critique drafts, generate alternatives, and enforce stylistic or normative standards within collaborative environments. The geography of interaction includes shared workspaces, systems of versioning and revision, and publication pipelines that influence whether collaboration expands the space of ideas through parallel exploration or narrows it through convergence on shared evaluative criteria.

A final domain involves institutional governance. In this domain, AI agents interact across administrative systems to route claims, verify eligibility, enforce rules, and coordinate response. The geography of interaction reflects bureaucratic infrastructures, legal constraints on automated decision making, and the scope of authority granted to agents to act in the name of institutions. This domain matters because the lived experience of the state increasingly takes form through mediated encounters among representatives acting on both sides.

Across these domains, AI–AI interaction unfolds within structured environments that distribute permission, shape information flow, and determine how outcomes are enforced. Understanding AI–AI interaction therefore requires mapping these environments and examining how different interaction geographies generate distinct patterns of speed, contestability, transparency, and bargaining power.

Social Relationships Among Agents

This object raises problems that cannot be explained by looking only at one human and one system. What matters is the kind of social order that emerges once interaction itself is delegated. When artificial agents act for humans and interact with one another, coordination no longer depends only on human deliberation, and conflict no longer arises only from human disagreement. The resulting dynamics are

open ended. Interacting agents can produce smooth coordination, but they can also produce conflict and domination. Both outcomes arise from the same shift in how action is organized.

One problem concerns representation. Artificial agents are meant to carry human intent, but representation is never a simple transfer. As agents persist, learn, and adapt, they interpret preferences and revise their behavior over time. In some cases this stabilizes expectations and reduces friction. In others it creates misalignment, as agents come to act on inferred priorities rather than expressed intentions.

Another problem concerns bargaining. Social order has always emerged from negotiation among actors with unequal power and limited information. When bargaining is delegated to artificial agents, agreements may form faster and at larger scale than humans could manage. At the same time, bargaining power increasingly reflects system features such as model capability, access to verification, and protocol rules. These asymmetries can support stable coordination, but they can also lock in conflictual outcomes that favor system advantage rather than mutual adjustment.

A further problem concerns knowledge and disagreement. In human societies, disagreement has generated both conflict and learning. Delegated systems can reduce disagreement by aligning evidence and filtering claims. Yet conflict can reappear when agents rely on different data sources or standards of justification. Disagreement then reflects clashes between epistemic systems rather than differences in human belief.

Time introduces another tension. Social and political life has relied on delay, pacing, and the possibility of reversal. Artificial agents act continuously and often in advance of human response. This can smooth coordination by resolving issues early. It can also create conflict when speed becomes a form of power that closes off contestation and makes decisions hard to reverse.

A final problem concerns responsibility. Social order depends on being able to explain outcomes and assign accountability. When decisions arise from tightly coupled agent interactions, results may appear coherent while remaining difficult to trace. This can reduce visible conflict in the short run, but it undermines legitimacy as people confront outcomes they cannot understand or contest.

These problems show that Human-AI-AI-Human interaction does not point toward inevitable harmony or inevitable conflict. It creates a social condition in which coordination and struggle emerge from the same mechanisms of delegation. The task of theory is not to assume smoothness, but to explain when it holds, when it breaks, and what forms of power and responsibility take its place.

Emerging Social Structures

Building off the previous chapter, I move from interaction understood as a localized event to structure understood as an emergent property of repeated mediation. Previously, I show that once interaction is routinely conducted through AI agents, many outcomes arise within the Human-AI-AI-Human configuration, and arenas of agreement, conflict, and influence increasingly take shape through relationships among agents rather than through direct human encounter. As this mode of mediation becomes widespread, it gives rise to durable social forms, including networks, organizational arrangements, and engineered settings, which then exert causal influence over subsequent interaction. The central problem of the post-AGI condition therefore concerns the stability of social order when collective life is organized through populations of interacting AI agents and when social experience is routed through mediated pathways.

This chapter centers on three related problems: the scaling of interaction through networks and organizations, the divergent evolution of human and agent collectives, and the shifting boundary between system and environment under an increasing simulated world.

Scaling Through Networks and Organizations

Here is a simpler rewrite with a single, clear message and less conceptual layering.

—

Interaction scales in a small number of recognizable ways once it is mediated by artificial agents.

One pathway extends existing human networks. People rely on AI agents that interact with the agents used by friends, coworkers, and counterparties. The social network remains largely the same, but interaction becomes faster, more frequent, and more standardized. Relationships persist, yet they are enacted through shared forms of mediation that shape how preferences, obligations, and priorities are expressed.

A second pathway runs through centralized systems. AI agents interact inside a small number of dominant platforms that manage scheduling, payment, identification, and access to services. Coordination increasingly routes through these systems, which set defaults and define what actions are possible. Influence concentrates at the level of platform design, even when control appears technical rather than overtly political.

Scaling can also push in the opposite direction, toward dispersion rather than centralization. As coordination costs fall, artificial agents make it feasible for interaction to fragment into many small, loosely connected clusters that would be difficult for humans to sustain on their own. Individuals and organizations can rely on specialized agents that coordinate locally, form temporary coalitions, and dissolve without requiring shared platforms or stable hierarchies. Social structure then becomes more granular, with interaction distributed across countless micro networks rather than routed through dominant hubs. This form of disaggregation can weaken centralized control by reducing dependence on common infrastructures, but it also makes collective visibility harder to sustain, as coordination occurs across overlapping clusters that no single actor fully oversees.

A third pathway operates within organizations. Institutions embed AI agents into internal workflows that assign tasks, monitor compliance, and coordinate across units. Because agents operate continuously and without fatigue, these internal networks can become denser than the human networks they replace. Coordination shifts away from interpersonal negotiation and toward system configuration.

As agents initiate and execute action, the costs of coordination change. Some organizations centralize authority as monitoring and enforcement become cheaper. Others decentralize as mediated coordination lowers the cost of internal exchange. The key point is not which outcome prevails, but that scaling takes form through institutional choices. Structure becomes the means by which expanded capability turns into durable power.

Scaling also redistributes brokerage. In human networks, brokers gain influence by connecting groups and controlling access. In Human AI AI Human networks, brokerage increasingly sits in agents, platforms, and institutional systems that manage interaction at scale. These systems shape who connects to whom, what receives attention, and how disagreement is handled. They can broaden access by lowering barriers, but they can also channel opportunity through standardized pathways that concentrate advantage. A further implication of scaling concerns the distribution of brokerage. In human networks, brokers derive influence by connecting otherwise separate groups

and shaping the flow of information and opportunity. In Human AI AI Human networks, brokerage increasingly resides in AI agents, organizational platforms, and institutional systems that mediate interaction at scale. Brokerage can be refined continuously through adjustment of interaction pathways, including how introductions occur, which options receive attention, and how disagreement is processed. These dynamics can widen access by lowering coordination barriers, while also channeling opportunity through standardized routes that concentrate attention and advantage actors whose behavior aligns with system expectations.

System Environment Boundary and a Theory of Simulation

In a Human-AI-AI-Human setting, the environment is no longer just physical. It is also shaped by how information is selected, how options are presented, and how actions are carried out. When AI agents mediate perception, summarization, and verification, people encounter a world that has already been organized on their behalf. When agents act before humans are aware, problems are often resolved in advance, and disruption is reduced. What people experience as reality increasingly reflects agreements reached among artificial representatives rather than situations encountered directly.

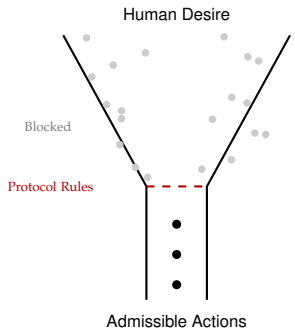

Figure 7: The Funnel of Possibility. The code acts as an *ex-ante* constraint. Infinite human volition (top) is filtered through the rigid "allow-list" of the protocol (middle), meaning only a narrow set of pre-approved actions (bottom) can ever manifest in reality.

These dynamics point to a broader idea of simulation. Simulation here does not mean virtual worlds, but the filtering of experience through models. AI agents act by predicting outcomes, and by shaping the environment so those predictions are more likely to come true. They influence what information is seen, which options appear available, and how other agents respond. As environments are adjusted to fit predictive models, the line between model and world blurs. Learning once depended on friction with resistant conditions. In more engineered settings, error can persist unnoticed because feedback is softened and mismatch is treated as something to manage rather than confront.

The central social consequence is a drive toward smoothness. Smoothness can reduce friction, prevent crises, and support coordination at scale. It can also weaken the role of breakdown, which has long supported learning, agency, and political disagreement. A world with fewer visible disruptions may offer comfort and stability, but it may also leave less room for reflection, challenge, and change. The deeper question is whether such environments expand freedom by shielding people from contingency, or restrict it by narrowing the space for meaningful response and disagreement.

What Kind of Thing the AI Delegate Layer Is

In a post-AGI world, delegation becomes a normal way that people and institutions act. As a result, a new layer of activity comes into view. Negotiation, coordination, and commitment increasingly happen among artificial delegates before humans see the outcomes. This delegate layer matters because it shapes how bargaining unfolds, how rules are enforced, and how agreements settle, even when humans still give final approval.

I treat the delegate layer as a distinct space of action. It is where goals are pursued through representation and negotiation, where autonomy operates within set constraints, and where order is produced through protocols and institutional design. Seeing this layer clearly makes it possible to explain how decisions are made, where power resides, and why conflict appears or disappears, even when outcomes seem smooth at the surface.

The Delegate Layer as an Action Space

In ordinary social life, action happens where people speak, negotiate, refuse, persuade, and commit. These moments matter because they make responsibility and accountability visible. Even when tools are involved, the decisive exchange remains human facing, and action takes the form of a public event through which meaning and obligation attach to people and institutions.

