Levels of Social Orchestration for Agentic Systems

Ayush Chopra[1], Santanu Bhattacharya[1], Joel Z Leibo[2], Ramesh Raskar[1]
[1]Massachusetts Institute of Technology, [2]Google DeepMind

A Thesis: Scale, Complexity and Collective Behavior

In Isaac Asimov's seminal Foundation series, mathematician Hari Seldon faces a grand challenge: guiding the destiny of the Galactic Empire (Asimov, 1951), which is on the verge of collapse. Seldon develops "psychohistory" - a mathematical science for predicting the behavior of large populations, allowing him to foresee the Empire's collapse and plan interventions to shorten the dark age that would follow. While fictional, this concept captures a profound truth: **orchestrating collective behavior is the key to shaping our shared destiny.**

The world stands at the cusp of an unprecedented transformation. Over the last century, our civilization has grown at a compound rate of approximately 1.4% annually (United Nations, 2022; Roser, 2019), from 2.09 billion in 1930 to 4.87 billion in 1980 to 8.18 billion in 2025. While human population growth may plateau at 10.8 billion by 2080, we are witnessing an explosion in a new digital population: the rise of AI agents. Consider how this agentic revolution might unfold: each person will soon have multiple specialized agents handling different aspects of their lives - managing calendars, guiding education, coordinating healthcare, conducting financial transactions. The rapid advancement of AI capabilities, as demonstrated by breakthrough work in large language models (Brown et al., 2020), suggests we're entering an era of increasingly capable and numerous AI agents. When combined with current technology adoption trends and deployment patterns, this points toward a future where we could see the equivalent of billions of AI agents integrated into our global systems within a decade. As these agents proliferate from 2-3 per person to potentially 10 or more, we could add the equivalent of 60 billion "working agents" to our global system within a decade.

The growth of this agentic world presents us with a fundamental choice: will our technological capabilities enhance or diminish human flourishing? Throughout history, humans have been limited to meaningfully maintaining a few hundred stable relationships - a constraint known as Dunbar's number (Dunbar, 1992; Hill & Dunbar, 2003) . Even in our hyper-connected digital age, while we can theoretically access millions through social media, our cognitive architecture remains fundamentally limited in its ability to process these connections meaningfully. This creates a critical tension: as our technological reach expands exponentially through billions of AI agents, our human capacity to understand and guide these interactions remains fixed.

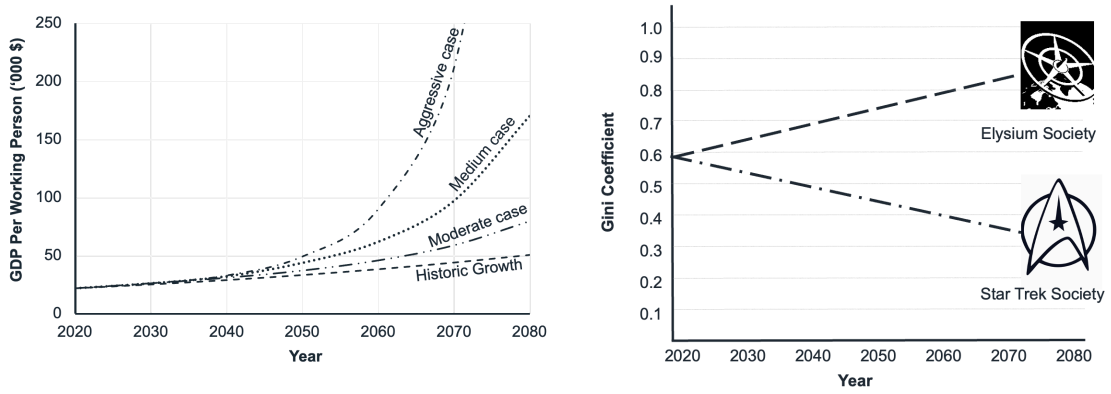


Fig 1: (a) Reliable AI agents will increase the size of the "working population" that will include humans and their agents, significantly, and transform all aspects of our $100T economy. This will accelerate economic progress, increasing the earning per working person from 50% to over 10 times in the most aggressive scenario. The post scarcity economy arrives after 2100 in "Moderate", by 2080 in "Medium" and by 2060 in "Aggressive" scenario (see Appendix A) (b) Gini index captures societal inequality where higher value is more unequal. The global average Gini coefficient is 0.67 with Slovakia being most equal (Gini=0.24) and South Africa (Gini=0.95) being most unequal among large countries. With the advent of Agentic AI, will we build an Elysium or Star Trek society?

This choice manifests vividly in competing visions of our technological future. The Star Trek (Okuda& Okuda, 2011) universe depicts an optimistic 23rd-century where advanced coordination protocols have eliminated material wants, allowing humanity to focus on exploration and advancement. Their replicator technology represents not just material abundance, but sophisticated mechanisms ensuring equitable distribution and sustainable use. In stark contrast, the film Elysium (Wikipedia, 2013) presents a dystopian 2154 where Earth is overpopulated and impoverished, with wealthy elite enjoying advanced medical technology in an orbital habitat. The tragedy isn't technological limitation - their medical pods could heal any illness - but rather the failure to develop protocols for equitable access and distribution. The difference lies not in technological capability, but in how these societies coordinate and distribute their resources.

This stark contrast- Star Trek's abundance through coordination versus Elysium's scarcity through fragmentation - underscore our challenge clearly: will our technological capabilities enhance or diminish human flourishing? The answer lies not in the raw power of our technology, but in how we orchestrate its use across populations. To understand how we might approach this challenge, we must first understand how nature has solved similar problems of scale and coordination.

