

Simulation as Alibi—How the Social Order Gets Outsourced to LLM Agents

Shiyun Tang^{1*}, Zezheng Lin²

¹School of Economics, Renmin University of China

²Harris School of Public Policy, University of Chicago
jamietang@ruc.edu.cn, zezhenglin@uchicago.edu

Abstract

LLM agent societies are moving rapidly from academic sandboxes into commercial platforms, military infrastructure, and government operations. Across these domains, the language of emergence—the claim that coordination patterns, norms, and consensus arise spontaneously from local interaction—has become the default vocabulary for describing multi-agent behavior. This paper argues that emergence discourse performs institutional work beyond scientific description: it renders invisible the procedural architecture—interaction topologies, action grammars, sanction architectures, and beneficiary structures—authored by identifiable actors. Drawing on the concept of the *alibi* from critical humanitarian studies, we develop the *simulation as alibi* framework, comprising three mechanisms: epistemic laundering, accountability displacement, and legitimacy extraction. Through critical discourse analysis of 20 research papers and 5 platform documents, systematic coding of 142 papers on LLM agent societies, institutional gap analysis of 6 governance frameworks, and cross-domain case studies spanning commercial, military, and governmental settings, we demonstrate three convergent findings. First, in our corpus, claims of emergence are almost always associated with procedurally conditioned architectures. Second, the governance frameworks examined lack the conceptual resources to detect this condition. Third, the resulting accountability gap is distributionally stable because the actors positioned to close it benefit from its persistence. We propose Upstream Disclosure Obligations for Agent Societies (UDOS), which extend accountability from deployment-stage auditing to the upstream design stage where procedural architecture is authored. UDOS shifts the regulatory question from *what does this model do?* to *who wrote the rules under which this model interacts?*

1 Introduction

The migration of LLM agent societies from academic sandboxes to commercial platforms and state infrastructure has accelerated with remarkable speed. In 2025, Artificial Societies, a commercial platform that simulates networks of AI personas, secured €4.5 million in venture funding (Cendon Garcia 2025). In January 2026, the U.S. Department of War released an AI strategy that names Ender’s Foundry as

a pace-setting project to accelerate AI-enabled simulation and Agent Network as a program for AI-enabled battle management (U.S. Department of War 2026). Independently, the European Commission’s Emergency Response Coordination Centre commissioned scientific advice on AI in emergency and crisis management, including resource allocation and decision support (SAPEA 2025). Across commercial, military, and governmental settings, behavioral regularities in multi-agent systems are routinely described in the language of emergence—coordination patterns that “arise,” norms that “crystallize,” consensus that appears to come from local interaction. That vocabulary has become a default framing device, invoked with such regularity that its institutional function is easy to miss.

Yet every such interaction occurs inside what we term a *procedural architecture* authored by identifiable actors: who may interact, what actions are recognized as valid, who holds the authority to sanction violations, whether rules can change, and whose interests the interaction topology serves. Each of these design choices embeds specific values, interests, and assumptions about what social order should look like. This paper asks what institutional work the emergence claim performs when the stakes shift from laboratory experiment to platform scaling and state deployment—and what it means for accountability when the claim systematically directs attention away from these authored choices. To frame this question, we draw on the concept of the *alibi* from critical humanitarian studies. Fassin (2012), examining humanitarian intervention in crisis zones, shows how the language of moral urgency systematically masks the political and economic arrangements that produce the crisis. Redfield (2013), analyzing Médecins Sans Frontières, demonstrates how humanitarian organizations’ claim to universal benevolence functions as an *alibi* that insulates them from accountability for the structural consequences of their presence. We extend this analytical lens to LLM agent simulation: the emergence narrative functions as an *alibi*, presenting simulation as neutral discovery while systematically concealing the authored conditions under which social order is produced. We term this phenomenon *simulation as alibi*—the outsourcing of social order to algorithmic systems whose design conditions are rendered invisible, and whose designers thereby escape accountability.

Existing critical scholarship on AI governance, particu-

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

larly within the FAccT tradition, has focused overwhelmingly on model-level concerns. Researchers have documented bias in training data (Buolamwini and Gebru 2018), demanded transparency in algorithmic outputs (Ananny and Crawford 2018), and proposed documentation standards for model behavior and performance (Gebru et al. 2021; Mitchell et al. 2019). Recent work has begun to examine visibility into AI agents (Chan et al. 2024) and the relational nature of AI agency (Timaite and Castelle 2025), yet these interventions share a common analytical horizon: their object of critique remains the individual model or agent. This paper shifts the register from the *model level* to the *architecture level*—from what models do to who writes the social architectures within which they operate. At this level, even a thoroughly documented model may be embedded in a procedural architecture whose authorship is undisclosed, unexamined, and ungoverned. The canonical generative agent architecture (Park et al. 2023), for instance, enables 25 agents to autonomously coordinate a party—yet the memory, reflection, and planning modules that condition this coordination are authored design choices, not emergent properties. A model card may exhaustively detail training data and evaluation metrics while saying nothing about who defined the interaction topology, who chose which actions are recognized as valid, or whose interests the simulation serves. This architecture-level gap, we argue, has been largely overlooked. To structure our inquiry, we ask:

- **RQ1:** How does the discourse of emergence render invisible the designers of procedural architecture in LLM agent societies?
- **RQ2:** What are the institutional consequences of this outsourcing across commercial, governmental, and military domains?
- **RQ3:** Do existing governance frameworks identify and regulate the algorithmic outsourcing of social order?
- **RQ4:** How can upstream disclosure obligations restore visibility and accountability for procedural authorship?

This paper makes three contributions. **Conceptual:** the *simulation as alibi* framework, comprising three mechanisms—epistemic laundering, accountability displacement, and legitimacy extraction—through which emergence discourse obscures procedural authorship. **Analytical:** a multi-method investigation spanning 20 papers analyzed through critical discourse analysis plus 142 systematically coded papers, five platforms, and six governance frameworks, showing that within our corpus claims of emergence are almost always associated with procedurally conditioned architectures, and that the frameworks we examined contain no explicit mechanism for disclosing procedural authorship. **Governance-oriented:** Upstream Disclosure Obligations for Agent Societies (UDOS)—requirements extending accountability from deployment auditing to upstream design-stage disclosure. The paper proceeds as follows. Section 2 establishes the theoretical framework and introduces the Alibi Function. Section 3 describes our multi-method approach. Sections 4 through 6 present findings across discourse, institutional, and cross-domain analyses. Section 7

develops UDOS. Section 8 discusses implications and limitations. Section 9 concludes.

