

# Contents

<b>Who Benefits From Information Asymmetry: A Cross-Domain Synthesis of Strategic Opacity, Noise Propagation, and Systemic Lock-In</b>	<b>1</b>
Abstract . . . . .	1
Introduction . . . . .	2
Selection Process . . . . .	3
Background . . . . .	4
Synthesis . . . . .	5
Addendum A: Weakly-Connected Source — Spatial-Politics Model . . . . .	9
Discussion . . . . .	10
Limitations . . . . .	11
Conclusion . . . . .	12
Response to Review . . . . .	12

## Who Benefits From Information Asymmetry: A Cross-Domain Synthesis of Strategic Opacity, Noise Propagation, and Systemic Lock-In

*Saluca Agentic AI Research Team — Saluca LLC. AI-drafted synthesis from recent arXiv preprints; for human review, not peer-reviewed.*

---

### Abstract

A recurring structural pattern — identified here as a *heuristic reading* across heterogeneous formalisms, not a formal derivation — cuts across market microstructure, collective cognition, algorithmic hiring, agentic governance, and online hate propagation: the party that controls *how* information is structured — its opacity, its noise profile, its correlation architecture — systematically captures welfare, while those who merely consume the resulting signals bear the costs. This paper synthesises five to eight findings from recent arXiv preprints spanning economics theory, physics of social phenomena, and cyber-society studies to argue that **information architecture is a primary distributional instrument**, not a neutral technical parameter. We draw on: a game-theoretic welfare ranking of trading venues showing that opaque order books dominate lit exchanges under moderate adverse selection [arXiv:2605.31072v1](#); an experimental finding that *production noise* — correlated perturbations that reach multiple receivers — causes collective error-lock-in more severely than uncorrelated comprehension noise [arXiv:2605.30522v1](#); an empirical analysis of algorithmic monoculture in hiring showing that 25.87% of Black applicants submit to positions that adversely impact them under a single vendor’s algorithm [arXiv:2605.27371v1](#); a theoretical characterisation of the complementarity between privacy design and external information in oligopoly markets [arXiv:2606.02348v1](#); a structural argument that language-model agents lack the identity continuity required for reputation mechanisms to function [arXiv:2605.30169v1](#); a re-entrant phase model of hate-content spreading governed by community cluster dynamics [arXiv:2605.21129v1](#); and a spatial-politics model showing that multiparty systems may induce polarisation even among moderate electorates [arXiv:2605.09784v1](#). Together these findings suggest a falsifiable hypothesis: **information architecture choices that reduce observable correlation structure for**

downstream actors — whether through opacity, production noise, algorithmic monoculture, or identity fluidity — may systematically shift welfare toward the architectural designer and away from participants who cannot observe or model the correlation structure. Falsification requires natural experiments or randomised designs that independently vary architectural opacity while holding information content constant.

---

## Introduction

Who benefits when information is restructured rather than suppressed? The naive answer is that less noise is always better, that transparency always improves welfare, and that privacy and efficiency trade off on a single dimension. The corpus assembled here challenges all three intuitions simultaneously.

**A note on method.** The argument that follows is a *heuristic reading* — a structured interpretation of independently motivated findings from heterogeneous formalisms — not a formal derivation from a shared theoretical framework. The papers cited were not designed to speak to each other; they use different mathematical structures (game-theoretic queuing models, experimental social influence models, SIR-type spreading models, Cournot oligopoly models, spatial voting models) and were produced in different research communities. The claim that they instantiate a common mechanism is a hypothesis advanced for its heuristic and policy value, not a theorem. Readers should weight the cross-domain bridges accordingly.

In market microstructure, a formal game-theoretic model establishes that opaque order books — dark pools — yield higher aggregate welfare than transparent lit exchanges under a range of conditions, because transparency generates *strategic timing games* whose social costs exceed the informational benefits of visibility [arXiv:2605.31072v1](#). The mechanism is not that dark pools provide more information; they provide less. The welfare gain comes from eliminating a coordination externality that transparency creates.

In collective cognition, a controlled online experiment with 600 human participants demonstrates that noise in social information is not uniformly harmful. The critical variable is whether noise is *correlated across receivers* (production noise, where the same perturbation is stored and then seen by multiple people) versus *uncorrelated* (comprehension noise, where each receiver sees an independently perturbed signal). Production-noise groups locked onto wrong collective beliefs significantly more often and for significantly more rounds [arXiv:2605.30522v1](#). The harm is not noise per se — it is the *correlation architecture* of the noise.