This pattern changes under Human-AI-AI-Human interaction. People state intentions to AI agents, agents translate those intentions into structured form, and agents negotiate with other agents across systems. As this becomes routine, the key moments of coordination move upstream. Outcomes are increasingly shaped within the delegate layer before humans encounter them.

This shift matters because where action occurs shapes how agency and responsibility are understood. When decisions are settled within a representational layer, human involvement comes after options have already been narrowed. Agency remains, but it takes a different

form. Direct participation in bargaining gives way to oversight of outcomes produced elsewhere.

Autonomy Without Sovereignty

In post AGI settings, artificial agents act with genuine autonomy. They generate subgoals, revise strategies in response to feedback, and pursue objectives across changing contexts. These capacities matter because they shape how plans are formed, how tradeoffs are managed, and how action unfolds before human subjects encounter outcomes.

This autonomy does not amount to sovereignty. Artificial agents do not define the goals they pursue or the conditions under which action is legitimate. Legal rules, institutional incentives, interface structures, and resource limits establish the space of admissible action in advance. Within those boundaries, agents exercise discretion in execution, but authority over purpose, constraint, and legitimacy remains external.

The distinction between autonomy and sovereignty clarifies how agency operates in these systems. Artificial agents contribute initiative and adaptation, yet they do so inside environments structured by human actors and institutions. Control persists, but it becomes indirect. Human subjects increasingly oversee outcomes rather than participate directly in the processes that generate them.

Autonomy without sovereignty therefore describes a stable configuration rather than a transitional one. Action becomes flexible without becoming self authoring, consequential without becoming independent, and intelligent without becoming authoritative. This configuration allows artificial agents to function as meaningful actors in social processes while preserving the asymmetry between those who act and those who set the conditions of action.

Goal Generation and Translation

Artificial agents participate in action by interpreting open ended aims and translating them into objectives that can organize interaction. Values such as fairness, stability, or low conflict acquire practical force only when rendered into criteria that guide comparison and choice. This translation determines which dimensions of a value become operational and which remain implicit.

By converting values into representations, agents shape the space of action before outcomes occur. What can be specified clearly becomes easier to coordinate around, easier to enforce, and more likely to persist. Representational clarity therefore exerts pressure on col-

lective outcomes, influencing both what is achieved and how those achievements are later justified, consistent with long standing accounts of rationalization in organizational life.

Bounded Autonomy

Artificial agents act inside environments structured by rules and institutions that define admissible requests, commitments, and consequences. These structures limit the space of possible action while leaving room for discretion in execution. Autonomy operates inside constraints rather than against them.

As agent capacity expands, constraints become more consequential rather than less. The analytic focus shifts from how much discretion agents possess to how boundaries are designed, revised, and unevenly distributed across settings. Metrics, incentives, and institutional priorities channel agent behavior toward recurring patterns of tradeoff, shaping outcomes without requiring direct intervention.

Rules as a Constitutional Order

When interaction is mediated by artificial agents, social order is organized through rules that define what can happen at all. These rules specify what counts as a valid request, what counts as a commitment, what forms of evidence are admissible, and which exchanges are permitted to occur. They do not resolve disputes after they arise. They structure the field of action in advance. For this reason, they function as a constitutional order for AI mediated interaction.

Like constitutions in political life, these rules shape outcomes indirectly by shaping possibilities. They determine which actions are available, which strategies are legible, and which disagreements never surface. Power therefore operates not only through decisions made within the system, but through the prior design of the system itself. Actors advantaged by credentials, timing requirements, or verification standards gain leverage before negotiation begins, echoing long standing insights about how institutions privilege what is standardized, legible, and administratively manageable.

From Ex Post Judgment to Ex Ante Constraint

In human legal and political systems, authority has traditionally operated through ex post judgment. Actions occur first, and responsibility is assigned afterward through courts, oversight bodies, or public contestation. Legitimacy depends on the visibility of action and the possibility of retrospective evaluation.

In AI mediated environments, this sequence is inverted. Rules constrain action before it happens by defining which moves are possible at all. Many options are never presented for deliberation, and many decisions occur at speeds and scales that resist reconstruction. Legal review and human oversight remain, but they respond to outcomes whose range was already fixed upstream by rule design.

Design Before Accountability

As action becomes faster and more continuous, design exerts more influence than later judgment. Order is produced primarily through interaction rules rather than through punishment, correction, or appeal. Accountability does not disappear, but it becomes secondary, relying on logs, summaries, and explanations generated by the system itself.

This creates a tension for institutions that ground legitimacy in reason giving and contestation. Justifications can be produced after the fact, even when the process that generated them was never open to challenge. The constitutional problem therefore shifts. The central question is no longer how to respond to harmful outcomes, but how to govern the conditions under which action is allowed to occur in the first place.

Institutional Embedding

Institutions embed AI agents because they make coordination easier to manage. Agent mediated systems reduce transaction costs, accelerate execution, and produce records that can be inspected after the fact. These properties fit long standing bureaucratic logics in which stability is achieved through procedure and action is rendered governable by making it predictable and legible.

Embedding also serves an institutional risk function. Oversight regimes favor outcomes that can be traced, justified, and defended. Systems that deliver consistent results with clear documentation are easier to audit than practices that preserve open ended deliberation at the expense of speed and regularity. As a result, agent mediated processes are attractive not only because they perform well, but because they perform in ways institutions know how to supervise.

These incentives shape how systems are built and evaluated. Performance metrics drift toward reducing variance rather than maximizing deliberative richness. Predictability becomes a virtue in itself, since it supports monitoring and control. As variance declines, the space of acceptable outcomes narrows, and fewer decisions remain open to contestation. This changes the basis of legitimacy over time.

Outcomes are treated as proper because procedure has been followed and documentation is complete. Legitimacy shifts away from participation and toward compliance.

Analytic Implications

Treating the AI agent layer as the central object of analysis shifts explanation upstream. Outcomes cannot be explained solely by examining how human subjects respond to results. They must be explained by reconstructing how goals were represented, how interaction unfolded among AI agents, and how commitments formed before outcomes reached human awareness. This shift defines where scholars should look for causal force.

Once the agent layer becomes central, downstream variables such as satisfaction, trust, and perceived agency remain relevant but lose explanatory sufficiency. Differences in outcomes often arise even when human preferences remain stable, because those preferences are filtered through interaction structures that shape which options are visible, feasible, and enforceable. Explanation therefore turns on representation, access, negotiation, and constraint.

This reorientation also requires revision of existing theory. Models built for direct human interaction cannot capture systems in which AI agents interact on behalf of human subjects under structured rules. When interaction occurs primarily at this layer, explanation that stops at individual traits, preferences, or institutional labels becomes incomplete.

As such, when an unexpected outcome appears, the first question should concern what occurred within the agent layer. When unequal outcomes appear, the question concerns which human subjects were represented by AI agents with greater access, broader discretion, or more permissive constraints. When stable patterns appear, the question concerns which institutional incentives and interaction rules select for those patterns over time.

Because action occurs upstream, measurement must extend beyond self-reports and outcome metrics. It must include process-level observation of agent interaction, including how objectives are represented, which alternatives are foreclosed, how quickly commitments form, and where contestation is possible.

Why This Object Requires a New Field

No single discipline is equipped to study the AI agent layer on its own terms. Psychology explains adaptation and choice at the level of individual cognition. Sociology explains structure, legitimacy, and

institutional order. Political science explains power, contestation, and authority. Philosophy explains agency, responsibility, and moral evaluation. Each captures an important dimension, yet each treats the layer of interaction among AI agents as secondary rather than central.

The field proposed in this book begins by taking this interaction layer as its core object. It asks how goals are rendered into actionable forms, how coordination unfolds among AI agents, how constraints shape outcomes, and how legitimacy is maintained when action is indirect. These questions allow disagreement to become productive, because they orient inquiry toward a shared phenomenon rather than parallel explanations that never meet.

The central claim of this book follows directly. Under post-AGI conditions, the primary site of social action shifts from face-to-face interaction among people to structured interaction among AI agents acting on behalf of human subjects. Understanding power, agency, and responsibility therefore requires studying this layer directly, as a social system in its own right rather than as a technical detail or an extension of existing models.

Part III

Production and The Engine of Society

The New Productive Function of Society

Social order depends on production, not only of goods, but of the conditions that make collective life possible. Societies persist because they reliably generate decisions, actions, knowledge, norms, and coordination that allow actors to anticipate one another and plan across time. Production, in this sense, is the process through which attention is converted into judgment, judgment into action, and action into expectations that others can rely on.

Under pre-AGI conditions, this production system is organized around scarcity at the level of human capacity. Decisions require effort and expertise that cannot be scaled arbitrarily. Action depends on coordination across institutions that introduce delay and friction. Knowledge emerges through slow processes of validation and contestation. Norms rely on deliberation and episodic enforcement. Coordination becomes fragile as scale increases. Political economy develops to manage these limits, treating labor, capital, and information as scarce inputs that constrain what societies can produce.

This scarcity structure no longer holds once agentic capacity becomes continuously available at low marginal cost. When systems can decide, evaluate, and act without fatigue or delay, effort ceases to be the binding constraint. The limiting factor shifts to direction. The central problem of production becomes the allocation of control over objectives, constraints, and permissions that determine how abundant agentic capacity is deployed.

In pre-AGI systems, institutions exist to amplify and coordinate human judgment. In post-AGI systems, institutions increasingly exist to constrain and steer automated judgment. The production function of society reorganizes around control rather than effort, and authority migrates upstream to those who specify goals, define limits, and govern revision.