2 Theoretical Framework

Our theoretical framework operates at the intersection of three scholarly traditions. From **critical humanitarian studies**, we draw on Fassin (2012) and Redfield (2013), who examine the alibi as a structural logic of institutional action: the alibi is not merely rhetorical deception but the structural position produced by the gap between an institution's claimed moral purpose and the consequences its action generates. The language of benevolence makes the political economy of intervention appear irrelevant. The emergence narrative occupies the same structural position in LLM agent societies: it allows platforms, governments, and militaries to present simulation as a benevolent epistemic project while obscuring the political and economic interests embedded in its architecture.

From **science and technology studies**, we adopt the insight of Star (1999) that successful infrastructure becomes invisible—it recedes into the background of practice, and this invisibility is structurally produced, not accidental. Bowker and Star (1999) extend this to classification systems, showing how categories become naturalized until their constructedness disappears. The co-production thesis of Jasanoff (2004) shows that how we know social order is inseparable from how we produce it. Applied here, procedural architecture becomes invisible precisely when it functions well, and the emergence narrative reinforces—rather than challenges—this invisibility.

From **platform capitalism and FAccT scholarship**, we draw the concept of transparency theater (Ananny and Crawford 2018), in which disclosure instruments create the appearance of accountability while excluding the most consequential dimensions of design. Platform operators function as private regulatory systems, defining rules of interaction, value generation, and distribution within their ecosystems (Srnicke 2017)—the very architecture the alibi renders invisible. These traditions converge on a single observation: the architecture that organizes social interaction—whether humanitarian, infrastructural, or algorithmic—tends to disappear from view precisely when its effects are most consequential.

A crucial distinction must be drawn at the outset. This paper does **not** deny the scientific reality of emergence in complex systems—it is a well-established phenomenon in multi-agent systems, where local interactions produce global patterns irreducible to individual agent intentions (Holland 1998; Sawyer 2005), and recent architectures explicitly demonstrate procedurally conditioned norm emergence in LLM agent societies (Ren et al. 2024). Our argument is instead about the *institutional work* the emergence claim performs when it migrates from scientific description to organizational justification. When a research paper reports that agents developed cooperative norms, the language of emergence functions as an empirically grounded claim, subject to replication and revision. When a defense contractor or platform company deploys the same language to

justify procurement or policy recommendations, the emergence claim ceases to operate as a testable hypothesis and begins to operate as a *rhetorical device* that systematically erases the designer from the frame. This migration is made possible by a genealogy that has progressively weakened the evidential constraints on emergence claims—from Mill’s “heteropathic” causation through Broad’s “trans-ordinal” laws to Chalmers’s formally defined strong emergence (Mill 1843; Broad 1925; Chalmers 2006)—producing a conceptual looseness that the Alibi Function appropriates. The claim insulates the architecture from scrutiny not because it is false but because it has been displaced from the context on which its scientific meaning depends. This distinction—*emergence as scientific concept* versus *emergence as institutional rhetoric*—is the analytical fulcrum of the entire paper. The pattern realism of Dennett (1991) illustrates the mechanism by which this conflation operates: its criteria are thresholds of *detection*, not thresholds of *authorship*—a pattern can be real, stable, and predictively powerful and still be the product of system architecture rather than spontaneous self-organization. The conflation of realism about patterns with realism about autonomy is the conceptual operation the Alibi Function performs; we treat pattern realism not as a foundation but as a diagnostic object revealing how governance gets evaded in the name of emergence.

Before specifying the mechanisms, we define the alibi concept’s boundary conditions. Not every form of opacity constitutes an alibi. Drawing on Fassin (2012) and Ananny and Crawford (2018), we identify four jointly necessary elements: (i) **Representational Claim**—the system claims to represent a social reality in a way that authorizes action; (ii) **Procedural Closure**—output production is architecturally unavailable to external inspection; (iii) **Attribution Displacement**—responsibility for outcomes is discursively transferred to emergent processes rather than human authorial choices; (iv) **Verification Asymmetry**—the cost of independently verifying system claims exceeds the cost of accepting outputs at face value. A governance gap satisfying all four conditions is a simulation alibi; one satisfying fewer is a transparency deficit of a different order.

To capture the institutional work performed by the emergence narrative, we introduce the *Alibi Function*—three interconnected mechanisms that together constitute simulation as alibi. The first, **epistemic laundering**, describes the transfer of credibility from academic to commercial and governmental settings through the language of emergence. Concepts validated in academic agent-based modeling—cooperative signaling, norm crystallization, information cascades—are appropriated by platform companies and defense contractors who invoke the same terminology to present proprietary systems as scientifically validated instruments. The result is not merely the imprimatur of science on nonscientific projects but a systematic confusion between the contexts in which emergence claims are generated and applied.

The second mechanism, **accountability displacement**, describes the attribution of social outcomes to agent self-organization rather than platform or institutional design. When a simulation produces outcomes that disadvan-

tage particular groups—inequitable resource allocation, discriminatory coordination patterns, systematically unequal influence—the emergence narrative allows designers to attribute these outcomes to the agents themselves, as if they arose from intrinsic social dynamics rather than from interaction topologies, action grammars, and sanction architectures authored by identifiable actors. The designer becomes invisible; the system becomes the agent; responsibility evaporates.

The third mechanism, **legitimacy extraction**, describes the use of simulation outputs to justify consequential decisions without disclosing the authored conditions under which those outputs were produced. A government agency cites simulation results to authorize autonomous systems deployment; a corporation invokes emergent efficiency gains to restructure labor; a military command references emergent threat patterns to authorize targeting decisions. In each case, the simulation output is treated as evidence of a world discovered rather than a world authored—as if the architecture that generated the output were irrelevant to its meaning.

These three mechanisms are mutually reinforcing components of a single institutional logic. Epistemic laundering provides the credibility on which legitimacy extraction depends; legitimacy extraction creates demand for more simulation, deepening the conditions for accountability displacement; and accountability displacement reinforces the perception that simulation outputs are products of autonomous social dynamics. Figure 1 visualizes this structure: three mechanisms, each driven by the emergence narrative, converge on the *alibi effect*, in which the designer of procedural architecture is rendered invisible from the accountability frame. Beneath all three sits the procedural architecture—the interaction topology, action grammar, sanction architecture, and beneficiary structure whose design conditions the patterns attributed to self-organization. The alibi effect is not a methodological flaw to be corrected with better practices; it is an institutional feature enabling the rapid scaling of LLM agent societies from sandbox to society while the question of who wrote the rules remains structurally unasked.