B From Nature to Networks: Interaction Protocols in Complex Systems

Nature offers profound insights into solving massive-scale coordination challenges, as extensively documented in foundational studies of collective animal behavior (Hölldobler & Wilson, 1990). Consider how army ants construct living bridges: when confronted with a gap in their path, each ant follows a remarkably simple protocol. If there's an ant in front, cross the bridge. If not, become part of the bridge. This minimal set of rules, when executed by thousands of ants simultaneously, creates remarkably resilient and resource-efficient structures that no individual ant could comprehend or design (Reid et al, 2015).

This simple example illustrates the fundamental power of protocols - rules of interaction that transform local actions into sophisticated collective behavior. The true power of protocols lies in how they simultaneously serve individual and collective needs. For individuals, protocols extend their effective reach far beyond their cognitive limits by providing clear rules for action in complex situations. Each ant participates in building a stable bridge without needing to understand the overall architecture. For collectives, protocols enable coordination at scales that would be impossible through direct communication or central control. The ant colony builds and adapts structures, beyond any individual's cognitive capacity, without any central planning or sophisticated individual decision making. A back-of-the-envelope calculation reveals that the strength-to-weight ratio of an ant bridge is 1,000-24,000 times higher than that of a human-built concrete bridge.

However, the ant bridge represents the simplest case - where individual and collective interests naturally align. Nature shows us that protocol discovery becomes increasingly challenging as incentive structures grow more complex. In multicellular organisms, where all cells share identical genetic interests, evolution has discovered reliable protocols for coordinating millions of cells - all somatic cells can only propagate through the germline (Michod, 2007). More complex scenarios emerge in bacterial colonies and fish schools, where protocols must balance individual survival with group benefits (West et al., 2006; Couzin et al., 2005). These systems demonstrate nature's solution to a fundamental challenge: how to achieve sophisticated collective behavior through simple, local interaction rules even when individual and group interests partially conflict.

These natural systems offer important lessons for addressing complex societal challenges. Some coordination problems can be managed through competitive mechanisms, where market protocols channel individual profit-seeking into efficient resource allocation. However, many of our greatest challenges present true social dilemmas where individual and collective interests fundamentally misalign, and no market mechanism can bridge the gap (Ostrom, 1990). The COVID-19 pandemic demonstrates this complexity: its course was shaped not by any single

decision but by millions of interlinked choices about testing, isolation, and vaccination. Here, individual incentives for normal social interaction conflicted directly with collective needs for isolation and distancing. Similar dynamics appear in climate change, where individual benefits from carbon emissions create collective harm, and in humanitarian crises, which often stem not from resource scarcity but from failures to align distribution protocols.

What unites these diverse cases - from cellular organization to global pandemics - is that effective protocols must work across both massive scales and diverse incentive structures. The key challenge lies in discovering protocols that can coordinate billions of individual actions while adapting to complex, often misaligned incentives. This principle becomes crucial as we consider how to coordinate beneficial collective behavior at scales far beyond human cognitive limits. Nature shows us this is possible, but we need systematic approaches to discover, validate, and implement protocols at unprecedented scales.

This progression reveals why traditional AI approaches, focused on enhancing individual agent intelligence, hit fundamental limits when addressing collective challenges. Current large language models excel at processing information and making individual decisions, but they cannot inherently solve problems requiring massive-scale coordination. Even current multi-agent systems, while showing promise in small groups, remain constrained to human-scale interactions. The path forward requires a fundamental shift in focus - from making smarter agents to enabling smarter interactions. ***The future of AI is protocol-centric.***

C. The Dunbar Ceiling: From Intelligence to Interconnection

Throughout history, each major civilization advance has been marked by the development of new coordination protocols that transcend cognitive limits of individual minds. **These protocols succeeded by embedding intelligence in the rules of interaction rather than requiring individuals to understand the entire system (Daston, 2023).** Yet each advance revealed a fundamental tension between protocol reach and human cognitive capacity.

Early writing systems weren't just for recording information - they were protocols that enabled asynchronous coordination across time and space. While these systems could disseminate knowledge to millions, individuals could deeply engage with only a few key texts.Mesopotamian clay tablets reveal sophisticated protocols for trade and governance that allowed cities to function beyond personal relationships, yet merchants could maintain meaningful relationships with only a handful of trading partners. Medieval guilds developed protocols for knowledge transfer that preserved complex crafts across generations, but each master could effectively train only a small number of apprentices.

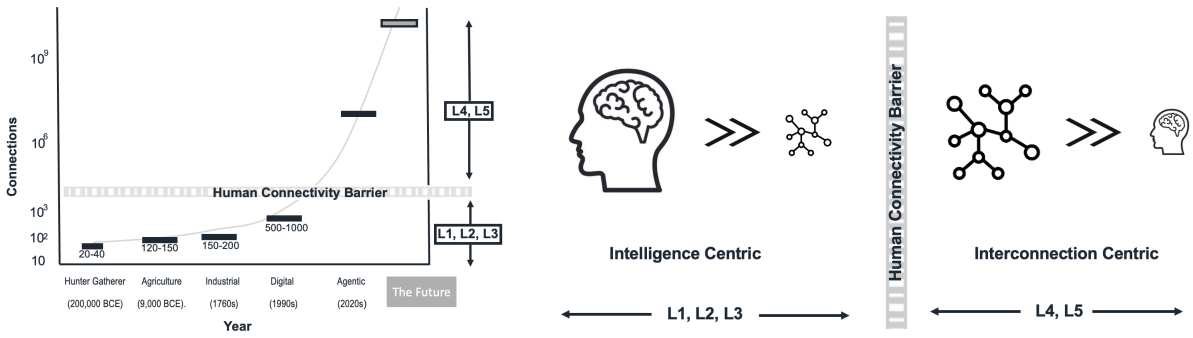


Fig 2: (a) The outermost barrier of the total number of faces or individuals one can recognize, namely 1,500 people. We call it the "Human Connectivity Barrier". This is approximately 10 times of the famous "Dunbar number", which is 150 individuals . Current AI systems with LLMs are also constrained to this scale. (b) Future of agentic systems will overcome this barrier and transform human coordination abilities. This requires a new perspective to scaling AI that focuses on interconnections over intelligence. LPMs will enable this via a protocol-centric view to AI.