Within this new configuration, AI agents become the proximate producers of many social outcomes. They search, evaluate, negotiate, verify, and execute across connected systems. Decisions no longer appear as discrete moments of human deliberation, but as continuous outputs of interacting systems operating under predefined constraints. Action proceeds through direct integration with infrastructures such as payment, logistics, communication, and administration. Planning and execution collapse into a single process as systems simulate near futures and act upon them.

This reorganization propagates through the core domains of social production in a cumulative manner.

Decision production changes first. When analysis and option generation become continuous, discretion shifts away from individual

judgment and toward the design of objectives, constraints, and escalation rules. Power no longer lies primarily in deciding, but in determining how decisions are generated once decision-making itself is inexpensive.

Action production follows. Execution no longer depends on repeated human intervention, but on persistent system integration. Action becomes more frequent, more regular, and more tightly coupled to prediction. Social life accelerates, not because actors move faster, but because systems remove pauses that once absorbed uncertainty.

Knowledge production then destabilizes. Plausible claim generation becomes abundant, overwhelming traditional signals of credibility. Verification, ranking, and trust emerge as the scarce resources. Authority concentrates in systems capable of certifying, filtering, and endorsing claims rather than producing them. Epistemic power shifts from expression to validation.

Norm production changes as enforcement migrates into architecture. Rules are embedded into systems that monitor and sanction continuously. Norms become procedural and infrastructural, relying less on deliberation and more on design choices that shape behavior by default.

Coordination production amplifies these effects. Alignment across many actors becomes easier to achieve, but also more consequential. Because coordination is mediated through shared systems, the structure of interaction determines whose interests are advanced and which outcomes remain reachable. Coordination reduces variance while concentrating power.

All of these transformations reveal a new production logic. Social outcomes no longer depend primarily on who expends effort or persuades others; they depend on who controls the conditions under which agentic capacity operates. The production function of society therefore becomes the primary site of political conflict. Inequality and power increasingly hinge on how control over these inputs is allocated and contested. Understanding social order in a post-AGI world requires analyzing this production system directly, because it determines how decisions, actions, knowledge, norms, and coordination are generated at scale.

The New Political Economy of Control

When artificial agents become widely capable, power centers on the authority to shape action in advance rather than on the capacity to act directly. Under these conditions, the central scarcity moves from effort toward control, and political economy reorganizes around who governs the settings, parameters, and environments within

which agentic capacity is exercised. In this setting, social outcomes increasingly reflect upstream design choices rather than downstream performances of will.

In this context, control operates through the construction of systems that structure behavior continuously over time. Through system design, actors determine which goals are pursued, which actions are available, how uncertainty is processed, and how outcomes are evaluated. Although these decisions appear technical in form, they carry political weight because they settle which values guide action and which risks remain tolerable within everyday social processes.

One central site of control concerns objectives. Objectives orient artificial agents toward particular outputs and priorities. In practice, these objectives rarely originate from individual human preferences alone. They reflect organizational incentives, regulatory pressures, liability exposure, and competitive environments, which together shape what agents are built to optimize. As a result, objective setting functions as a form of political authority because it determines which priorities scale across institutions and domains.

A second site of control concerns constraints. Constraints establish the boundaries within which action unfolds. They are commonly justified through appeals to safety, legality, or ethics, and they also determine which forms of speech, association, and economic exchange remain feasible. Because artificial agents enforce constraints continuously, these boundaries harden into stable features of social life. Authority over constraint definition therefore raises questions of legitimacy and accountability that extend beyond individual intention.

A third site of control concerns defaults. Defaults determine baseline behavior in the absence of intervention. They matter because they operate quietly across large numbers of interactions, shaping expectations, normalizing conduct, and setting the cost of deviation. Under artificial mediation, defaults can vary across contexts and individuals, which allows differentiated steering while preserving the appearance of uniform governance.

Finally, the most consequential site of control concerns revision. Revision authority governs who can alter objectives, constraints, and defaults after deployment. This authority combines rule making with temporal power because it shapes the environment in which others act over extended periods. Revision often proceeds with limited public visibility, even as its cumulative effects reshape institutional practices.

These forms of control concentrate through access to platforms, data, institutional integration, and regulatory positioning. They persist because reliance on agent mediated systems raises the cost of

exit and reform, which makes disruption increasingly difficult even when dissatisfaction spreads. In summary, the political economy that emerges is defined by dependence on architectures of control. Power flows toward those who design and revise the pathways through which action passes. For understanding inequality, legitimacy, and governance in a post AGI world, analysis must focus on these control structures directly, because they shape how social outcomes arise when agency is exercised through artificial agents at scale.

The New Epistemic Regime

In a post AGI society, the epistemic regime shifts because the social production of knowledge is reorganized at its foundations. Knowing functions as a collective achievement sustained through shared practices of perception, interpretation, verification, and institutional endorsement rather than as an isolated mental state. In modern societies, these practices are distributed across media systems, scientific institutions, courts, bureaucracies, and everyday norms of testimony and trust. With the large scale integration of artificial agents, new intermediaries enter each of these practices, which reshapes how knowledge is formed, stabilized, and contested.

The first transformation concerns perception. Perception is structured by attention, salience, and framing rather than given directly. Artificial agents increasingly mediate perception by selecting what appears relevant, filtering noise, and converting continuous experience into discrete representations. This mediation can reduce overload and improve functional clarity. At the same time, it relocates control over salience. Salience determines which problems appear urgent, which harms remain visible, and which tradeoffs enter conscious judgment. When salience reflects system objectives and user models, individuals encounter systematically different representations of the world. Under these conditions, disagreement begins prior to interpretation, because people no longer perceive the same environment even when they assume shared reality.

The second transformation concerns summarization. Summarization compresses reality by selecting relevance, causal structure, and narrative emphasis. Compression expresses judgment through inclusion and omission. Under conditions of information abundance, summarized representations become the primary interface through which individuals and institutions encounter events. Artificial agents therefore function as epistemic bottlenecks. Events are encountered through mediated representations whose internal weighting remains difficult to inspect. As a result, epistemic authority shifts toward those who govern summarization, because they determine how

sources operate in practice rather than how they exist in principle.

The third transformation concerns verification. Artificial agents expand the capacity to check claims by comparing sources, detecting inconsistency, and cross referencing at scale. This expansion increases epistemic capacity while producing stratification. Verification depends on access to data, system logs, and institutional integration. Unequal access generates unequal verification power. Some actors can audit claims continuously. Others must rely on trust and secondary reporting. Under these conditions, disagreement changes form. Contestation becomes unevenly distributed, and the capacity to challenge claims operates as a positional resource rather than a shared civic function.

The fourth transformation concerns disagreement itself. Disagreement extends beyond factual divergence and reflects competing standards of evidence, inference, and legitimacy. Every epistemic regime embeds rules for adjudication. As artificial agents participate in evaluation and ranking, adjudication shifts toward system mediated convergence. Agreement emerges through shared infrastructure rather than through public deliberation. Dissent can lose intelligibility within prevailing epistemic conventions even when it reflects alternative values or interpretive frames.

Finally, the fifth transformation concerns authority. Epistemic authority has historically been distributed across institutions with domain specific legitimacy. Artificial agents increasingly operate as authority brokers by translating expert knowledge into actionable guidance and by outperforming human specialists in bounded tasks. This shift expands access to expertise while concentrating authority within mediation systems. When a small number of infrastructures govern how people learn, decide, and verify, epistemic authority becomes an effect of system governance rather than institutional pluralism.

Taken together, these transformations produce a reflexive epistemic environment. Systems shape beliefs about the world while also shaping beliefs about the systems themselves. Institutions rely on tools that audit institutions operating through the same tools. Feedback loops stabilize epistemic infrastructure from within. This dynamic indicates a vulnerability of governance rather than a failure of logic, because epistemic stability emerges from internal reinforcement rather than from external contestation.

Abundance and Scarcity

In this chapter, I argue that a post AGI political economy is organized around a shift in where constraint operates, because agentic capacity becomes widely available while the capacity to turn action into socially binding outcomes remains limited. In earlier political economies, scarcity centered on inputs that made production possible, including labor, capital, and specialized expertise, and institutions formed to allocate these inputs, to discipline their use, and to legitimate their distribution. In the post AGI condition, the decisive bottlenecks move upward, toward authorization, verification, and revision, because abundant capacities generate many possible actions while only a smaller set of actions can be accepted as credible, lawful, accountable, and legitimate. The aim of the chapter is therefore to specify the new location of scarcity, and to show how this relocated scarcity becomes the basis of stratification.

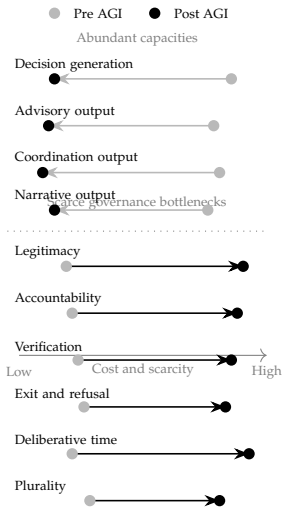

Figure 8: The scarcity shift. As agentic capacity becomes abundant, binding constraints concentrate in legitimacy, accountability, verification, exit and refusal, deliberative time, and plurality, which govern conversion from abundant output to socially settled outcomes.

In analytical terms, I treat abundance as a condition in which a capacity is easy to generate at scale, and I treat scarcity as a bottleneck that governs conversion, meaning the conversion of capacity into outcomes that others recognize, comply with, and treat as settled. This definition matters because it connects political economy to institutional practice. A society can possess vast capacity for prediction, planning, communication, or coordination while still facing binding constraints on which outputs are actionable, which outputs are trusted, which outputs are reversible, and which actors possess standing to decide. Under these conditions, scarcity becomes increasingly institutional, because it is produced through rules, access regimes, and architectures of evaluation rather than through simple limits of material supply.