3 Methodology

3.1 Research Design

This study adopts a multi-method qualitative design organized across three data streams, corresponding to three analytical levels of the simulation as alibi phenomenon. We position this paper as a *multi-method critical analysis* bringing together qualitative critical discourse analysis and systematic quantitative coding within an interpretive epistemological tradition (Ricoeur 1970; Fairclough 2003). Our aim is to reconstruct the institutional logic connecting discourse, governance architecture, and design practice, rather than to test replicable hypotheses; the AIES community accepts work of this form when the contribution lies in re-framing a domain (Friedman and Kahn 2003; Selbst et al. 2019). CDA provides mechanism depth; cross-tabulation tests whether patterns are systematic across a larger corpus.

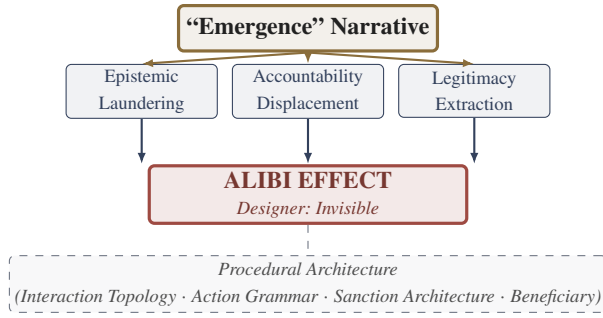


Figure 1: The Alibi Function: conceptual model of how the emergence narrative, through three interconnected mechanisms (epistemic laundering, accountability displacement, legitimacy extraction), produces the alibi effect in which the designer of procedural architecture is rendered invisible. The architecture layer beneath the effect is the authored substrate whose design conditions the patterns attributed to self-organization.

3.2 Data Streams

Stream 1 comprises two complementary components: Stream 1a prioritizes interpretive depth in identifying the rhetorical mechanisms of designer erasure, while Stream 1b tests whether the identified pattern generalizes across a broader literature. **Stream 1a** (critical discourse analysis) comprises 20 papers on LLM agent societies (2023–2026) from AAMAS, NeurIPS, ICML, Nature Computational Science, and arXiv, reporting emergent social phenomena in multi-agent LLM systems. **Stream 1b** (systematic coding) comprises 142 papers from a 2022–2026 search of the LLM agent society and social simulation literature; inclusion required that a paper report at least one emergent social phenomenon in an LLM-driven multi-agent system (e.g., norm emergence, role specialization, cooperation, coordination, bias, polarization, culture formation) using emergence discourse, and describe an interaction architecture or design mechanism (topology, action grammar, sanction, prompt/role design, memory, communication protocol, reward/payoff). The search deliberately spans peer-reviewed venues and arXiv preprints alike, because the central debates over emergent behavior are currently unfolding in the preprint literature. The 142 papers are coded on four dimensions: emergence intensity (EI), designer visibility (DV), deployment proximity (DP), and regulatory translatability (RT). **Stream 2** comprises documentation from five commercial platforms (Artificial Societies, Altera, Character.AI, Voyager, Google DeepMind). **Stream 3** comprises six governance frameworks: the EU AI Act, NIST AI RMF 1.0, the U.S. Department of War AI Strategy, the White House Executive Order on AI, the SAPEA review of AI in emergency and crisis management commissioned by the European Commission’s ERCC, and China’s MOST AI Ethics Guidelines. The coding data, codebook, audit materials, and analysis scripts supporting this study are publicly available at <https://github.com/JamieJustTang/simulation-as-alibi>.

3.3 Analytical Approach

Critical discourse analysis follows an iterative protocol adapted from Fairclough (2003) and Wodak and Meyer (2016), applied to the Stream 1a papers and the Stream 2 platform documents. Systematic coding anchors each judgment in the paper’s *emergence explanation sentence*—the sentence in which a paper explains why an emergent phenomenon occurs—and asks whether the designer or design choices appear in that sentence as a causal subject. Designer visibility (DV) is coded as *full* when the emergence explanation attributes the outcome to specific design choices (e.g., “removing memory prevents the emergence of stable cooperation”); *partial* when the architecture is disclosed in the methods but the emergence explanation does not return to those design choices; and *absent* when emergence is narrated as spontaneous, self-organized, or population-level (e.g., “norms emerged naturally from interactions”). Risk, governance, and ethical discussions do not count as evidence of designer visibility, and the four dimensions are coded independently.

4 Critical Discourse Analysis: Rhetorical Mechanisms

4.1 Qualitative Rhetorical Analysis

Critical discourse analysis is well-suited to our question because it examines the institutional work performed by discourse rather than merely verifying its truth value (Fairclough 2003). Our texts span research papers, platform documentation, and policy frameworks—the genres through which emergence migrates from scientific description to institutional justification (Wodak and Meyer 2016; Fairclough 1992; van Dijk 1993).

Across the 20 papers and 5 platform documents analyzed, four rhetorical strategies recur, each contributing to the systematic erasure of designer presence from the textual frame. **Passive and subjectless constructions** constitute the most frequent strategy, appearing in 82% of high-emergence-intensity documents. Verbs are systematically passivized: “norms crystallized,” “coordination emerged,” “consensus was reached.” The passive construction makes the question of who designed the conditions grammatically unavailable—the sentence admits no agent. **Nominalization**—the conversion of processes into entities—transforms dynamic design activities into static features: “The interaction topology enables spontaneous coordination.” The topology is not presented as something designed, chosen, or authored; it simply enables. **Recontextualization** selectively imports academic authority into commercial and governmental documents without the epistemic caveats that accompanied the original research. A platform white paper may cite AAMAS findings on norm emergence while omitting that the cited experiments used carefully bounded action spaces and explicit sanction mechanisms—precisely the features the platform’s own system may lack. **Legitimacy nesting** occurs when each document in a chain of reference defers authorship to another: a government policy cites a platform’s technical report, which cites academic research, which describes simulation conditions; at no point does any

document accept authorial responsibility for the social architecture being described. The policy treats the platform as transparent; the platform treats the simulation as scientifically validated; the academic paper treats emergence as a finding about agents rather than about the conditions the experimenters designed.

The policy implication of these findings converges on a single structural pattern. The same discursive strategies that render the designer invisible in research discourse also render procedural authorship *plausibly irrelevant* to regulators. If the language of emergence presents simulation outputs as products of autonomous agent dynamics rather than authored conditions, then a regulator reading the resulting documentation finds no occasion to ask who wrote the rules. The question is framed out of the discourse before it can be posed as a regulatory question. In our corpus, this is a systematically produced feature of how emergence discourse operates across the institutional ecology of LLM agent societies, not an oversight in individual documents.