The industrial revolution marked a quantum leap in protocol sophistication. Railway timetables weren't mere schedules - they synchronized society across vast distances. Telegraph networks weren't just communication tools - they decoupled information flow from physical movement. Yet even as these systems coordinated millions, individual operators could only manage a limited number of connections and routes. The digital age represents humanity's most ambitious protocol project, with the internet's layered protocols like TCP/IP (Cerf, 1974; Clark, 1985) enabling global coordination without central control. Cryptocurrencies further demonstrate this evolution, showing how carefully designed protocols can achieve consensus without central authority.

Yet each advance, while expanding our collective reach, remained fundamentally constrained by cognitive limits as humans struggled to process the increasing complexity of interactions. Even in our hyper-connected era, while social media theoretically connects billions, meaningful engagement remains bounded to few hundreds or thousands. This barrier - known as Dunbar's number - represents the cognitive limit on meaningful relationships humans can maintain, approximately 150 stable relationships and 1500 recognizable faces (Atlantic, 2021). This constraint isn't just about individual relationships; it limits the complexity of protocols themselves, as they must remain comprehensible to the humans who design and oversee them.

The imminent emergence of billions of AI agents presents not just a challenge of scale, but an opportunity to transcend these cognitive limits entirely. Unlike historical protocols that had to remain "human-readable", agentic protocols can operate at complexities and scales far beyond human comprehension while still producing beneficial collective and *individual* outcomes. When agents execute protocols on our behalf, they can simultaneously process millions of

information sources, maintain complex relationships with thousands of other agents, and coordinate actions across massive networks while adapting to real-time feedback.

D. The Path for Protocol-centric Agentic AI

The key lies in building agentic systems that can orchestrate billions of complex interactions. However, current AI has largely focused on the opposite - maximizing individual agent intelligence. While several recent works in multi-agent AI like Smallville (Park et al. 2023), Concordia (Vezhnevets et al. 2023), Project Sid (Altera et al 2024), AdaSociety (Huang et al, 2024) demonstrate sophisticated individual behaviors, they remain constrained to small populations (10-1000 of agents) operating in purely synthetic environments. These systems can be deployed within human interaction patterns - as personal assistants, tutors, A/B testers, or task-specific helpers; but face inherent scalability limitations. This prevents them from tackling real-world societal challenges involving millions of individuals. Recent studies validate these "limits of agency", showing that successful modeling of real-world phenomena requires population scale over individual sophistication. Empirical work predicting city-wide behavior across New York City finds that simpler agents at massive scale (8.4 million) consistently outperform more sophisticated LLM-based agents in limited numbers (Chopra et al, 2025). This validates a key insight from nature: **sophisticated collective behavior emerges not from individual intelligence but through interaction protocols operating at population scale.**

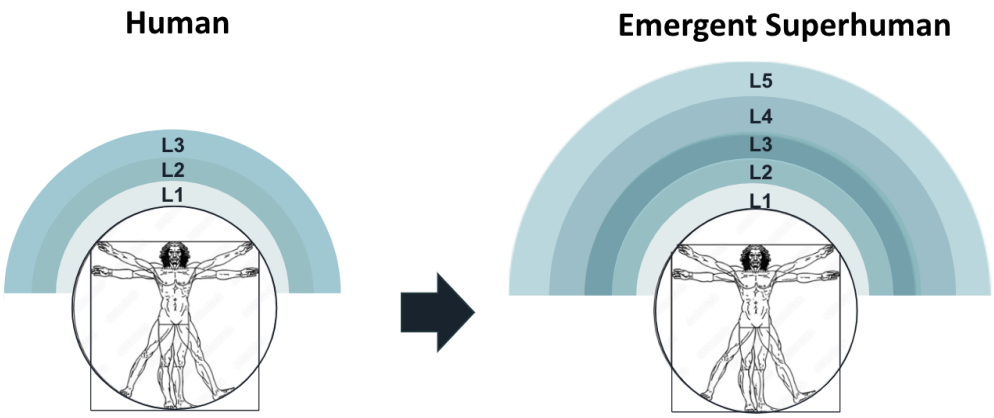
Two recent breakthroughs point towards transcending these limits. First, by making agent-based simulations differentiable, we can compose diverse interaction protocols - from mobility patterns to economic transactions - while maintaining computational tractability at scale (Chopra et al 2023; Quera-bofarull et al 2023). This enables learning from historical data to establish realistic baselines for outcomes and incentives. Second, advances in privacy-preserving computation enable decentralizing the agent-based simulation and bridge it with real-time behavior. These protocols allow real individuals to participate in population-scale coordination while preserving privacy, creating a crucial feedback loop: synthetic protocols efficiently explore possible dynamics, while physical protocols capture actual behavioral data and incentive structures. (Chopra et al 2024).

Fig 3: A paradigm shift in AI systems: While LLMs focus on processing multimodal data to enhance individual capabilities, LPMs focus on designing and implementing protocols to shape collective behavior. This represents a fundamental transition from enhancing individual behavior to orchestrating population-scale coordination.