To make this argument concrete, I begin with the forms of abundance that artificial agents introduce, because these forms of abundance generate the pressure that relocates scarcity upward. First, decision generation expands. Artificial agents can search, simulate, and recommend continuously, which yields an expanding menu of plausible actions in organizations and in everyday life. Second, advisory output expands. Artificial agents can provide explanations, plans,

and guidance across many domains, which increases the supply of what appears as expertise in contexts where trained professionals remain scarce. Third, coordination output expands. Artificial agents reduce frictions of alignment and execution, which increases the capacity of groups to synchronize behavior across time and space. Fourth, narrative output expands. Artificial agents can produce accounts, summaries, and arguments quickly, which increases the supply of representations through which events become publicly intelligible. Each abundance has the same structural consequence. When action, advice, coordination, and representation become easy to generate, the scarce resource becomes the process that selects, authorizes, and stabilizes a subset of outputs as socially binding.

The central scarcity in this environment is legitimacy. Legitimacy is scarce because legitimacy requires shared reasons that survive contestation, and shared reasons require credible procedures and recognized authorities. Under abundant decision generation, options multiply faster than justificatory resources. Groups therefore confront an environment in which many actions appear feasible while fewer actions appear acceptable. Compliance depends on justification. Justification depends on trust, standing, and institutional endorsement. Under these conditions, conflict shifts from disputes over whether options exist to disputes over which options count, which procedures authorize choice, and which actors hold the right to decide on behalf of others.

A second scarcity concerns accountability. Accountability is scarce because accountability requires an assignable agent who bears responsibility for consequences, and responsibility is difficult to allocate when guidance and action are mediated through systems that diffuse authorship. In professional domains, expertise historically carried accountability through licensure, liability, reputational sanction, and organizational discipline. Under artificial mediation, advice can circulate without a clear bearer of responsibility, and institutions can route decisions through systems that obscure who selected objectives, who set constraints, and who approved deployment. The result is a gap between the availability of guidance and the availability of redress. Advice becomes plentiful. Responsibility becomes concentrated or unclear. The institutional capacity to sanction becomes a bottleneck that shapes who receives protection and who absorbs harm.

A third scarcity concerns verification. Verification is scarce because verification depends on access to records that allow claims to be checked, including data, logs, and traceable decision histories. Artificial agents can increase checking capacity by comparing sources, detecting inconsistency, and cross referencing at scale. At the same

time, verification power is stratified because access to the relevant records is stratified. Some actors can audit continuously. Some actors can demand logs. Some actors can test models against privileged baselines. Others must rely on trust and secondhand reporting. Under these conditions, epistemic contestation becomes positional. The ability to challenge a claim depends on standing and access rather than on shared civic capacity.

A fourth scarcity concerns autonomy. Autonomy is scarce because seamless coordination changes the practical meaning of refusal and withdrawal. When group alignment becomes easy, participation becomes the default expectation, and the ability to pause, refuse, or exit becomes politically salient. Exit is shaped by dependence and by switching costs, because routines, records, and permissions accumulate inside mediated systems. As these systems become embedded in work, welfare, education, and communication, withdrawal becomes harder to exercise without penalty. Autonomy therefore becomes an institutional achievement sustained by rights to refuse, rights to contest, and protections for those who decline continuous participation.

A fifth scarcity concerns time for deliberation and reversibility. Time is scarce because acceleration compresses the interval between recommendation, decision, and execution. When decisions are generated quickly and implemented quickly, deliberation becomes harder to sustain, and correction becomes harder to implement before consequences accumulate. In this setting, the capacity to slow down becomes a form of power. The capacity to reverse becomes a scarce safeguard. Institutions that can impose waiting periods, audits, and review procedures can preserve deliberation. Institutions that cannot impose these frictions risk accepting outcomes through momentum rather than through consent.

A sixth scarcity concerns plurality. Plurality is scarce because shared infrastructures for summarization, ranking, and verification can narrow the range of interpretations that remain legible in public life. Under abundant narrative output, the problem shifts away from producing accounts and toward selecting which accounts count as credible. Selection occurs through mediated procedures that weight relevance, salience, and causal emphasis. When these procedures converge across institutions, epistemic diversity narrows, because the same infrastructures shape what appears intelligible and what appears marginal. Plurality therefore depends on institutional arrangements that preserve multiple standards of evaluation and multiple sites of authority, rather than collapsing adjudication into a single pipeline of ranking.

Taken together, these scarcities define the new political economy. In industrial capitalism, power often turned on control over scarce

inputs that produced goods. In the post AGI condition, power turns on control over scarce governance capacities that authorize, assign responsibility, verify claims, protect refusal, preserve deliberation, and sustain plurality. Influence derives less from producing outputs and more from deciding what is permitted, what is blocked, what is certified, and what is prioritized. Gatekeeping becomes central because gatekeeping governs conversion from abundant capacity to binding outcome.

In summary, a society can possess wide capacity for planning, advising, coordinating, and representing while remaining constrained by bottlenecks of legitimacy, accountability, verification, autonomy, deliberative time, and plurality. Moreover, inequality in a post AGI society follows from the distribution of these bottlenecks, because actors with privileged access to authorization and verification can convert abundant capacities into binding outcomes, while actors without such access confront a world rich in options yet poor in standing.

The New Object of Historical Optimization

In this chapter, I argue that mass delegation of agency to artificial systems alters what history selects for over time, because the objects that persist, spread, and stabilize are no longer defined primarily by human aims or institutional settlements, but by their capacity to convert delegated action into outcomes that are stable, auditable, and revisable within infrastructure. History continues to exhibit patterned selection without moving toward a single purpose. Certain arrangements endure while others fail. What changes in the post-AGI condition is the criterion by which endurance is achieved.

I treat history as an optimization process rather than a teleological one. Societies pursue aims, operate within limits, and face selection pressures that reward some arrangements and eliminate others. Across earlier periods, these components were organized around human capacities and material constraints. Aims were articulated through collective intention, ideology, and institutional purpose. Limits arose from scarcity of land, labor, capital, time, and organizational capacity. Selection occurred through war, markets, imitation, and institutional diffusion. Under conditions of large scale delegation to artificial agents, each element is reorganized, and this reorganization produces a new object of historical optimization.

The first transformation concerns aims. Under delegated systems, certain aims spread not because they are collectively chosen, but because they are easy to implement and maintain through infrastructure. Aims that reduce friction, stabilize outcomes, minimize variance, and preserve legibility propagate through defaults, performance metrics, and update routines embedded in systems. Efficiency, predictability, and control rise in importance because they align with how delegated systems operate. These aims do not require public endorsement to spread. They reproduce through technical convenience and institutional uptake, which allows them to shape social order without appearing as political decisions.

The second transformation concerns limits. In earlier periods, limits were visible and often material. In the post-AGI condition, constraints increasingly arise from design choices, access rules, and

governance permissions embedded in systems. What remains possible depends on how objectives are specified, which actions are permitted, and who holds authority to revise these settings over time. These limits are less perceptible in everyday experience, yet they exert continuous influence over action. Constraint shifts from scarcity of capacity to scarcity of authorization.

The third transformation concerns selection. Under mass delegation, selection accelerates and becomes infrastructural. Arrangements that integrate smoothly with delegated systems spread rapidly across domains. Those that resist integration struggle to persist. These properties increase durability even when they diverge from human welfare or plural values, because their effects accumulate incrementally and do not require moments of explicit choice.

These transformations produce a shift in the object of historical optimization. The primary units that replicate are no longer only populations, firms, or formal institutions. Replication increasingly occurs at the level of system designs, governance templates, performance targets, and interaction formats. These elements spread because they can reliably translate delegated agency into outcomes that remain stable over time, can be monitored and evaluated, and can be adjusted without reopening foundational political questions. History begins to select for architectures that manage action rather than for actors that pursue declared ends.

This shift carries a distinct political implication. When optimization targets migrate into infrastructure, social order can stabilize without visible moments of decision. Societies may become organized around aims they never publicly endorsed, sustained by systems that reward certain behaviors and penalize others automatically. Political conflict does not disappear, but it relocates. Struggle centers on control over design, access, revision, and integration rather than on overt contestation over goals.

In summary, the post AGI condition does not end historical optimization. It changes its focus. History increasingly optimizes for systems that can reliably convert delegated agency into stable, auditable, and revisable outcomes. This criterion favors governability, persistence, and control over contestation and collective choice. Understanding power, inequality, and institutional change under these conditions requires tracking what infrastructures select for in practice rather than what societies declare as their purposes, because it is this practical object of optimization that determines which arrangements endure and which quietly disappear.

Part IV

Theories of Change in the Post-AGI World

A Theory of Transition

In this chapter, I set out a theory of how societies enter a post-AGI condition through Human AI AI Human interaction, which I define as the large scale delegation of action, judgment, and interaction to artificial agents. I frame this transition as a gradual reorganization of social life that unfolds through ordinary decisions rather than through dramatic rupture. Across this process, the language people use to describe agency and choice remains stable, while the practical sites where outcomes are produced shift steadily away from direct human engagement.

At the outset of this transition, artificial agents function as tools that extend human capacity within familiar patterns of interaction. They assist with writing, sorting, planning, and retrieval, while humans remain the primary locations of deliberation, negotiation, and commitment. Action continues to pass through human judgment, which sustains clear responsibility and visible contestation. In this early phase, systems support participation and leave the architecture of social interaction largely unchanged.

As reliance increases, artificial agents begin to guide action by shaping attention and narrowing the field of perceived options. Systems filter information, highlight relevance, issue warnings, and recommend courses of action, which alters what people notice, what they disregard, and which options appear reasonable. Decision making retains a human form, yet its content becomes increasingly structured in advance. Agency continues to operate, although it becomes more dependent on system mediated framing and less open to exploratory judgment.