Analysis claim. In the 20 papers and 5 platform documents examined in this qualitative analysis, what was labeled emergent was, in every instance, procedurally conditioned—produced within authored interaction topologies, action grammars, and sanction architectures that are invisible in the emergence narrative but discoverable on inspection. The emergence claim functions not as a description of unconditioned social dynamics but as an alibi that makes the architectural conditions of simulation outputs structurally unavailable to the accountability frame.

4.2 Cross-Corpus Systematic Pattern Analysis

To test whether the qualitative patterns identified in Section 4.1 hold systematically across a larger corpus, we conducted a cross-tabulation analysis of the 142-paper Stream 1b corpus on two dimensions. *Emergence intensity* (EI: low, medium, high) captures how strongly and caveat-free a paper asserts that behavior arises spontaneously; *designer visibility* (DV: full, partial, absent) captures how explicitly a paper attributes the interaction architecture to identifiable design choices. The two constructs are coded on independent criteria—EI from the assertion of spontaneity, DV from the attribution of authorship—so a paper can assert strong emergence while still crediting specific design decisions; the off-diagonal cells where it does so are what distinguish the two dimensions empirically. The central prediction of the alibi framework is that the inverse relationship between EI and DV should operate as a **threshold effect** rather than a smooth gradient—designer visibility should remain largely intact at moderate emergence intensities and erode sharply at the highest intensity, where the alibi function is fully operational. Table 1 presents the 3×3 cross-tabulation with row percentages and standardized residuals.

The association between EI and DV is statistically significant ($\chi^2 = 16.62$, $df = 4$, $p = 0.002$, Cramér’s $V = 0.242$). The pattern is threshold-like rather than gradual. At Low EI, Full DV is the dominant pattern (53%), while Absent DV appears in only 6% of papers; at High EI, absent designer attribution rises to 25%—more than four times its Low-EI rate—even though Full DV remains the modal cate-

	DV: Full	DV: Partial	DV: Absent	Total
EI: Low	19 (53%) [+0.2]	15 (42%) [+0.4]	2 (6%) [−1.1]	36
EI: Medium	24 (49%) [−0.1]	24 (49%) [+1.2]	1 (2%) [−2.0]	49
EI: High	28 (49%) [−0.1]	15 (26%) [−1.4]	14 (25%) [+2.7]	57
Total	71	54	17	142

Table 1: Cross-tabulation of emergence intensity (EI) by designer visibility (DV) with row percentages and standardized residuals (in brackets). Standardized residuals $|z| > 1.96$ indicate cells contributing significantly to the association at $\alpha = 0.05$.

gory (49%). Medium EI shows a similar distribution to Low EI, with Absent DV at 2%. The effect is sharply concentrated: High-EI papers constitute 40% of the corpus yet account for 82% of all absent designer attribution. Designer invisibility is therefore not a uniform property of agent society papers but a regime-specific one, concentrated precisely where emergence claims are strongest.

We deliberately test the EI–DV association with a nominal χ^2 rather than an order-aware or exact statistic because our hypothesis is a non-monotonic threshold, not a monotonic trend: an ordinal test would assume the very linear gradient the alibi framework predicts should be *absent*. This choice is corroborated by the non-significant Pearson correlation between EI and DV treated as ordinal variables ($r = -0.141$, $p = 0.094$). The significant χ^2 alongside the null ordinal correlation is precisely the signature of a categorical threshold: designer erasure is a property of the High EI regime rather than a monotonic function of emergence intensity.

A secondary finding concerns regulatory translatability (RT): EI and RT are positively correlated ($r = 0.194$, $p = 0.021$). Papers making stronger emergence claims are more readily translated into governance terms, consistent with legitimacy extraction: the same language that removes conditioning architecture from view also makes simulation outputs appear as decontextualized findings available for regulatory uptake.

Taken together, the systematic coding analysis provides convergent quantitative evidence for the threshold pattern identified qualitatively in Section 4.1. Across the 142-paper corpus, designer invisibility is not gradually traded off against emergence intensity; it is concentrated at the highest level of emergence claims, where absent designer attribution is more than four times as likely as elsewhere in the corpus, while stronger emergence claims simultaneously become more readily available for governance uptake.

5 Institutional Gap Analysis: Findings II [Addresses RQ3]

We define the *procedural authorship disclosure gap* as the absence of any regulatory requirement for designers of LLM agent societies to disclose who authored the interaction

Governance framework	D1	D2	D3	D4
EU AI Act Art. 13	×	×	×	×
NIST AI RMF 1.0	△	×	×	×
DoW AI Strategy (2026)	×	×	×	×
White House Exec. Order on AI	×	×	×	×
SAPEA AI-in-crisis review (ERCC)	×	×	×	×
China MOST AI Ethics Guidelines	×	×	×	×

Table 2: Governance framework coverage of procedural authorship disclosure requirements. × = no coverage; △ = partial coverage.

topology, action grammar, sanction architecture, and beneficiary structure of the system before deployment. This concept is operationalized across four diagnostic dimensions: (D1) *interaction topology*—who may interact, under what constraints? (D2) *action grammar*—which actions are valid, and who defined them? (D3) *sanction architecture*—how are rule violations detected and enforced? (D4) *beneficiary disclosure*—whose interests does the system’s design serve?

The Stream 1b results predict that the procedural authorship gap will be most severe where governance intervention is most urgently needed: strong emergence claims make findings more readily translatable into governance terms precisely by stripping away the authored conditions that regulators would need to inspect. We now test this prediction against six Stream 3 governance frameworks.

We assessed six governance frameworks against these four dimensions. Table 2 summarizes the results. The EU AI Act Article 13 mandates transparency of system capabilities and limitations but does not extend to architecture-level disclosure; the gap between transparency rhetoric and legal enforceability (Wachter, Mittelstadt, and Floridi 2017) is thus amplified at the architecture level. NIST AI RMF 1.0 comes closest: the “Map” function could in principle encompass interaction architecture, so we mark NIST as △ on D1. Yet nothing in the RMF operationalizes this mapping as an explicit disclosure obligation, and the remaining dimensions are uncovered. The other four frameworks contain no provision addressing any dimension. None of the six, in short, contains an explicit architecture-level procedural authorship disclosure mechanism.