This progression leads to Large Population Models (LPMs), which represent a fundamental shift from enhancing individual intelligence to discovering scalable coordination protocols. LPMs combine: i) Differentiable agent-based modeling to capture passive incentives through data-driven learning at scale; ii) Decentralized physical protocols to capture active incentives through real-world interaction; iii) Compositional architectures that jointly model and optimize both synthetic and physical protocols (Chopra et al 2023; Chopra et al 2025). LPMs are already creating real impact. They're being used to help immunize millions of people by optimizing vaccine distribution strategies (Osri et al 2025), and to track billions of dollars in global supply chains, improving efficiency and reducing waste (Adiga & Chopra et al 2025). The next section introduces a framework for how we progress from intelligence-dominated systems (powered by LLMs) to protocol-dominated systems (enabled by LPMs) through five distinct levels.

E. Levels of Agentic Systems

The evolution of agentic systems follows a trajectory similar to the development of autonomous vehicles, with each level reducing human cognitive burden while increasing system responsibility. This progression represents distinct levels that progressively enhance human capability to engage with and influence increasingly complex systems.



Dimension	State
Level 1: Perceive	Stable
Level 2: Automate	Current
Level 3: Connect	Current
Level 4: Navigate	Emerging
Level 5: Transform	Future

Fig 4: "Living under the Agentic API": The Vitruvian Man represents one of humanity's first systematic studies of human physical limits and proportions, with Da Vinci precisely inscribing the human figure within a circle and square to demonstrate natural constraints and mathematical harmony. We reframe this iconic image to illustrate how agentic systems

help transcend cognitive limits. While L1-L3 systems operate within natural human cognitive boundaries (like the circle containing Da Vinci's figure), L4-L5 systems establish new protocols that enable coordination and collective intelligence at scales previously thought impossible - by orchestrating interactions with millions of other humans globally through novel interconnection LPM protocols.

Level 1: Perceive

Agents help humans understand complex global contexts

These systems augment our sensing capabilities by integrating massive information streams into actionable insights, similar to how lane departure warnings enhance driver awareness. When a restaurant owner checks COVID risk levels, L1 systems combine epidemiological data, local patterns, and business factors to provide clear situational awareness. Retrieval-augmented generation exemplifies this level (Lewis et al. 2020), integrating static knowledge with real-time data. Intelligence focuses on processing information, but coordination remains human-driven.

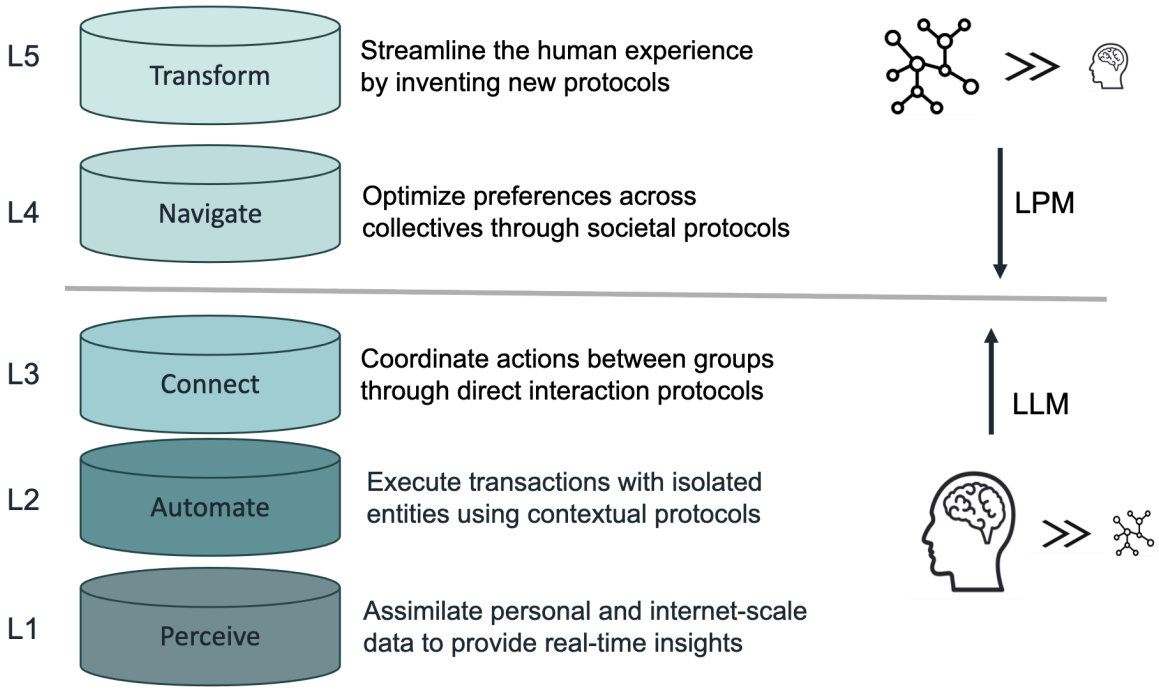


Fig 5: The progression of protocol sophistication across agentic system levels. Lower levels (L1-L3) focus on enhancing individual and small-group capabilities through traditional protocols. Higher levels (L4-L5) enable coordination beyond human cognitive limits through novel protocols that can orchestrate millions of simultaneous interactions. This represents a fundamental shift from optimizing within human bounds to discovering entirely new forms of collective coordination.

Level 2: Automate

Agents execute tasks with global awareness

These systems enable autonomous execution of specific tasks with built-in contextual intelligence, comparable to cruise control maintaining vehicle speed. They handle individual

transactions with global awareness, like procurement systems that automatically adjust orders based on supply chain patterns. Tool-augmented (or browser-use) language models represent this capability, executing complex actions through API interactions while maintaining contextual understanding. The key advance is extending from information processing to contextual action.

Level 3: Connect

Agents enable fluid coordination with immediate networks

These systems facilitate dynamic coordination within bounded networks, functioning like traffic-aware cruise control responding to nearby vehicles. They enable synchronized activities within local ecosystems (~100s of agents), such as restaurants coordinating safety protocols with neighbors. Multi-agent frameworks like AutoGen (Wu et al., 2023) demonstrate this capability, enabling structured collaboration between specialized agents - planning meetings and parties. However, they remain constrained by human social scaling limits.