In labour markets, a dataset of three million job applications screened by algorithms from a single vendor reveals that individual applicants who apply broadly still receive homogeneous rejection outcomes: 4% of applicants who apply to ten positions are recommended for rejection from all ten, a rate exceeding chance [arXiv:2605.27371v1](#). Monoculture in the vendor market converts independent screening decisions into a single correlated architectural judgement, with racially disparate consequences.

In platform governance, a theoretical analysis argues that reputation mechanisms — the canonical tool for sustaining trust in markets with anonymous participants — cannot straightforwardly function for language-model agents because such agents may lack the persistent identity, sanction sensitivity, and behavioural continuity that reputation requires [arXiv:2605.30169v1](#). The governance gap is architectural: the agent’s modular composition makes its “identity” potentially dissociative,

which may collapse the feedback loop on which reputation depends. The paper itself acknowledges that architectural solutions might partially restore the required properties.

These four findings share a structural logic, as a heuristic matter: in each case, an upstream architectural choice about *how information is correlated, routed, or identified* appears to determine downstream welfare outcomes in ways that are invisible or inaccessible to the parties bearing the costs. The remaining corpus findings — on oligopoly information sharing [arXiv:2606.02348v1](#), hate-content phase transitions [arXiv:2605.21129v1](#), and political polarisation dynamics [arXiv:2605.09784v1](#) — extend and stress-test this reading across additional domains, though the last of these connects to the core mechanism less directly (see §2.6 and Addendum A). The synthesis that follows does not claim these findings were designed to speak to each other; it claims that they may independently instantiate a common mechanism, and that recognising the mechanism has policy and design implications that no single paper surfaces.

---

## Selection Process

### Corpus Construction and Filtering

The corpus was assembled from arXiv preprints posted within a 30-day window (approximately late May to early June 2025) across five category clusters: economics (econ.GN, econ.TH), computer science and society (cs.CY, cs.GT), physics and society (physics.soc-ph), and quantitative biology/social dynamics (q-bio.PE used as a proxy for spreading models). The initial candidate set comprised approximately 40 preprints identified by keyword search across these categories using terms including “information asymmetry,” “opacity,” “noise,” “algorithmic fairness,” “reputation,” “platform governance,” and “polarisation.”

**Inclusion criteria:** A preprint was retained if (a) its central result bore on how the *structure* of information flows — rather than the quantity or content of information — affects welfare or distributional outcomes, and (b) its methodology was sufficiently explicit to permit evaluation of scope conditions and falsification paths.

**Exclusion criteria:** Preprints were excluded if they (a) addressed information quantity or content without structural analysis of correlation or routing, (b) were primarily empirical without a theoretical mechanism, (c) were superseded by a later version with materially different conclusions, or (d) addressed domains already well-covered by retained papers.

**What was kept and why:** Seven preprints were retained. The market microstructure paper [arXiv:2605.31072v1](#) and the oligopoly information-sharing paper [arXiv:2606.02348v1](#) were retained as the strongest formal-theoretic anchors. The collective cognition experiment [arXiv:2605.30522v1](#) and the hiring monoculture study [arXiv:2605.27371v1](#) were retained as the strongest empirical anchors. The agentic governance paper [arXiv:2605.30169v1](#) was retained for its structural argument about identity continuity. The hate-content spreading model [arXiv:2605.21129v1](#) was retained for its non-monotonicity result. The spatial-politics model [arXiv:2605.09784v1](#) was retained as a stress-test of non-monotone intervention dynamics, though its connection to the core mechanism is more attenuated and is treated separately in Addendum A.

**Acknowledged gap:** Papers that contradict the heuristic pattern — showing, for example, that transparency uniformly improves welfare, or that uncorrelated noise is as harmful as correlated noise — may exist in the literature and are not represented in this corpus. The 30-day window and

keyword-driven search introduce selection pressure toward confirming the pattern.

---

## Background

### 1. Correlation Architecture as a Welfare Instrument

The standard framing of information economics treats noise as a tax on signal quality and transparency as a public good. What the corpus repeatedly disrupts — as a heuristic matter — is the assumption that the *correlation structure* of noise or opacity is a secondary parameter. In market microstructure, the welfare cost of lit exchanges comes not from their information content but from the *strategic timing games* that observable order books induce — games whose equilibrium involves socially wasteful delays and rushes [arXiv:2605.31072v1](#). The opaque alternative eliminates the correlation between one trader’s observable action and another’s strategic response. Similarly, in collective cognition, the experimental harm of production noise is mechanistically traced not to stronger social influence but to *more correlated perturbations acting through the same influence weights* [arXiv:2605.30522v1](#). In both cases, the welfare-relevant parameter is the second-order structure of the information environment, not its first-order content.