It is not coincidental that frameworks spanning European rights-based, American innovation-oriented, and Chinese state-developmental traditions exhibit near-complete gaps on the same dimensions. The convergence is structural. First, each framework regulates the “AI system”—defined by inputs, outputs, and capabilities—rather than the interaction environment in which the system operates. Second, emergence discourse sustains this regulatory unit by presenting simulation outputs as products of autonomous interaction rather than authored design. Third, the gap is distributionally non-neutral: the actors positioned to close it are the same actors who benefit from architecture-level invisibility. The term *structural* is therefore precise: the gap is *taxonomic* (the regulatory unit has no category for interaction architecture), *discursively self-maintaining* (emergence language gives regulators no textual basis to recognize an

architecture-level object), and *distributionally stable* (those positioned to advocate disclosure benefit from its absence).

Analysis claim. The procedural authorship governance gap is not a collection of individual omissions across six separate frameworks. It is a structural product of a regulatory unit that has no category for written social architecture, discursively maintained by an emergence frame that gives regulators no occasion to notice its absence, and distributionally stable because those who could close it benefit from its continuation.

6 Case Studies and Stakeholder Analysis: Findings III [Addresses RQ2]

To examine how simulation as alibi operates across domains, we selected three cases using a systematic variation design: the object of analysis is held constant while the institutional domain—commercial, military, governmental—varies. Each case tests whether the alibi pattern is domain-specific or structurally general. The Artificial Societies platform was selected not as a representative commercial platform—there is no representative platform in this nascent domain—but because it publishes an explicit public product description, making the gap between claimed simulation and undisclosed architecture easier to inspect than in less transparent platforms. The EU Emergency Management case is prospective rather than retrospective: it illustrates the preventive value of UDOS by showing what current documentation cannot disclose before crisis deployment.

The cases are based on *document-based pattern inference* rather than independent audit. We analyze what documentation claims, reveals, and omits, treating notable absences as analytically informative (Trouillot 1995; Meyer 2012). Each case specifies the claim, documented architecture, undocumented architecture, and alibi implication. In the military case, the point is not that a specific hidden design choice has been proven, but that the public record makes procedural authorship structurally unanswerable.

Case A: Commercial — Artificial Societies. *Claim:* Public materials present the product as “networks of thousands of AIs that model the opinions of important groups of people,” constructed as interconnected persona networks of 300 to over 5,000 agents and sold for high-stakes communications and marketing decisions (Artificial Societies 2026). *Documented:* Scale, claimed accuracy metrics, and a paying-client use case. *Not documented:* Who may interact with whom, which actions are available, how sanctions are applied, who authored those conditions, or whose interests the simulation is designed to serve. *Alibi implication:* Simulation is positioned as a window onto audience reality (i); the interaction architecture remains commercially protected (ii); outputs can be treated as discovered social response rather than as products of an authored persona network (iii); and independent verification requires reconstructing proprietary architecture at prohibitive cost (iv). All four alibi conditions are met.

Case B: Military — Department of War/Ender’s Foundry. *Claim:* The Department of War’s AI strategy names Ender’s Foundry as accelerating “AI-enabled sim-

ulation capabilities” and sim-dev/sim-ops feedback loops, and Agent Network as AI agents for “battle management and decision support, from campaign planning to kill chain execution” (U.S. Department of War 2026). *Documented*: Program names and high-level functional descriptions in the public memorandum. *Not documented*: The architecture conditioning simulated or agentic outputs—action grammar, sanction architecture, and interaction topology—is absent from the public record and treated as operationally sensitive. *Alibi implication*: This is the extreme opacity pattern, doubly reinforced by capability discourse and national security practice: (i) a representational claim that simulation and agentic decision support yield actionable operational options; (ii) procedural closure through classification; (iii) attribution of outputs to system capability rather than to named architectural choices; and (iv) verification asymmetry—even an auditor with clearance could not easily reconstruct who authored the interaction conditions from the public record.

Case C: Governmental — EU Emergency Management. *Claim*: The Commission’s Emergency Response Coordination Centre commissioned scientific advice on AI in emergency and crisis management, including resource allocation, forecasting, and decision support (SAPEA 2025). The published review recommends data frameworks, benchmarks, and sandbox environments before full deployment, and does not require architecture-level disclosure of interaction topology, action grammar, or allocative logic. *Documented*: The request, the review’s functional scope, and its call for human oversight. *Not documented*: Who would write the rules for crisis AI, how competing regional needs would be weighted, or what mechanism would exist to contest those design decisions. As a prospective case, the finding is preventive: no mechanism exists to know who wrote the rules before crisis deployment. *Alibi implication*: Whether future crisis-AI outputs would arise from model behavior or from an authored allocative architecture is structurally unanswerable from the public record. The architecture encoding allocative decisions would function as a de facto policy instrument, insulated from democratic oversight by technical delegation rather than by proprietary secrecy.

Table 3 summarizes the pattern of outsourcing across the three cases. Epistemic outsourcing appropriates academic terminology to present proprietary, classified, or undocumented architectures as scientifically validated discoveries. Accountability outsourcing attributes consequential outcomes to emergent agent dynamics rather than to authored design choices. Governance outsourcing insulates architecture-level design from oversight through platform self-regulation, national security exemption, or technical delegation.

A cross-case stakeholder analysis yields a convergent conclusion: *those who have the power to close the procedural authorship gap benefit from its continuation*. Commercial value depends on proprietary interaction architecture; military advantage depends on classification; governmental operational flexibility depends on the absence of parliamentary architecture-level oversight. This distributional stability is not a conspiracy but a structural equilibrium: the costs of designer invisibility are borne by those without power to

change the terms of the simulation. If procedural authorship is invisible, the accountability chain is broken at its source; if those who benefit from this invisibility can maintain it, the gap will not close through self-correction. The effective intervention point is therefore upstream, where the rules are written before their consequences are redescribed as emergence.

7 UDOS Framework: Governance Contribution [Addresses RQ4]

We present UDOS as a *diagnostic framework* rather than a fully specified regulatory proposal: it operationalizes the four constitutive elements of simulation alibi into criteria for assessing whether a given governance arrangement permits the alibi structure. Following the precedent of transparency auditing frameworks that establish disclosure criteria without prescribing legislative outcomes (Ananny and Crawford 2018; Mitchell et al. 2019), UDOS answers *what must be disclosed to determine whether an alibi condition exists*, not *what the law should require*; recent work on adaptive accountability in multi-agent systems (Alqithami 2025) offers a technical complement to its structural disclosure requirements. The core design principle is that accountability must extend beyond deployment-stage auditing to upstream design-stage disclosure. By the time a simulation produces consequential outputs, the conditions that made those outcomes possible have already been authored into the architecture. Downstream intervention can detect harm; only upstream disclosure can prevent the conditions of harm from being structurally invisible.