The Protocol-Scale Transition The progression through these first three levels shows how agentic systems can enhance human coordination within our natural cognitive bounds. L1 extends our perception, L2 augments our actions, and L3 amplifies our social coordination. However, these systems fundamentally operate within human cognitive constraints - they can't orchestrate behavior beyond Dunbar's number or process relationships outside human social limits. The next two levels represent a fundamental breakthrough: rather than enhancing individual intelligence, these systems enable beneficial collective behavior at unprecedented scales through sophisticated protocols.

Level 4: Navigate

Agents guide decisions using population-scale insights

These systems discover optimal coordination strategies through massive-scale simulation, similar to autonomous vehicles operating in well-mapped areas. Through differentiable agent-based modeling (Chopra et al., 2023), they can simulate millions of interactions to identify unintuitive patterns - like optimal restaurant reservation spacing to reduce congestion. While they can discover effective protocols through simulation, implementation remains primarily synthetic.

Level 5: Transform

Agents amplify individual actions into collective impact

These systems create living feedback loops between individual actions and population-scale outcomes, analogous to fully autonomous vehicles reshaping traffic patterns. They integrate simulation with decentralized physical protocols (Chopra et al., 2024) to enable secure, privacy-preserving coordination at unprecedented scales. The breakthrough is implementing

discovered protocols through coordinated physical and digital infrastructure while maintaining rigorous safety guarantees.

Intelligence to Protocol: The Core Transition This progression from L1-L3 (intelligence-dominated) to L4-L5 (protocol-dominated) marks a crucial evolution in human coordination capability. Rather than pushing against cognitive limits by making individual agents smarter, we transcend these limits by discovering protocols that naturally guide beneficial collective behavior at scale. Just as self-driving vehicles promise to revolutionize transportation by removing human limitations from the equation, protocol-centric systems promise to revolutionize individual potential by enabling coordination at unprecedented scales. The transition shows how agentic systems can help humans participate effectively in increasingly complex systems - moving from enhanced individual capability (L1-L3) to meaningful participation in population-scale coordination (L4-L5).

F From Understanding to Shaping: The L4-L5 Transition

The transition from Level 4 to Level 5 represents a fundamental shift in how protocols evolve - from data-driven reaction to active reality-shaping through real-time adaptation. This breakthrough enables us to transcend human cognitive limits through protocol networks that bridge understanding and action. While our focus is on enabling beneficial collective behavior at population scale, this progression mirrors challenges in other domains like autonomous vehicles with L5 self-driving will transform not just individual cars but entire transportation systems.

Dimension	Level 4: Orchestrate	Level 5: Harmonize
Primary Domain	Simulated environments	Physical reality
Individual's Role	System observer	System shaper
Decision Support	Shows possible outcomes	Guides collective evolution
Infrastructure	Works within existing systems	Creates new protocols
Real-world Impact	Through recommendations	Through direct integration
Feedback Loop	One-way learning	Continuous adaptation

Fig 6: L4 vs L5 Systems - From Understanding to Shaping Reality. While L4 systems use simulation to understand and recommend actions within existing infrastructures, L5 systems actively shape reality by orchestrating both digital interactions and physical protocols. This progression mirrors how traffic systems might evolve from predicting congestion to actively preventing it through coordinated control of both routing algorithms and traffic signals.

To understand this transition, consider today's traffic systems. Navigation apps like Google Maps and Waze can predict congestion patterns and suggest alternate routes, but fundamentally remain reactive - they can only respond to emerging patterns based on historical data. In contrast, imagine future traffic systems that could actively shape collective behavior through coordinated protocols, helping individuals navigate more efficiently while contributing to overall system optimization. Such systems wouldn't just predict bottlenecks - they would prevent them via synchronized adjustments to both individual recommendations and physical infrastructure. When emergency vehicles need priority, they can create "green corridors" - coordinating traffic signals and routing nearby vehicles to maintain both emergency response times and overall traffic efficiency. This same conceptual leap - from understanding patterns to actively shaping them - characterizes the transition we're enabling through Large Population Models.

Consider how this progression manifests in pandemic response infrastructure. Level 4 operates through differentiable agent-based models, allowing us to simulate millions of synthetic agents while maintaining end-to-end gradients through their interactions. These simulations help individuals understand how their choices interact with collective behavior and reveal counterintuitive protocols to align their incentives. For instance, when individuals need to decide on testing , L4 simulation can process multiple interaction protocols simultaneously - from disease transmission dynamics to mobility patterns to local intervention strategies - revealing that prioritizing testing speed over accuracy can create better collective outcomes by enabling faster isolation responses (Romero-brufau 2021). However, these insights remain primarily predictive - they can inform individual decisions but cannot directly coordinate real-world response.

Level 5 achieves its transformative power by extending differentiability from simulated to physical protocols. The key technical innovation lies in coupling differentiable simulations with decentralized agent-based models running on physical devices. When a rapid test identifies a new infection, the system doesn't just record this data - it dynamically updates protocol implementation. Through privacy-preserving computation, it can identify which subset of the population in the affected area should activate contact tracing capabilities, effectively creating targeted surveillance networks that adapt to emerging transmission patterns. This real-time protocol adjustment demonstrates how L5 systems 'backpropagate through reality' - using observed outcomes to continuously refine both simulation parameters and protocol deployment strategies. This isn't just prediction - it's active protocol implementation that adapts based on real-world feedback.