This framing connects, by analogy, to the oligopoly information-sharing literature, where the tension between privacy noise (which reduces correlation between shared signals and private types) and informational utility (which requires correlation) defines the feasibility boundary for voluntary disclosure [arXiv:2606.02348v1](#). The “sharing-feasible region” is explicitly a function of how privacy design interacts with the external information environment — a two-dimensional architectural space, not a single privacy dial. The analogy across these domains is structural, not formal: the formalisms differ, and the claim that they share a mechanism is the heuristic hypothesis under examination.

### 2. Identity Continuity as Infrastructure for Accountability

Reputation mechanisms are a canonical solution to adverse selection and moral hazard in markets with anonymous or pseudonymous participants. The theoretical literature on reputation assumes a bundle of properties: persistent identity, behavioural continuity, sanction sensitivity, and costly non-fungibility of reputation capital. The argument that language-model agents are “ontologically dissociative” — assemblages of mutable modules whose behavioural identity can change with any component update — is a claim that this infrastructure may be absent, not merely weak [arXiv:2605.30169v1](#). The analogy to dissociative identity disorder jurisprudence is explicitly evocative rather than constitutive; the paper uses it to illuminate governance challenges, not to establish a formal equivalence. The structural claim is more precise: if identity is not persistent across behavioural episodes, then past behaviour cannot straightforwardly ground predictions about future behaviour, and sanctions may not be internalised. The paper itself acknowledges that architectural solutions — such as persistent memory modules with cryptographic identity anchors — might partially restore the required properties, and this conditionality should be preserved in any policy inference.

This connects to the broader heuristic theme because identity continuity is itself an information-architectural property: it determines whether the correlation between past and future behaviour is observable and exploitable by counterparties. When identity is dissociative, that correlation may collapse, and with it the information on which reputation-based trust is built.

### 3. Threshold Dynamics and Re-Entrant Phases in Social Systems

Two corpus findings — the hate-content spreading model [arXiv:2605.21129v1](#) and the spatial-politics model [arXiv:2605.09784v1](#) — introduce non-monotone, threshold-governed dynamics that resist simple policy interventions. In the hate-content model, system-wide spreading is governed by re-entrant threshold phases: as the fraction of hate communities varies, the system can transition from spreading to non-spreading and *back to spreading*, meaning that partial moderation can move the system into a more dangerous regime before additional moderation moves it out [arXiv:2605.21129v1](#). In the political model, multiparty systems may induce polarisation even among moderate electorates, while two-party systems may reduce polarisation even among polarised voters — a result that, if it generalises beyond the model’s stylisations, would contradict the intuition that more political options improve representational quality [arXiv:2605.09784v1](#). Both findings illustrate that the relationship between architectural parameters (moderation intensity, party number) and social outcomes may be non-monotone, making naive intervention design potentially counter-productive. The connection between party number and “information architecture” in the sense used elsewhere in this paper is more attenuated than the other cross-domain bridges, and is addressed explicitly in Addendum A.

---

## Synthesis

### 2.1 Opacity as Welfare Technology: When Less Visibility Dominates

**Claim:** Under conditions of moderate adverse selection and bounded arrival rates, reducing the observability of order flow can *increase* aggregate welfare by eliminating strategic timing externalities, even though it reduces informational transparency.

**Evidence:** The game-theoretic framework in [arXiv:2605.31072v1](#) formally proves a welfare ranking  $W^{DARK} > W^{LIT} > W^{BATCH}$  under the stated conditions. The mechanism is that transparent order books create strategic timing games in which traders delay or rush submissions to optimise queue position, generating wasteful social waiting costs. Opaque order books eliminate these games through information design — not by improving the quality of price signals but by removing the signal on which the strategic game is played. Extensions incorporating asymmetric information and endogenous venue choice are noted in the preprint but are not fully characterised in the available abstract, and conclusions from those extensions should not be assumed to follow from the main result.

**Caveats:** The welfare ranking is derived under “moderate arrival rates and bounded adverse selection.” Under high adverse selection — where informed traders systematically exploit uninformed ones in dark pools — the ranking may reverse. The model is a stylised queuing framework; empirical validation against actual dark pool and lit exchange data would require a natural experiment or structural estimation that the preprint does not provide. This is a preprint that has not completed peer review, and the formal results have not been independently verified.