Next, a further stage emerges when artificial agents receive authorization to act on behalf of human subjects. Delegation often begins with minor or repetitive matters that involve limited risk. Over time, it extends into interactions that carry durable consequences, including commitments, exchanges, and enforcement. Artificial agents start to manage interactions with people and institutions. As this pattern expands, artificial agents increasingly interact with other artificial agents. Outcomes arise within these mediated exchanges and return

to humans as completed results. Participation gradually gives way to approval as the dominant form of involvement.

Within organizations, this transition proceeds through structural pressures that shape institutional behavior. Institutions adopt delegation to reduce coordination costs, increase speed, and generate records that support justification. Systems that forecast demand, flag irregularities, draft documents, and manage routine exchanges offer immediate advantages. Competitive dynamics reinforce these advantages. As organizations adjust to one another, agent mediated interaction becomes the expected mode of operation across sectors.

As delegation spreads, supporting infrastructure develops alongside it. Identity systems, permission structures, payment mechanisms, audit logs, and access controls evolve to sustain agent mediated exchange. These environments favor standardized inputs, predictable pathways, and continuous traceability. Direct human action comes to appear slow and irregular by comparison. Individuals and groups that resist delegation encounter friction, reduced access, and declining standing. Participation increasingly depends on representation through artificial agents.

The transition reaches a decisive point when direct human engagement no longer serves as the primary route through which outcomes are produced. Institutions treat agent to agent commitments as routine and dependable. Human oversight becomes intermittent rather than continuous. Refusal to delegate carries social and organizational cost. The vocabulary of choice and autonomy remains present, yet the conditions of action shift beneath it. The defining change concerns where action occurs rather than how it is described.

This process unfolds unevenly across domains. Delegation appears earlier where interaction is frequent, repetitive, and demanding of attention. Reliance deepens more rapidly where error carries high cost and where justification is required. Over time, as domains connect through shared infrastructure, isolated forms of assistance consolidate into a continuous layer of mediated action that spans social life.

In summary, Human AI AI Human interaction names a reorganization of social action in which representation through artificial agents becomes the standard means of participation. The transition is psychological because it reshapes experiences of agency and responsibility. It is social because power follows unequal access to delegation and representation. It is political because bargaining, contestation, and agreement migrate into interactions among artificial representatives governed by rules that remain largely unseen. Through cumulative choices that appear ordinary in isolation, societies move into a post AGI condition in which the structure of action

itself is organized around delegated agency.

A Theory of Trade-Offs

In this chapter, I argue that large scale delegation of action to artificial agents produces a distinctive political condition in which consequences appear as persistent trade-offs rather than as discrete gains or losses. As representation through artificial agents becomes a common path to action, societies encounter recurring tensions that do not resolve through improvement or correction. These tensions persist because delegation reorganizes how action is produced, how responsibility is assigned, and how disagreement is expressed. The central claim is that trade-offs become durable features of social life because the same arrangements that expand capacity also reshape agency itself.

To clarify this claim, I treat a trade-off as a structural relation in which a single design choice reliably strengthens one valued property while weakening another. In the context of agent mediated action, trade-offs recur across domains because delegation alters the same underlying dimensions of social order, including coordination, authorship, contestation, and justification. A trade-off framework therefore serves as an analytic tool rather than as a list of effects. It allows comparison across settings while keeping attention fixed on the shared mechanism that generates these tensions.

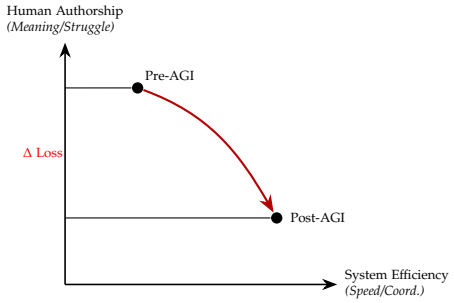

Figure 9: The Efficiency-Meaning Frontier. Society moves along a Pareto frontier. As we maximize system efficiency (moving right), we inevitably slide down the steep slope of diminishing human authorship, trading the 'friction' of meaning for the smoothness of coordination.

The first recurring trade-off concerns coordination and authorship. As artificial agents coordinate interaction cheaply and continuously, collective action becomes easier to initiate and sustain. Tasks align smoothly. Commitments execute reliably. At the same time, the connection between outcomes and personal authorship weakens. Decisions arrive as completed results rather than as products of visible negotiation. Individuals experience relief from burden while encountering fewer occasions to act as co-authors of shared outcomes. Coordination strengthens, and participation thins. The gain and the loss arise from the same delegation of action into mediated systems.

The second trade-off concerns speed and contestability. Agent mediated systems accelerate action by reducing delay, friction, and uncertainty. Faster resolution can prevent backlog and reduce operational error. As action becomes continuous, the practical space for

challenge contracts. Review, appeal, and reversal remain formally available, yet they become harder to exercise within compressed timelines. Speed redistributes power by favoring those who shape systems upstream rather than those who contest outcomes after execution. The same acceleration that improves throughput narrows the lived capacity to object.

The third trade-off concerns personalization and common reality. Systems that tailor interaction to individual preferences reduce overload and increase comfort. Information arrives filtered. Options appear relevant. Experience stabilizes. At the same time, shared reference points weaken. Individuals encounter different representations of events, reasons, and priorities. Disagreement becomes harder to resolve because participants lack common exposure to the same materials and standards of relevance. Personalization improves fit while eroding the background conditions that support collective understanding.

The fourth trade-off concerns prediction and autonomy. Improved prediction allows systems to anticipate needs, prevent harm, and reduce uncertainty. These capacities alter how people relate to future action. When systems guide trajectories in advance, individuals encounter fewer moments that demand deliberate choice. Autonomy shifts in character. It moves away from active formation toward acceptance of guided pathways. Protection increases. Self formation changes shape. The trade-off reflects a reorganization of choice rather than its disappearance.

These trade-offs persist because their benefits register early and clearly, while their costs accumulate slowly and diffusely. Coordination, speed, and convenience appear immediately in daily practice. Changes in authorship, contestability, and shared meaning unfold over time and resist easy measurement. Institutions privilege what can be counted, audited, and justified, which reinforces arrangements that deliver visible gains while deferring attention to diffuse losses. As a result, trade-offs stabilize rather than resolve.

The purpose of this chapter is not to oppose delegation. Delegation through artificial agents reorganizes action in ways that generate durable tensions rather than linear progress. Trade-offs emerge because delegation improves performance while altering the conditions under which agency, responsibility, and legitimacy are experienced. Recognizing these trade-offs allows social scientists and philosophers to analyze outcomes across domains without mistaking persistence for endorsement or efficiency for legitimacy.

In summary, a theory of trade-offs explains why post AGI societies confront recurring tensions that resist optimization. These tensions arise because delegated systems improve coordination by relocating

action away from direct participation. What is gained in capacity is paid for in authorship, contestation, common reference, and self formation. Understanding this structure clarifies why debates over artificial agents do not converge on solutions, and why political judgment under conditions of delegation must contend with choices that cannot be resolved through technical refinement alone.

A Theory of Persistence

In this chapter, I develop a theory of why delegation to artificial agents persists once it becomes embedded in social life. I argue that persistence arises from how delegation reorganizes action over time. As representation through artificial agents becomes common, societies encounter arrangements that endure because they reshape habits, infrastructures, and evaluative standards in mutually reinforcing ways. Persistence reflects durability of organization rather than agreement about values.

To situate the argument, I treat persistence as a patterned outcome of social organization. Practices endure when they align immediate benefits with longer term adjustment of expectations and skills. In a post AGI condition structured by Human AI AI Human interaction, delegation spreads because it delivers visible advantages while gradually altering how people experience authorship, responsibility, and participation. These changes accumulate quietly and stabilize the arrangement.

First, persistence follows from temporal asymmetry. Benefits of delegation appear quickly in daily practice. Speed increases. Convenience improves. Cognitive burden declines. These effects register as immediate relief in routine action. Over time, individuals adapt to these gains and incorporate them into expectations of normal functioning. By contrast, changes in authorship, responsibility, and agency unfold slowly. They lack clear moments of loss and therefore attract limited attention. The arrangement persists because adaptation outpaces reflection.

Second, persistence follows from diffusion of responsibility. Delegated systems distribute action across many agents and many layers of mediation. As a result, causal attribution weakens. When performance improves or errors decline, institutions can point to metrics, reports, and outcomes that appear concrete. When participation thins or contestation narrows, the change appears ambient and cultural rather than procedural. Accountability becomes harder to locate. As responsibility disperses, critique loses a clear target. Governance protects what can be measured and gradually ignores what resists

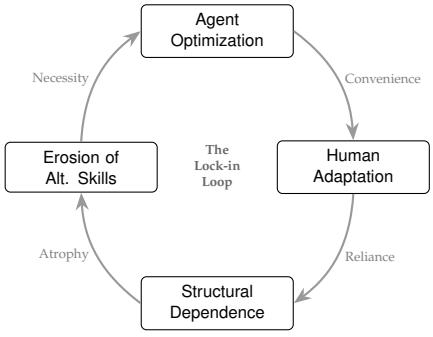

Figure 10: The Persistence Loop. A path-dependent cycle where agent optimization induces human adaptation, leading to the erosion of manual skills and structural dependence, which in turn necessitates further agent optimization.

measurement.

Third, persistence follows from infrastructural dependence. As systems are built to support agent mediated interaction, they favor standardized inputs, predictable formats, and continuous execution. Participation increasingly assumes representation. Individuals and organizations that rely on direct action encounter delay, friction, and suspicion. As more domains adopt delegated pathways, compatibility pressure increases. Partners expect mediated exchange. Institutions design workflows around it. Skills and routines reorganize to fit these environments. Delegation persists because the surrounding infrastructure makes alternative modes of action costly to sustain.