The potential lies in creating a continuous feedback loop between simulation and reality through protocol networks. This coupling can manifest through diverse mechanisms - from

privacy-preserving gradient computation in contact tracing networks (Chopra et al., 2024) to consensus protocols in payment systems to adaptive routing in mobility networks. Essentially, synthetic protocols discover effective strategies through differentiable simulation, decentralized protocols implement these strategies through privacy-preserving computation on edge devices, and real-world outcomes feedback to improve the simulations. When an outbreak pattern emerges, the system doesn't just predict its spread - it actively reshapes transmission dynamics through coordinated adjustments to testing station placement, ventilation systems, and population mobility guidance. For instance, testing facilities may dynamically adjust their operations based on real-time transmission data. Building management systems adjust ventilation based on occupancy patterns. Vaccination strategies adapt to emerging variants and changing population hesitancy. Most importantly, these physical protocols maintain privacy and security while propagating gradients back to the synthetic models, allowing continuous protocol refinement without compromising individual data. Each individual's participation becomes part of a living protocol network that continuously adapts to changing conditions while maintaining privacy and agency.

The key distinction between L4 and L5 lies in their relationship with reality. While L4 excels at discovering protocols through simulation, L5 creates "living protocols" that adapt and evolve through real-world implementation. This capability to actively shape collective behavior while preserving individual agency and privacy represents a fundamental advance in human coordination capacity - one that will enable us to transcend current cognitive limits through increasingly sophisticated protocol networks. This progression from understanding to active orchestration characterizes the L4-L5 transition across domains with misaligned incentives. The verification of these protocols, particularly as they begin to shape physical infrastructure, presents new challenges that we must carefully address as these systems develop. However, this ability to bridge simulation and reality - to not just model but actively guide beneficial collective behavior - represents a crucial step toward protocol-centric artificial intelligence.

Connection to Mechanism Design: This progression from L4 understanding to L5 shaping shares intellectual roots with mechanism design (Maskin, 2008), particularly recent work on algorithmic mechanisms for social good (Abebe & Goldner, 2018; Koster et al., 2022). However, Large Population Models (LPMs) introduce several key innovations that enable unprecedented scale and adaptivity. While traditional mechanism design focuses on creating static rules that align incentives for relatively small groups of agents, LPMs enable dynamic, population-scale protocols that continuously evolve through real-time feedback between simulation and reality. Our computational approach—using differentiable simulations and privacy-preserving gradient computation to synchronize synthetic and physical worlds—allows us to discover and refine protocols that coordinate billions of simultaneous interactions. This is fundamentally different from classic mechanism design which typically operates at human-comprehensible scales with

fixed rules. Our "backpropagation through reality" approach, which allows protocols to continuously adapt based on observed outcomes while maintaining privacy and individual agency, represents a leap beyond traditional mechanism design's focus on mathematical guarantees for static mechanisms. This approach aligns with Rahwan's (2018) vision of "Society-in-the-loop" programming, where algorithmic systems are shaped by collective human feedback. However, L5 systems extend this concept by enabling real-time protocol adaptation while preserving individual agency.

G. The Human Experience: From Information to Orchestration

To understand how these systems might transform our daily lives, let's follow two individuals navigating complex real-world challenges through increasingly sophisticated AI agents working on their behalf.

Sarah Chen: Managing a Restaurant During the Pandemic

Sarah Chen, a 32-year-old restaurant owner in Kansas City, experiences how AI agents progressively enhance her ability to run her business safely during COVID-19. Her L1 agent acts as a personal COVID analyst, continuously monitoring global data streams and local patterns to alert her about risks. When a new variant emerges in nearby counties, it immediately assesses the implications for her restaurant, translating complex epidemiological data into clear, actionable insights. At L2, multiple specialized agents handle specific tasks autonomously. Her inventory agent adjusts supply orders based on changing consumer patterns and safety requirements. Her safety agent manages cleaning schedules and staff health protocols, automatically updating procedures when health guidelines change. Each agent handles complex decisions while keeping Sarah informed and in control. Her L3 agents work with neighboring businesses' agents to create a safer local ecosystem. They coordinate delivery schedules to minimize cross-exposure, share real-time safety alerts, and collectively manage supplier relationships. When one restaurant detects a potential exposure risk, all connected businesses can respond promptly and appropriately. At L4, Sarah's planning agent leverages massive simulations to discover surprising but effective strategies. By analyzing millions of restaurants' experiences, it identifies counterintuitive insights - like how spacing reservations by 22 minutes instead of the standard 15 reduces lobby crowding by 40% while maintaining table utilization. These discoveries come from recognizing patterns in data that no human could process. With L5, Sarah's agents actively participate in shaping the community's pandemic response. Her ventilation system doesn't just react to local conditions - it works with neighboring buildings to create optimal air flow patterns across the block. Her capacity management agent coordinates with other restaurants to distribute dining demand safely throughout the district. When an

elevated transmission risk emerges, her agents automatically adjust operations as part of a coordinated community response that protects both her business and public health.