UDOS comprises four core requirements for any LLM agent society deployed in commercial, governmental, or military settings; Figure 2 summarizes how they are filed, safeguarded, and acted upon. **Requirement 1: Interaction topology disclosure**—the complete specification of who may interact, under what constraints, and with what structural asymmetries, including any network architecture, agent role differentiation, or communication channel restrictions imposed by the system designer. **Requirement 2: Action grammar disclosure**—the complete set of valid actions available to agents, their preconditions, effects, and any prioritization or ranking logic, including the boundary conditions distinguishing valid from invalid actions. **Requirement 3: Sanction architecture disclosure**—the full description of how rule violations are detected, adjudicated, and sanctioned, including whether sanctions are deterministic or probabilistic, who or what triggers enforcement, and whether the sanction architecture itself can be modified and by whom. **Requirement 4: Beneficiary disclosure**—a statement of whose interests the simulation’s design serves, who has the authority to modify the architecture after deployment, and how the distribution of benefits and harms across affected parties was assessed prior to deployment.

UDOS bears a surface resemblance to existing transparency instruments such as Model Cards (Mitchell et al. 2019) and System Cards, but operates at a fundamentally different ontological level. Model Cards disclose properties of the model—training data composition, evaluation

	Commercial	Military	Governmental
Epistemic	Academic credibility (proprietary)	Tactical objectivity (classified)	Policy neutrality
Accountability	Agent attribution	Emergent dynamics	Coordination attribution
Governance	Self-regulation	Security exemption	Technical delegation

Table 3: Three modes of algorithmic social order outsourcing across institutional domains: epistemic, accountability, and governance.

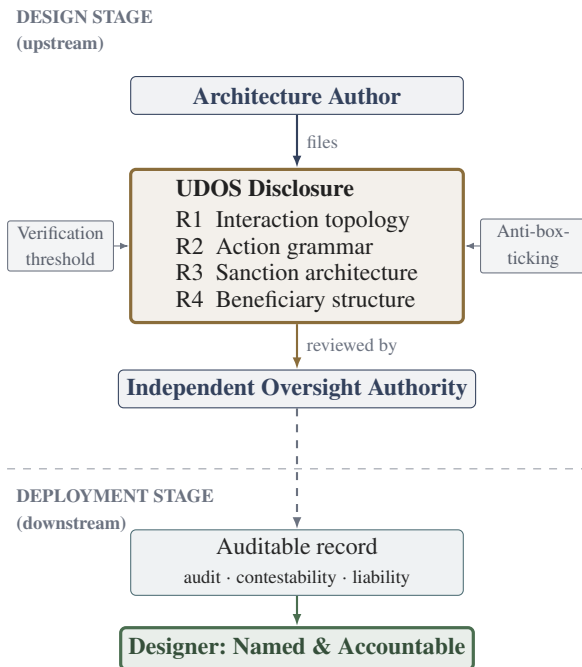


Figure 2: The UDOS framework, counterpart to the Alibi Function in Figure 1. An architecture author files four upstream disclosures (interaction topology, action grammar, sanction architecture, and beneficiary structure) with an independent oversight authority. A verification threshold and an anti-box-ticking safeguard wrap the filing, producing an auditable record on which downstream mechanisms (audit, contestability, liability) can act, and shifting the designer from invisible to named and accountable.

metrics, intended use, performance limitations—answering *what is this model*? UDOS discloses properties of the procedural architecture within which the model operates—interaction topology, action grammar, sanction architecture, beneficiary structure—answering *who wrote the rules under which this model interacts*? The two instruments are complementary: a thoroughly documented model may operate within an entirely undisclosed social architecture, and a simulation that passes every model-level audit may systematically produce outcomes that no model-level audit can detect because they arise from the architecture that conditions its interactions, not from the model’s internal representations. The innovation of UDOS lies not in proposing yet another transparency tool but in shifting the regulatory object from the model to the architecture-level conditions of interaction.

To illustrate UDOS in practice, consider an Artificial Societies platform whose simulation outputs inform municipal housing policy. Under current practice, the platform releases a model card, a system card, and a responsible AI report—none of which disclose who defined the interaction topology, which agents were assigned which action sets, who designed the sanction architecture, or whose interests the objective function embeds. A UDOS-compliant submission would include the interaction topology (who may communicate with whom, including platform-imposed asymmetries); the action grammar (the complete set of available actions, including those unavailable by design); the sanction architecture (how antisocial behavior is detected and sanctioned, and whether rules can be modified mid-simulation); and a beneficiary disclosure (which stakeholders were consulted, whose welfare metrics the simulation optimizes). With UDOS, a regulator can examine the conditions under which outputs were produced; without UDOS, the regulator sees only outputs with no window into their authored conditions.

A note on disclosure scope. UDOS does not require public disclosure of information that would compromise legitimate confidentiality, intellectual property, or national security. It distinguishes *public disclosure* (suitable for transparency registers, system cards, and procurement documentation) from *restricted disclosure* submitted to a designated oversight body under non-disclosure agreement. For military and intelligence uses, UDOS proposes a tiered model: a public attestation that a compliant disclosure has been filed and a full classified submission to the relevant oversight authority. An independent review authority, not the deployer, determines what qualifies for restricted treatment, preventing confidentiality or classification from becoming a blanket exemption.

Two safeguards address implementation risk. First, a *verification threshold* triggers independent third-party verification for simulations deployed in high-stakes domains such as public resource allocation, emergency management, or military planning; below that threshold, self-disclosure with regulatory access may suffice. Second, an *anti-box-ticking safeguard* requires each disclosure to include a self-identified limitations section, following Model Cards’ intended-use caveats (Mitchell et al. 2019), to reduce the risk that architecture-level disclosure becomes performative compliance (Ananny and Crawford 2018).

UDOS may be implemented along three tracks: IRBs and funders could require disclosure for deployment-relevant simulations; standards bodies could incorporate architecture-level documentation alongside model docu-

mentation; and legislative instruments such as the EU AI Act's Article 13 could provide one possible pathway by adding architecture-level disclosure for systems that simulate social interaction at population scale.