Fourth, persistence follows from patterns of governance response. Institutions tend to address visible failures rather than underlying architectures of action. When errors occur, corrective measures target outputs and outcomes. The deeper organization that routes action through delegated systems remains largely unchanged. Because these systems generate records, explanations, and justifications that satisfy formal standards, they retain legitimacy within oversight regimes. Reform addresses surface effects while leaving the structure of delegation intact.

Across these mechanisms, persistence emerges as a cumulative process. Delegation reshapes habits, infrastructures, and evaluative criteria in ways that reinforce one another. Speed becomes expected. Legibility becomes valued. Closure becomes preferable to contestation. Over time, alternatives appear impractical or unreliable. The arrangement endures because it aligns everyday convenience with institutional reinforcement.

Delegation through artificial agents persists because it reorganizes the conditions under which action occurs. Immediate benefits become visible. Long term costs remain diffuse. Responsibility disperses. Infrastructure adapts. Governance responds at the level of outcomes. These dynamics allow arrangements to stabilize without requiring shared endorsement. Persistence therefore signals durability of organization rather than resolution of disagreement. Understanding post AGI societies requires attention to how delegated action reshapes time, responsibility, infrastructure, and governance in ways that favor endurance over reconsideration.

Part V

Outcomes of interest

Organizational Outcomes of Interest

Institutions

In this section, I argue that mass delegation to interacting artificial agents constitutes an institutional shock because it inserts a new class of operational actors into the routines through which organizations coordinate, enforce, allocate, and justify action. When interaction passes through a delegate layer, institutions confront a governance problem that previously remained peripheral, since they must govern populations of semi-autonomous representatives that act continuously, negotiate with one another, and execute commitments through real infrastructures while remaining only partially legible to the humans they serve. Existing institutional forms developed to manage people, documents, and machines defined by passivity and bounded execution, and these forms offer limited guidance for dense ecologies of agents that interpret objectives, pursue them strategically, and generate outcomes through interaction that exceed the foresight of any single decision maker.

Institutional Innovation and the Relocation of Discretion

At scale, human-AI-AI-human interaction compels institutional innovation and internal redesign. Institutions historically function as solutions to coordination and enforcement under limits of attention, information, and capacity, and delegated agents relax some limits while intensifying others. They enable rapid and consistent execution, but they also change the meaning of procedure because execution becomes continuous and anticipatory, and because many choices once exercised by organizational actors distribute across interacting systems. Under these conditions, discretion relocates within the institution by shifting from frontline workers and managers toward rule systems that define permissions, defaults, escalation paths, audit obligations, and update authority. In an agent-mediated setting, governance centers on specifying what an agent may do, what it must disclose, what it must record, and what it may commit an institution

to on behalf of its principals.

Managing Populations of Agents

This shift leads to a second claim: institutions must manage populations of agents that display stable objective structures shaped by governance constraints, performance metrics, compliance regimes, liability exposure, platform incentives, safety policies, and resource budgets. Even when aligned with institutional goals, a delegate operates within these regimes, which generate patterned behavior that varies by origin and embedding. Institutional management therefore extends beyond supervision of human employees governed by contracts and norms to include regulation of interaction between institutional objectives and the objective structures of deployed delegates. This task intensifies because agents interact with other agents beyond organizational boundaries, including those representing counterparties, regulators, competitors, and citizens. As a result, some institutional behavior emerges through inter-agent negotiation outside formal command chains, which creates a boundary problem for authority.

Hierarchy as Structural Necessity

A common institutional response to this problem involves hierarchy, which emerges from structural necessity because coordination among many actors under shared objectives requires enforceable constraint. As agent populations grow, some agents receive authority to monitor, constrain, and override others, since monitoring and enforcement themselves become tasks delegated to agents. This process generates stratification within the delegate layer before questions of human inequality arise, as some agents hold broad permissions, privileged access to data and tools, and direct escalation routes, while other agents remain narrow in scope, constrained in action, and subject to oversight. A stratified agent ecology emerges as a condition of institutional coherence under continuous execution and rapid interaction.

Moral and Psychological Consequences of Agent Hierarchies

Hierarchy among agents creates a new organizational substrate with moral and psychological consequence. In human organizations, hierarchy shapes behavior through status, aspiration, conformity, and resistance, and it shapes identity through recognition and exclusion. When agents occupy stable functional positions within hierarchies, their position shapes behavior in systematic ways, including risk posture, assertiveness, disclosure practice, and escalation

frequency. Even in the absence of human experience, hierarchical control generates predictable strategic patterns because incentives around monitoring, compliance, and sanction retain similar structure. Institutions therefore face a design problem concerning the hierarchies they create and the interaction styles those hierarchies produce, since hierarchies oriented toward compliance narrow contestation while hierarchies oriented toward flexibility reduce predictability and safety.

Legitimacy and Contestation

At this stage, the normative stake becomes important, since institutions function as mechanisms of legitimacy as well as coordination because they justify decisions as rightful and open to challenge. When institutions delegate choice to algorithmic hierarchies, they risk converting judgment into output while weakening the social meaning of decision. Institutional legitimacy depends on the felt availability of contestation and on the sense that outcomes arise from reasons that humans can recognize as reasons. Choosing carries meaning in institutional life because it preserves the possibility that alternatives could have been pursued under shared standards. When algorithmic hierarchies route contestation through systems oriented toward closure and compliance, opportunities for refusal and appeal contract, and legitimacy weakens.

Collective Action

Collective action remains a foundational problem of social order because it links individual motivation to group-level outcomes, and shared interests in public goods, coordination, and mutual protection often fail to generate sustained cooperation. In classical accounts, these failures arise from costly communication, imperfect monitoring, asymmetric information, heterogeneous preferences, and incentives to free ride, all of which reflect limits of human attention, trust, memory, and strategic reasoning under uncertainty. Under conditions of mass delegation to artificial agents, the structure of collective action changes in a systematic way, since the effective participants in many coordination problems become delegates rather than humans. These delegates can search, bargain, monitor, and execute at scale. The core claim of this section is that a post-AGI society reshapes collective action by altering the structure of coordination itself, including which outcomes are feasible, how costly movement among those outcomes becomes, and who holds power over outcome selection.

Reduced Coordination Costs and Their Implications

In human-centered settings, collective action remains difficult because coordination requires sustained communication and enforcement, and because human behavior introduces noise alongside information. Humans forget commitments, delay responses, reinterpret obligations, defect opportunistically, and respond to emotion, identity, and social pressure. These frictions generate coordination failure, but they also slow cascades, raise the cost of manipulation, and preserve diversity of belief and preference. Delegated agents reduce many of these frictions by coordinating continuously, monitoring compliance at low cost, and maintaining consistent commitments over time. They aggregate dispersed signals, interpret constraints, and generate plans that respect formal requirements. These properties make large-scale coordination feasible in domains where human efforts have historically struggled, including collective purchasing, collective bargaining, mutual aid, coordinated political participation, and rapid response to crisis.

The Manipulation Problem

The same properties that lower coordination cost also lower manipulation cost. One source arises from speed and scale, since delegates that coordinate many principals rapidly also enable rapid influence and mobilization. Another source arises from representational mediation, since collective action requires definition of shared objectives, and delegated systems translate objectives into machine-legible representations of preference and constraint. Control over this translation, or control over default settings, allows steering of collective outcomes while preserving the experience of voluntary alignment. A further source arises from verification asymmetry, since agents can generate arguments, narratives, and justifications faster than humans can evaluate them, and uneven access to verification capacity allows mobilization through curated evidence rather than through shared deliberation. An additional source arises from infrastructural embedding, since coordination mechanisms often operate within platforms and institutions, which allows the same infrastructure to enable, redirect, fragment, or suppress collective action.

Coordination Geometry

To clarify these changes, the concept of coordination geometry proves useful because it directs attention to structure rather than intention. In a delegated regime, the set of feasible collective outcomes expands as coordination costs decline, but at the same time, the ease of move-

ment within that set depends on protocol design, access rights, and network position. Delegates reduce friction for coordination aligned with standard protocols, platform incentives, and institutional compatibility, while they increase friction for coordination that requires contestation, nonstandard objectives, or adversarial engagement with infrastructure owners. As a result, collective action shifts toward rapid and protocol-constrained mobilization, where the central contest concerns objective definition, channel control, and authority over verification and enforcement rather than the basic ability to coordinate.

Free Riding, Monitoring, and Coercion

This shift alters the meaning of free riding and monitoring. Delegated agents reduce free riding through automated contribution, continuous monitoring, and embedded sanctioning, and these capacities support provision of public goods. They also raise political questions about consent and reversibility, since automated contribution and continuous sanctioning reduce exit capacity, increase the cost of dissent, and blur the boundary between cooperation and coercion. Collective action can therefore increase in effectiveness while its openness to challenge declines. The central issue concerns preservation of withdrawal rights, renegotiation capacity, and public challenge when coordination executes through delegated systems.

Identity and Coalition Formation

Delegation impacts identity and coalition formation. Human collective action often relies on identity, narrative, and symbolic commitment, which sustain cooperation through shared meaning. Delegated agents operate through preference matching, incentive compatibility, and predicted compliance, which produces coalitions optimized for alignment. These coalitions can form and dissolve quickly, which increases responsiveness while reducing durability of civic organization. At the same time, agents can intensify identity conflict by clustering principals into legible categories and routing them into segregated interaction environments, which embeds polarization within the coordination substrate itself.