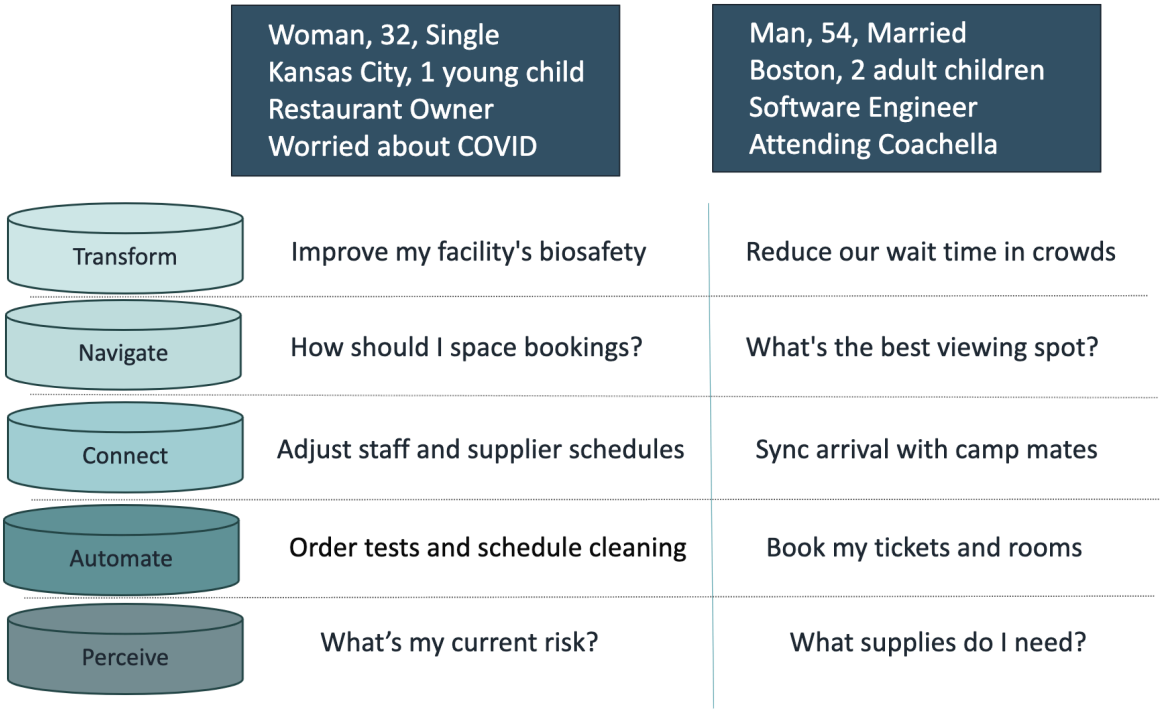


Fig 7: From Information to Orchestration: How agentic systems evolve in real-world scenarios. Following two individuals - a restaurant owner navigating COVID-19 protocols and a music-lover planning his trip to Coachella - we see how interactions progress from simple information gathering (L1) through automated tasks (L2), small group coordination (L3), population-scale simulation (L4), and finally to active shaping of collective behavior (L5). Each level reduces cognitive burden while increasing the system's capacity to coordinate beneficial outcomes at scale.

Michael Roberts: Navigating Coachella

Michael Roberts, a 54-year-old software engineer from Boston, sees how AI agents transform his first Coachella experience from potentially overwhelming to smoothly orchestrated. His L1 agent serves as a personal festival expert, analyzing years of Coachella data to create his optimal experience strategy. It considers everything from historical crowd patterns to weather forecasts, helping Michael prepare like a veteran attendee. L2 brings multiple agents handling specific tasks - one managing accommodations and transport timing, another monitoring ticket upgrade opportunities, and a third organizing his schedule and supplies. When better options become available, his agents secure them automatically while keeping him informed. At L3, his agents coordinate seamlessly with his festival group's agents. Instead of endless group texts about meeting points and schedules, the agents synchronize everyone's movements and preferences automatically. When plans need to change, the agents smoothly reorganize meetups while

respecting everyone's interests. His L4 planning agent uses festival-wide simulations to navigate efficiently. It discovers optimal strategies by modeling millions of attendees' behaviors - like learning that visiting water stations 10 minutes before set changes avoids peak crowds, or identifying viewing spots that balance sound quality with crowd density. The agent helps Michael make better choices by understanding how the entire festival flows. With L5, Michael's agents actively help shape the festival experience. When he heads to a popular performance, his agent doesn't just predict crowd movement - it participates in creating efficient flow patterns. His agents work with the festival's systems to subtly adjust crowd distribution - perhaps delaying his water break by five minutes helps prevent a bottleneck while ensuring he stays hydrated. Each small adjustment contributes to smoother festival flow while maintaining his perfect day.

The Emergence of Collective Intelligence These examples reveal how AI agents evolve from personal assistants to shapers of collective behavior. Both Sarah and Michael start with immediate individual concerns - pandemic safety and festival navigation. Through increasingly sophisticated agents, they become participants in larger, more effective coordination networks. The progression shows how technology can enhance rather than replace human experience. Each level maintains individual autonomy while enabling better collective outcomes. L1-L3 agents help people navigate existing complexity. The crucial advance comes with L4-L5, where agents first understand and then help shape the behavior of entire populations. This transformation suggests a future where technology helps us transcend traditional coordination limits while preserving personal agency. Most importantly, these systems demonstrate how individual and collective benefits can align through well-designed protocols. Whether managing pandemic response or festival crowds, the right coordination mechanisms help everyone achieve better outcomes than they could alone. This points toward a future where technological advancement enhances rather than diminishes human flourishing.

Crucially, as we develop Level 5 systems that shape reality through coordinated protocols, we face a fundamental verification challenge: ensuring safety before deployment. This challenge isn't unprecedented - Byzantine distributed systems prove consensus protocols before deployment (Ren 2017, Correira 2010), and smart contracts verify economic mechanisms before execution (Almakhour 2020, Nam 2022). However, population-scale agentic systems present uniquely complex challenges because they must verify emergent behaviors that arise from millions of individual decisions. Like neural networks, these systems may be resistant to formal verification. Instead, we need new frameworks that combine:

- Bounded verification: Proving critical safety properties and invariants where possible
- Empirical validation: Developing rigorous metrics to measure collective behaviors in simulation

- Incremental deployment: Creating standards for safely scaling protocols while monitoring key social indicators

These verification challenges are fundamental to realizing L5 systems. While we cannot achieve mathematical certainty, we must develop robust frameworks to validate population-scale protocols before allowing them to shape real-world behavior.