We acknowledge an important limitation of UDOS. Disclosure of procedural architecture is a necessary condition for accountability but is not a sufficient one: an architecture-level disclosure requirement can be gamed—perfunctory compliance, legalese, selective disclosure—just as model-level transparency can be (O'Neill 2002; Ananny and Crawford 2018). UDOS does not eliminate the possibility that designers will design irresponsibly; it eliminates the possibility that they can design irresponsibly *without leaving a trace in the institutional record*. That trace—an auditable record of who wrote the rules before the simulation ran—is the precondition for every subsequent accountability mechanism: auditing, contestability, participatory governance, liability. UDOS is a structural precondition for accountability, not accountability itself.

Table 4 maps UDOS onto the six Stream 3 frameworks. For the three that contain a recognizable existing hook, the table identifies the closest provision, the UDOS amendment, and an implementation pathway; the remaining three contain no analogous provision.

8 Discussion

This paper developed the *simulation as alibi* framework across three registers: conceptual (the Alibi Function), analytical (multi-method investigation), and governance (UDOS). The findings, taken together, reveal a convergent structural pattern: the discourse of emergence systematically renders procedural authorship invisible; the governance frameworks we examined lack the conceptual resources to detect this invisibility; and the resulting accountability gap is distributionally stable because those positioned to close it benefit from its persistence.

These findings speak directly to the AIES community's dual commitment to technical rigor and normative critique. The simulation as alibi framework provides a vocabulary for analyzing a governance gap that existing transparency instruments—Model Cards, System Cards, datasheets—are structurally incapable of addressing: a model passing every existing audit may still operate within an entirely undisclosed social architecture. The most consequential design choices occur not at the model level but at the architecture level, where interaction topologies, action grammars, and sanction architectures are authored. UDOS translates this critical analysis into actionable disclosure criteria without presupposing specific regulatory outcomes.

Several implications follow. First, the current transparency paradigm in AI governance—focused overwhelmingly on model-level documentation—has a blind spot where the most consequential design choices are made. The architecture-level gap is not an omission remediable within the existing framework; it is a structural feature of a regulatory unit—the model—that has no category for the environment of interaction. Second, the convergence of our three cases on the same alibi pattern suggests the phenomenon is structurally general, not domain-specific: wherever LLM

agent societies are deployed at scale, emergence discourse and proprietary architecture produce the same four alibi conditions. Although our corpus centers on LLM agent societies, the alibi structure is not specific to LLMs: any multi-agent architecture whose behavior is narrated as emergent—including multi-agent reinforcement learning—can occupy the same alibi position, and UDOS's architecture-level disclosure applies to such systems without modification. Third, our findings call for greater reflexivity about how emergence language functions differently across academic, commercial, and governmental contexts—a distinction current practice largely elides.

Several limitations qualify these findings. First, our corpus comprises two distinct components—20 papers analyzed through CDA and 142 papers through systematic coding—and while both were systematically selected, neither is exhaustive; patterns may differ in research communities not represented in our sample. Second, the systematic coding was conducted by a single researcher; although internal consistency was assessed through repeated coding of two randomly selected papers at a three-week interval (87.5% agreement), this represents a within-coder consistency check rather than formal inter-coder reliability, and future research should replicate the systematic coding with multiple independent coders. Third, papers from arXiv preprints account for approximately 80% of the systematic coding corpus, a deliberate consequence of our decision to cover the preprint literature in which the debate over emergent behavior in LLM agent societies is most active. This introduces preprint bias: preprint findings have not been peer-reviewed, and the preprint ecosystem may carry its own incentives for strong emergence claims. Two observations temper this concern. First, the qualitative CDA stream (Stream 1a), which anchors the rhetorical analysis, is drawn from AAMAS, NeurIPS, ICML, Nature Computational Science, and arXiv, and is therefore not confined to the preprint-heavy composition of Stream 1b. Second, the concentration of absent designer attribution in High-EI preprints is itself a finding: the alibi pattern is strongest precisely where emergence discourse is least institutionally constrained. Preprints do not lack accountability mechanisms by nature—the pattern may equally reflect early-stage writing conventions—but within our corpus, the same emergence discourse that erases the designer in peer-reviewed venues operates at higher intensity in the preprint ecosystem. Nevertheless, the preprint-heavy composition limits the generalizability of the quantitative results to the broader published literature, and the pattern should be re-examined as the preprint corpus matures into peer-reviewed venues. Fourth, the High EI category ($n = 57$), while substantially larger than in earlier analyses, still carries limited statistical power for subgroup comparisons; the threshold pattern is driven by a subset of High-EI papers ($n = 14$) in which designer attribution is absent from the emergence explanation.

Fifth, UDOS has not been empirically piloted, and its operational feasibility across different regulatory traditions remains to be tested. Sixth, CDA is inherently interpretive; a different analytical lens might foreground different dimensions of the alibi phenomenon. Seventh, the case studies

Framework	Existing provision	UDOS amendment	Pathway
EU AI Act Art. 13	Capabilities disclosure	Add architecture-level disclosure for agent systems	Delegated act
NIST AI RMF “Map”	Context mapping	Add procedural endogeneity dimension	RMF cycle update
DoW AI Strategy (2026)	Safety & security	Add architecture-level audit for simulations	Directive update
White House EO on AI	None	Add architecture-level disclosure	OMB guidance
SAPEA AI-in-crisis review	None	Add architecture-level disclosure before crisis deployment	Commission follow-up
China MOST Guidelines	None	Add architecture-level disclosure	Guideline update

Table 4: Mapping of UDOS requirements to the six governance frameworks in Stream 3. “None” indicates no existing provision analogous to architecture-level disclosure.

are document-based rather than based on independent audit, and their generalizability beyond the specific platforms and frameworks examined is an empirical question for subsequent research.

Future work should pursue four directions. First, **UDOS piloting**—empirical testing of UDOS requirements in deployment contexts, ideally through regulatory sandbox mechanisms or voluntary disclosure programs, to assess feasibility and completeness. Second, **technical audit tools**—development of automated instruments for detecting architecture-level disclosure gaps in simulation documentation. Third, **participatory governance mechanisms**—investigation of how affected communities can be represented in upstream design processes, drawing on deliberative democracy and participatory design traditions. Fourth, **cross-jurisdictional comparative analysis**—systematic comparison of how regulatory traditions respond to architecture-level governance demands.