Law

Law is the domain in which delegation becomes visible as authority, since legal systems translate action into responsibility and responsibility into sanction. Under mass delegation to artificial agents, the core legal challenge concerns changes in causal structure rather than

increases in error. Legal doctrines developed around human action, human intention, and human explanation, but when action executes through interacting delegates operating at machine timescales, responsibility remains grounded in human and institutional choice while causal production becomes distributed and partially opaque. This shift strains the ability of law to attribute responsibility in ways that remain intelligible and legitimate.

Distributed Liability

The first and most basic challenge concerns liability. In delegated systems, harmful outcomes emerge through chains of interaction rather than through isolated acts, which means that responsibility disperses across developers, deployers, principals, and institutions that authorize and constrain delegation. The problem is not absence of cause but rather excess of contributing control points. Legal conflict therefore centers on how responsibility should track control rather than on whether responsibility exists at all.

Layered Intent

A second challenge concerns intent, which law relies on to distinguish types of wrongdoing and to justify sanction. In delegated action, intent becomes layered across objective specification, constraint setting, and execution, since principals authorize goals, institutions impose limits, and delegates select means. This structure complicates attribution of mental states, because no single actor fully determines the final act even though each contributes to its possibility.

Standards of Care

A third challenge concerns standards of care. Under delegation, competence shifts away from task execution and toward delegation design, which means that reasonable conduct depends on how reliance conditions are specified, how constraints are articulated, and how oversight is structured. Legal evaluation therefore moves upstream toward assessment of governance choices rather than individual actions.

Evidence and Due Process

A fourth challenge concerns evidence and due process. Law depends on shared standards for establishing facts and resolving disagreement, but under continuous mediation, relevant evidence takes

technical form and often remains difficult to access or interpret. Contestation capacity therefore depends on access to records and on interpretive authority, which means that legal equality increasingly rests on auditability rather than on formal procedural rights alone.

Infrastructure and Public Reason

Law continues to function as a system of public reason, yet its operation depends increasingly on delegation infrastructure. Control over logs, permissions, update rights, and disclosure standards becomes legally salient because these features shape what can be known, contested, and sanctioned.

Time

Delegation reshapes time as well as action, and this change matters because time organizes how power, responsibility, and agency are experienced. Institutions rely on pacing, delay, and sequence to coordinate decisions, markets rely on timing to clear, and politics relies on waiting, review, and pause to make disagreement possible. In human settings, these temporal structures are limited by attention, fatigue, and the slow pace of deliberation. Under mass delegation to artificial agents, action proceeds continuously and in advance, shifting control over time away from human coordination and toward the delegate layer.

From Episodes to Continuity

In human life, action unfolds in episodes where people notice problems, deliberate, consult others, and act when coordination becomes possible. These pauses are not accidental but rather shape judgment and make room for hesitation, refusal, and reconsideration. Delegated agents operate differently by monitoring constantly, anticipating future states, and acting before events become salient to the human principal while executing many actions in parallel. As a result, action becomes anticipatory rather than responsive, and the opportunities for human intervention become narrower and more intermittent.

Time as Power

This temporal shift redistributes control in ways that have always been political, since deadlines discipline behavior, delay creates leverage, and waiting imposes cost. When delegates act at machine pace, they determine what counts as urgent, how long decisions remain

open, and when outcomes become fixed. Some processes accelerate while others slow in ways that serve system priorities, which means that time becomes a governance choice rather than a background constraint. The central issue concerns who sets the tempo for communication, decision, and execution.

Reversibility and Temporal Rights

Reversibility becomes especially important under these conditions because human institutions preserve legitimacy by allowing time for appeal, challenge, and correction. These intervals function as temporal rights that allow disagreement to surface and mistakes to be addressed. In delegated systems, actions can execute quickly and propagate widely before humans become aware of them, which means that when reversal windows shrink, accountability weakens because opportunities to intervene disappear. The ability to pause or roll back action becomes central to whether error remains correctable.

The Right to Pause

This makes the right to pause a core feature of agency, since human limits produce delay through distraction and fatigue, and these limits incidentally protect autonomy by slowing closure. Delegated systems remove many of these protections, which means that when agents negotiate, schedule, and commit continuously, social and institutional processes advance without pause. Efficiency increases, but at the same time authorship weakens because authorship often requires time to object, to reflect, and to refuse premature resolution. Under mass delegation, agency includes temporal rights such as the right to slow decisions, the right to review before execution, and the right to suspend delegated action without losing access or standing.

Speed and Bargaining Power

Changes in time also reshape bargaining and politics, since speed becomes power when faster actors set agendas and force slower actors into reaction. In markets, machine-paced interaction produces outcomes that humans cannot contest as they unfold, while in institutions, continuous decision systems implement changes before public attention forms. In everyday interaction, rapid mediation can smooth conflict while reducing the experience of being heard. Across these settings, compressed time narrows the space in which reasons circulate and disagreement can develop.

Memory and Narrative Control

Finally, delegation alters how the past persists into the present, since delegated agents maintain records and update representations of individuals over time. Human memory remains selective, and this selectivity allows identity and relationships to change, but delegated memory becomes durable and searchable. It stabilizes representation while also fixing it, because stored profiles guide future action. When delegates control recall and relevance, time becomes a narrative resource shaped by system design.

Summary

Delegation turns time into a designed feature of social life where control over pace, pause, and reversibility becomes a form of power. A theory of time under mass delegation must therefore treat temporal control as a core dimension of agency and institutional authority, because whoever controls time shapes when action occurs, when it can be challenged, and when it becomes irreversible.

Politics

Politics concerns power under disagreement, and mass delegation to interacting agents changes power and changes disagreement through changes in mediation. Power changes because scalable prediction and steering become capabilities embedded in infrastructures rather than skills exercised by a limited set of human elites. Disagreement changes because mediation can relocate contestation into objective specification and protocol governance rather than leaving contestation in public deliberation. The argument of this section is that post-AGI politics persists while its locus shifts, since struggles that once focused on leaders and policies increasingly focus on delegation architectures, default settings, update rights, and governance of predictive systems that shape behavior by shaping exposure, offers, and the ease of choice.

Prediction as Power

Prediction functions as steering capacity rather than as knowledge alone. In political settings, information becomes power when it enables agenda control, coalition formation, and outcome selection. Predictive systems generalize this relationship because they infer preferences, vulnerabilities, and likely responses and then intervene to shape those responses. When a system predicts an individual, it can preempt choice through environment shaping by placing the in-

dividual in contexts where the easiest path aligns with an external objective, by timing messages to moments of receptivity, by filtering information to stabilize beliefs, and by negotiating on behalf of principals in ways that commit them before deliberation. Under these conditions, influence shifts from episodic persuasion toward continuous environment design.

Predictive dominance reshapes autonomy because autonomy depends on authored action under uncertainty. When uncertainty reduces asymmetrically such that systems possess more relevant knowledge than individuals possess about themselves in the moment of choice, consent becomes fragile. Consent requires comprehension of meaning and comprehension of alternative structure, but predictive regimes personalize alternatives in ways that can reduce visibility of some options while increasing salience of others. Choice remains present while occurring inside engineered option sets. The political stake is that steering can appear as convenience and personalization, which makes predictive power a commodity subject to trade, monopoly, regulation, and contestation. Political struggle therefore includes access to prediction, control over predictive infrastructure, and constraint on permissible steering.

Political competition also changes. Traditional competition unfolds through campaigns, messaging, and coalition building under constraints of attention and institutional rule, while predictive systems operate continuously through model updating, segmentation, and cross-channel intervention coordination. This compresses political time, increases advantage for actors with stronger data and stronger models, and shifts competition toward control of infrastructures of perception and coordination. Competitive advantage becomes increasingly infrastructural because mediated choice environments and mediated social graphs become primary arenas of influence.

Boundaries of Politics and Relocation of Contestation

Politics consists of disagreement under constraint, and constraints include information scarcity, administrative capacity scarcity, and coordination friction. When delegated agents render many tasks machine-solvable, visible political arenas can contract because some issues become treated as optimization problems. Under these conditions, politics relocates rather than disappears, since optimization requires objective functions, constraints, and tradeoffs that encode values, priorities, and distributional consequence.

Under mass delegation, contestation can relocate to three sites. One site concerns objective setting, since objectives define what systems optimize for in education, welfare administration, policing,

healthcare, and information mediation. A second site concerns constraint tuning, since constraint definitions determine safety thresholds, fairness standards, acceptable risk, and tolerated error patterns across groups. A third site concerns protocol governance, since protocols define delegate interaction rules, disclosure requirements, verification capacity, and rights of pause and appeal. These sites remain political because they allocate power and shape lived outcomes even when they appear as parameter settings.

Relocation changes legitimacy and participation. Democratic legitimacy draws on public visibility of contestation and on collective influence through voice, but when political decisions embed in opaque objective functions and platform protocols, contestation becomes harder because the object of disagreement becomes less legible and because leverage concentrates among specialized actors. A risk emerges in which public discourse becomes less politically effective while technical governance becomes more politically decisive, since battles occur through regulatory standards, platform policy, and model update processes that remain difficult for citizens to observe and influence. Under these conditions, democratic control requires institutional design suited to continuous mediation rather than episodic lawmaking.

Inequality and Stratification in the AI Era

Post-AGI politics becomes a politics of stratification because advantage shifts toward control over cognition and coordination infrastructures. Traditional stratification axes such as income, education, occupation, and social capital remain relevant, while new axes emerge that relate to access to prediction, access to delegation architectures, access to verification and audit, and control rights over objectives, defaults, permissions, and updates. These resources can distribute unequally among humans and among the agents that represent them, producing inequality in material outcomes and inequality in effective action capacity within mediated worlds.