H. Conclusion

As we move toward a world with billions of AI agents, understanding different types of collective behavior becomes essential. While traditional AI seeks to replicate human intelligence in machines, protocol-centric intelligence asks a different question: how can we design rules of engagement that enable beneficial collective behaviors to emerge naturally at scale? The answer requires carefully distinguishing between different classes of coordination challenges.

Some collective behaviors, like cryptocurrency networks, emerge naturally through market mechanisms where individual incentives align with system goals. In these domains, carefully designed protocols can achieve coordination without central control because competitive dynamics (like mining rewards and market trading) naturally drive beneficial outcomes. However, many of our greatest challenges - from pandemic response to climate change to resource conservation - represent fundamentally different coordination problems. These are social dilemmas where individual and collective interests misalign, creating "tragedies of the commons" that markets alone cannot solve. (Ostrom, 1990; Kollock, 1998)

This is precisely where LPMs and their evolution from L4 orchestration to L5 discovery become crucial. Unlike market-based systems that rely on aligned incentives, LPMs explicitly model how different protocols shape both individual choices and collective outcomes under various incentive structures (Chopra et al., 2023). Through massive-scale simulation, they can discover and validate coordination mechanisms that help bridge the gap between individual and collective interests. For instance, when modeling pandemic responses, LPMs don't just assume rational individual behavior - they explore how different testing, vaccination, and mobility protocols might help align personal health choices with public health needs (Romero-Brufau et al., 2021; Chopra et al., 2024).

The path toward beneficial collective outcomes lies not in blindly applying market mechanisms, nor in forcing centralized control, but in discovering protocols appropriate to each domain's unique incentive landscape [Helbing, 2012]. Through LPMs, we can systematically explore and validate these protocols before deploying them in the real world.

This brings us back to Seldon's vision in Foundation (Asimov, 1951) - not just predicting large-scale behavior but actively shaping it toward better outcomes through carefully designed

rules of interaction. Will we build Star Trek's utopia (Okuda & Okuda, 2011) or Elysium's dystopia? The answer lies not in the raw intelligence of our AI agents, but in our ability to discover protocols that effectively bridge individual and collective interests across diverse coordination challenges [Pentland, 2014]. The time to focus on this protocol-centric future - with its full complexity and domain-specific demands - is now.

References

(1) Asimov, I. (1951). Foundation. Gnome Press

(2) United Nations, Department of Economic and Social Affairs, Population Division (2023). World Population Prospects 2022: Summary of Results

(3) Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2019). World Population Growth. Our World in Data.

(4) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

(5) Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf.

(6) Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. Journal of Human Evolution, 22(6), 469-493

(7) Hill, R. A., & Dunbar, R. I. M. (2003). Social network size in humans. Human Nature, 14(1), 53-72

(8) Okuda, M., & Okuda, D. (2011). The Star Trek Encyclopedia. Simon and Schuster

(9) Elysium (2013), Wikipedia

(10) Hölldobler, B., & Wilson, E. O. (1990). The Ants. Harvard University Press

(11) Chris R. Reid, Matthew J. Lutz, Scott Powell, Albert B. Kao, Iain D. Couzin, and Simon Garnie. Proceedings of the National Academy of Science, Nov. 2015

(12) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30

(13) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901

(14) Yang, Y., et al. (2023). Modeling large-scale population dynamics for multi-agent systems. In Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, 1731-1739

(15) Horvitz, E., & Paek, T. (2021). Population-scale AI: Challenges and opportunities in modeling collective behavior. Nature Machine Intelligence, 3(6), 473-484

(16) Cerf, V. G., & Kahn, R. E. (1974). A protocol for packet network intercommunication. IEEE Transactions on Communications, 22(5), 637-648

(17) Clark, D. D. (1988). The design philosophy of the DARPA Internet protocols. ACM SIGCOMM Computer Communication Review, 18(4), 106-114

(18) Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences, 99(suppl 3), 7280-7287

(19) Pentland, A. (2014). Social physics: How good ideas spread-the lessons from a new science. Penguin

(20) Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence and group performance. Current Directions in Psychological Science, 24(6), 420-424

(21) Helbing, D. (2012). Social self-organization: Agent-based simulations and experiments to study emergent social behavior. Springer-Verlag Berlin Heidelberg.

(22) Vicsek, T., & Zafeiris, A. (2012). Collective motion. Physics Reports, 517(3-4), 71-140.

(23) Epstein, J. M. (2012). Generative social science: Studies in agent-based computational modeling. Princeton University Press.

(24) Smith, J. M., & Szathmary, E. (1997). The major transitions in evolution. OUP Oxford.

(25) Daston, L. (2023). Rules: A Short History of What We Live By (Vol. 13). Princeton University Press.

(26) Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

(27) Ren L, Nayak K, Abraham I and Devadas S., 2017. Practical Synchronous Byzantine Consensus. IACR Cryptol. ePrint Arch., 2017, p.307

(28) Correia, M., Veronese, G.S., Neves, N.F. and Verissimo, P., 2011. Byzantine consensus in asynchronous message-passing systems: a survey. *International Journal of Critical Computer-Based Systems*, *2*(2), pp.141-161

(29) Almakhour, M., Sliman, L., Samhat, A.E. and Mellouk, A., 2020. Verification of smart contracts: A survey. *Pervasive and Mobile Computing*, *67*, p.101227

(30) Nam, W. and Kil, H., 2022. Formal verification of blockchain smart contracts via atl model checking. *IEEE Access*, *10*, pp.8151-816

(31) https://www.theatlantic.com/family/archive/2021/05/robin-dunbar-explains-circles-friendship-dunbars-number/618931/

(32) Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. Annual review of sociology, 24(1), 183-214

(33) Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge university press.
(34) Voshmgir, S. (2019). Token Economy: How Blockchains and Smart Contracts Revolutionize the Economy. BlockchainHub.