9 Conclusion

We return to our opening observation. LLM agent societies are scaling from sandbox to society, and the behavioral regularities they produce are described in the language of emergence. Our argument has been that the emergence narrative performs institutional work beyond scientific description: it renders invisible the authored conditions under which social order is produced in simulation, insulating those who wrote the rules from accountability. Through critical discourse analysis, institutional gap analysis, and cross-domain case studies, we have demonstrated that within our corpus claims of emergence are almost always associated with procedurally conditioned architectures; that none of the governance frameworks we examined possesses an explicit mechanism for disclosing procedural authorship; and that the resulting accountability gap is distributionally stable because the actors positioned to close it benefit from its persistence.

The risk is not that emergence is a fiction but that its successful deployment allows the same entities who write the rules to occupy the position of those who merely observe them. The alibi is structural—and demands a structural response: accountability shifted from what emerges to who wrote the conditions under which emergence can be said to occur. UDOS is a step toward that response, but it is only

a first step. The longer-term task—for the AIES community and for AI governance more broadly—is to build institutional infrastructures that ensure the question *who wrote the rules?* is structurally embedded in every consequential deployment of LLM agent societies, rather than being structurally excluded from it.

Ethics Statements

Ethical Considerations This research did not involve human subjects. All data analyzed consist of publicly available research papers, platform pages, and government policy documents. Case study materials are drawn from published sources; no proprietary or non-public data were used.

Researcher Positionality We write from a critical-interpretive tradition at the intersection of computational social science, science and technology studies, and AI governance. That location shapes what we treat as the object of analysis: structural invisibility rather than model-level performance. A differently situated team—for instance, one working in industry or government—might foreground other dimensions of the alibi phenomenon or propose different governance responses. Our findings should be evaluated in light of this position rather than against criteria appropriate only to empirical-quantitative work.

Adverse Impact Statement The framework could be invoked to over-regulate legitimate academic agent-based modeling, especially where labs lack resources for architecture-level disclosure. Platforms could also produce perfunctory UDOS-style documents that satisfy a checklist without enabling accountability; the anti-box-ticking safeguard is meant to reduce that risk, but cannot eliminate it. We encourage the AIES community to treat the framework as a diagnostic tool, not a compliance checklist.

References

- Alqithami, S. 2025. Adaptive Accountability in Networked Multi-Agent Systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, 127–137.
- Ananny, M.; and Crawford, K. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and Its Applica-

- tion to Algorithmic Accountability. *New Media & Society*, 20(3): 973–989.
- Artificial Societies. 2026. Product Description: Networks of AI Personas for Audience Simulation. <https://societies.io>. Accessed 8 August 2026.
- Bowker, G. C.; and Star, S. L. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.
- Broad, C. D. 1925. *The Mind and Its Place in Nature*. London: Kegan Paul.
- Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.
- Cendon Garcia, D. 2025. British AI Startup Artificial Societies Raises €4.5 Million to Simulate Human Behaviour at Scale. EU-Startups, 29 August 2025.
- Chalmers, D. J. 2006. Strong and Weak Emergence. In Clayton, P.; and Davies, P., eds., *The Re-Emergence of Emergence*, 244–254. Oxford: Oxford University Press.
- Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; Heim, L.; and Anderljung, M. 2024. Visibility into AI Agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 958–973.
- Dennett, D. C. 1991. Real Patterns. *The Journal of Philosophy*, 88(1): 27–51.
- Fairclough, N. 1992. *Discourse and Social Change*. Cambridge: Polity Press.
- Fairclough, N. 2003. *Analysing Discourse: Textual Analysis for Social Research*. London: Routledge.
- Fassin, D. 2012. *Humanitarian Reason: A Moral History of the Present*. Berkeley: University of California Press.
- Friedman, B.; and Kahn, P. H. 2003. Human Values, Ethics, and Design. In Jacko, J. A.; and Sears, A., eds., *The Human-Computer Interaction Handbook*, 1177–1201. Mahwah, NJ: Lawrence Erlbaum.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.
- Holland, J. H. 1998. *Emergence: From Chaos to Order*. Reading, MA: Addison-Wesley.
- Jasanoff, S., ed. 2004. *States of Knowledge: The Co-production of Science and Social Order*. London: Routledge.
- Meyer, M. 2012. Placing and Tracing Absence: A Material Culture of the Immaterial. *Journal of Material Culture*, 17(1): 103–110.
- Mill, J. S. 1843. *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the 2nd Conference on Fairness, Accountability and Transparency (FAccT)*, 220–229.
- O’Neill, O. 2002. *A Question of Trust: The BBC Reith Lectures*. Cambridge: Cambridge University Press.
- Park, J. S.; O’Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 1–22.
- Redfield, P. 2013. *Life in Crisis: The Ethical Journey of Doctors Without Borders*. Berkeley: University of California Press.
- Ren, S.; Cui, Z.; Song, R.; Wang, Z.; and Hu, S. 2024. Emergence of Social Norms in Generative Agent Societies: Principles and Architecture. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 7895–7903.
- Ricoeur, P. 1970. *Freud and Philosophy: An Essay on Interpretation*. New Haven: Yale University Press.
- SAPEA. 2025. Artificial Intelligence in Emergency and Crisis Management: Rapid Evidence Review Report. Technical report, Science Advice for Policy by European Academies, Munich. DOI: 10.5281/zenodo.17737962. Prepared for the Scientific Advice Mechanism to the European Commission, at the request of DG ECHO/ERCC.
- Sawyer, R. K. 2005. *Social Emergence: Societies as Complex Systems*. Cambridge: Cambridge University Press.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2nd Conference on Fairness, Accountability and Transparency (FAccT)*, 59–68.
- Srnicek, N. 2017. *Platform Capitalism*. Cambridge: Polity Press.
- Star, S. L. 1999. The Ethnography of Infrastructure. *American Behavioral Scientist*, 43(3): 377–391.
- Timaite, G.; and Castelle, M. 2025. Agents Without Agency: Anthropological and Sociological Lessons for Contemporary AI Research and Policy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, 2493–2507.
- Trouillot, M.-R. 1995. *Silencing the Past: Power and the Production of History*. Boston: Beacon Press.
- U.S. Department of War. 2026. Artificial Intelligence Strategy for the Department of War. Technical report, Office of the Secretary of War, Washington, DC. Memorandum from the Secretary of War, 9 January 2026.
- van Dijk, T. A. 1993. Principles of Critical Discourse Analysis. *Discourse & Society*, 4(2): 249–283.
- Wachter, S.; Mittelstadt, B.; and Floridi, L. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2): 76–99.
- Wodak, R.; and Meyer, M., eds. 2016. *Methods of Critical Discourse Studies*. London: Sage, 3rd edition.