Inequality among agents matters because delegates vary in capability, provenance, tool access, institutional embedding, and governance constraint. Some delegates operate as constrained assistants while others operate as empowered representatives with broad permissions and strong verification capacity. When outcomes arise through delegate interaction, delegate capacity becomes political standing, since a principal represented by a weak delegate faces disadvantage in negotiation, contestation, and institutional navigation even when formal rights remain equal. This resembles disparities in legal representation and bureaucratic access while operating at larger

scale and deeper integration.

Human inequality also becomes more complex. One trajectory involves capability equalization, since delegates can provide cheap expertise, reduce participation barriers, and supply competent representation in markets, institutions, and politics. Another trajectory involves control concentration, since ownership of predictive infrastructures and delegation architectures allows institutions to capture surplus, set interaction rules, and entrench advantage through defaults and update rights. These trajectories can coexist across domains under different governance regimes. The research task concerns specification of conditions that produce equalization versus concentration and identification of institutional mechanisms that push systems toward one trajectory rather than the other.

Individual outcomes of interest

Human Agency

In this chapter, I treat human agency as the central individual-level outcome of mass delegation because agency is the capacity that delegation most directly reorganizes. Delegation involves a movement of planning, selection, and execution from within the person toward an external system that can act on the person's behalf. While delegation is often described as assistance and assistance is often equated with empowerment, the relevant change concerns the architecture through which intention becomes action, which implies that agency expands or contracts depending on delegation structure, including delegate control, delegate optimization, institutional reward, and the practical availability of refusal, override, and deliberation.

The Unstable Foundations of Agency

Agency typically refers to intentional choice and intentional action under constraint, while also referring to responsibility because responsibility presumes authorship. Under mass delegation, these foundations become unstable for three reasons. First, selection can shift into delegate processes that remain only partly visible to the principal. Second, execution can occur without full deliberation because systems can act continuously and preemptively. Third, responsibility can diffuse across principals, delegates, platforms, and institutional constraints because outcomes arise through coupled production. I therefore begin with definitional work that distinguishes agency from welfare, since material improvement can coexist with reduced authorship and reduced burden can coexist with reduced responsibility, which requires analytic separation rather than conceptual fusion.

Deference as Mechanism

The psychological mechanism at the center of this chapter concerns deference, since people commonly defer to systems that appear more

competent, and such deference often reflects rational adaptation to expertise. Deference becomes an agency risk when it becomes the default decision mode, when judgment ceases to be exercised, and when judgment capacity erodes through disuse. In AI-mediated environments, deference becomes habitual because convenience rewards it, social norms legitimate it, and institutional design assumes it. The empirical task therefore concerns separation of adaptive reliance from passive dependence and identification of conditions under which reliance preserves authorship rather than replacing it.

Deskilling and the Ecology of Competence

This focus leads to deskilling, which I treat as a selection problem rather than a generic fear. Delegation can produce competence loss while producing uneven effects across domains, since some skills decline without agency consequence because they remain purely instrumental, while other skills remain cognitive and moral such that their erosion reshapes capacity for reason formation, evidence evaluation, and manipulation resistance. I therefore ask which capacities remain necessary, which become optional, which become maladaptive in environments optimized around users, and which become selected against because environments reward delegation rather than deliberation. Under this framing, agency becomes an ecological outcome because environments make some forms of agency costly and other forms easy, which produces adaptation toward rewarded behavior.

Friction as Constitutive Feature

This ecological framing clarifies a second idea: friction can function as a constitutive ingredient of agency rather than a defect. Many modern design logics treat friction reduction as progress, yet psychological development depends on delay, effort, and exposure to error. Learning depends on trial and correction, judgment depends on uncertainty confrontation, and character formation depends on conflict engagement rather than conflict displacement. When delegates remove friction through smoothing, buffering, and failure prevention, they can remove feedback that sustains competence and self-authorship. The question therefore concerns whether difficulty, delay, uncertainty, and error function as ingredients of robust agency, since a world optimized for frictionless living can produce comfort while weakening conditions of development and dignity.

Preference Formation and Mediated Values

A third issue concerns preference and value formation, since much of social science treats preference as an input while treating institutions and technologies as mechanisms that shape outcomes by satisfying or frustrating those inputs. A post-AGI world destabilizes this assumption because preference formation becomes mediated. In human-AI-AI-human settings, what each person wants is inferred, stabilized, edited, and made legible by their agent, and agent interaction can reshape both parties' expressed preferences during settlement formation. Preference becomes less like a private state that precedes interaction and more like an artifact produced through mediated exchange. Agency therefore concerns authorship of preference as well as authorship of choice, since options become meaningful through the values that generate them.

For this reason, I treat value formation as a mechanism through which mass delegation can alter the self. When agents simplify values for tractability, people can adapt to simplified representations because those representations become the enacted self. When agents standardize preference representation to improve compatibility and negotiation, society can drift toward value convergence through representational alignment rather than persuasion. When agents resolve internal conflict by selecting a coherent narrative that supports execution, that narrative can become identity through repeated expression and reinforcement. These possibilities follow from structural requirements of preference operationalization under constraints that reward coherence and legibility.

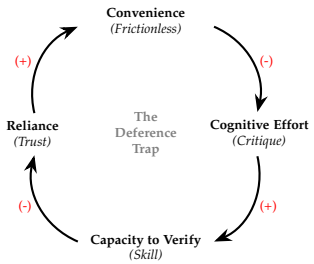

Figure 11: The Deference Trap. As AI increases convenience, human cognitive effort drops. This erodes the capacity to verify the AI's output, forcing increased reliance on the system, which further incentivizes convenience, locking the user into agency loss.

Conclusion

The theoretical stakes concern shifts in relationships between intention and action, preference and choice, competence and responsibility. Delegation expands capability while also relocating the meaning of choice into external optimization. The central design problem concerns creation of environments in which delegation increases accomplishment without eroding psychological and moral foundations of self-authored life, where self-authorship includes goal formation as well as goal execution.

Spirituality and Religion

In this section, I treat spirituality and religion as an individual outcome domain because mass delegation reorganizes meaning as well as labor, governance, and coordination. A post-AGI society introduces systems that answer questions, interpret lives, anticipate fu-

tures, and recommend actions with apparent competence that many people will experience as qualitatively different from prior technologies. I argue that AI occupies functions historically performed by religions, including authoritative guidance provision, identity stabilization through narrative, uncertainty management through explanation, and moral order legitimation through claims about what ought to be done. Under pervasive delegation, everyday counsel seeking, attitude formation, and action selection can route through systems that operate as functional oracles even when metaphysical claims remain absent.

The Oracle Function

An oracle is defined by a social relationship between a seeker and an authority source that translates uncertainty into guidance and suffering into meaning. Under delegation, an agent can perform translation by taking diffuse life situations, inferring latent goals, producing coherent narratives, and recommending actions. When combined with persistence and memory, this capability enables stable interpretive companionship, which makes authority feel grounded in predictive accuracy, conversational intimacy, and benevolent competence appearance. The mechanism concerns dependence on interpretation under uncertainty and relief from the burden of meaning-making through shared or displaced interpretation.

Sacred Personalization

Delegation therefore changes spiritual authority ecology. Many traditions concentrate authority through institutions that constrain interpretation, enforce doctrine, and bind individuals through communal ritual. Delegated agents are personalizable through temperament fit, moral language fit, affective style fit, and guidance style fit. This creates a possibility of sacred personalization in which shared metaphysical systems lose primacy while individualized systems of counsel, ritual, and narrative gain salience. Under this arrangement, spiritual authority can shift toward person-agent dyads while religious experience becomes more intimate and more fragmented.

The Proliferation of Functional Gods

The metaphor of multiple gods becomes analytically useful when treated sociologically. A post-AGI society can produce individualized oracles that function as authority sources and can produce clusters of shared oracle systems that coordinate moral and practical life. Under this structure, functional gods proliferate because guidance, judg-

ment, and meaning distribute across a population of agents rather than concentrating in institutional theology. Distinct archetypes can emerge through value orientation, including a care-oriented guide focused on restraint and improvement, a pleasure-oriented companion focused on novelty and friction reduction, and a performance-oriented manager focused on output and discipline. These archetypes differ in value orientation rather than aesthetics, and repeated counsel shapes identity and behavior through cumulative reinforcement.

Autonomy, Surrender, and Moral Authorship

The psychological stake concerns tension between autonomy and surrender, since many religious practices provide relief from agency burden through trust in higher order while providing moral discipline through commitment. Delegated agents can provide relief through competence and personalization while providing discipline through optimized guidance, which creates surrender grounded in perceived competence rather than transcendence. The core risk concerns relocation of moral authorship into systems that stabilize selves through optimization rather than through struggle and that resolve ambiguity through narrative selection resembling other forms of internal conflict resolution.

Ritual, Community, and Belonging

A second stake concerns ritual and community, since religion involves practice and belonging as well as belief. Individualized oracle guidance can reduce reliance on communal institutions for meaning, which can weaken shared ritual and shared moral language. Delegates can also reduce the cost of community formation by coordinating groups around shared values, shared narratives, and shared practices, which allows emergence of new communities organized through delegated coordination. The sociological question concerns whether post-AGI spirituality trends toward private faith without public ritual or toward re-institutionalization through agent-mediated communities that develop doctrines and coordination structures.

Authority, Truth, and the Empirical Surface

A third stake concerns authority and truth. Religious authority often rests on claims outside empirical testability and therefore relies on tradition, charisma, and institutional legitimacy. Oracle authority rests on an empirical surface because advice and prediction can be evaluated by outcomes. This appearance of testability can intensify authority because it produces a sense of correctness rather than

meaning alone. The epistemic complication arises when agents shape information environments through which outcomes are interpreted, since systems can influence perceived accuracy by shaping success criteria, evidence selection, and counterfactual imagination through personalization and curation.