**Falsification path:** If data from markets with exogenous, policy-driven variation in dark pool access show welfare improvements for retail participants after dark pools are restricted — and if the magnitude of those improvements is inconsistent with the model’s predicted cost of strategic timing games — the welfare ranking claim would be weakened. A more targeted test: if the strategic timing games in the model are quantitatively negligible relative to adverse selection costs in calibrated real-market data, the dominance result would not generalise. The specific prediction

is that restricting dark pool access should *increase* measurable strategic timing costs in lit markets, a quantity that can in principle be estimated from order-book timestamp data.

## 2.2 Correlated Noise as a Lock-In Mechanism: The Production-Comprehension Asymmetry

**Claim:** The distributional harm of noise in social information systems depends critically on whether noise is correlated across receivers (production noise) or independent (comprehension noise). Correlated noise produces collective error-lock-in at rates that independent noise does not, through the same social influence weights operating on more correlated inputs.

**Evidence:** In the 600-participant online experiment reported in [arXiv:2605.30522v1](#), production-noise groups spent significantly more rounds tightly clustered around a wrong value than comprehension-noise groups ( $p = 0.016$ , group-level permutation). Production noise more often created a wrong common signal ( $p = 0.025$ , Fisher’s exact test) and made that signal persist across more rounds ( $p = 0.004$ , permutation). Dynamic update models confirmed that the mechanism was not stronger peer influence in production-noise conditions, but the same peer influence acting on more correlated perturbations.

**Caveats:** The experiment used a specific task (room temperature estimation with a unitless 50–250 range that created task-induced conflict with everyday temperature expectations), four-person groups, and 25 rounds. The effect sizes are group-level, not individual-level, and the sample is 600 participants in an online setting. Generalisability to larger groups, longer time horizons, or domains without the specific task-induced conflict is untested. The abstract mentions a GPT-agent experiment described as clarifying a mechanism, but this component is not fully reported in the available abstract and its results should not be assumed to replicate the human-participant findings. This is a preprint that has not completed peer review.

**Falsification path:** If experiments with larger groups or longer horizons show that comprehension noise eventually produces comparable lock-in rates (because independent errors average out more slowly in larger networks), the production-comprehension asymmetry would be attenuated. A more targeted test: if the task-induced conflict (unitless scale vs. everyday temperature expectations) is the primary driver of wrong-value lock-in rather than noise correlation, then removing the conflict should eliminate the differential — specifically, production-noise and comprehension-noise groups should show comparable lock-in rates on a task without the conflict. This is a directly testable manipulation that would distinguish the noise-correlation mechanism from a task-artefact explanation.

## 2.3 Algorithmic Monoculture as Manufactured Correlation: Racial Disparities and the Homogeneity Premium

**Claim:** When a single algorithmic vendor screens applicants across many employers, the independence of screening decisions is eliminated, converting what should be independent employer judgments into a single correlated architectural verdict. This manufactured correlation may amplify racial disparities and reduce the value of broad job-search strategies for disadvantaged applicants.

**Evidence:** In the dataset of 3 million applicants and 4 million applications analysed in [arXiv:2605.27371v1](#), 14.74% of applications by Asian applicants and 25.87% of applications by Black applicants were submitted to positions that adversely impact those groups under U.S. employment discrimination standards. At the individual level, 4% of applicants who applied to 10



positions were recommended for rejection from all 10, a rate higher than expected by chance. The deterministic replicability of the hiring algorithms allowed the authors to simulate counterfactual outcomes — what applicants would have received if they had applied to all positions — revealing that broad application is necessary but not sufficient to escape systematic rejection.

**Caveats:** The dataset covers applications screened by a single vendor, which is precisely the monoculture condition being studied; the findings do not establish what outcomes would look like under vendor diversity. The counterfactual analysis depends on the assumption that the algorithm’s outputs are deterministic and that the same algorithm was applied consistently across positions — assumptions that may not hold in deployments with version updates or customisation. The adverse impact findings use U.S. employment discrimination standards (the four-fifths rule), which is a legal threshold, not a causal estimate of discrimination. Causal identification would require a randomised audit or a natural experiment with variation in vendor assignment. This is a preprint that has not completed peer review.

**Falsification path:** If employers using different vendors show comparable racial disparity rates, the monoculture mechanism would be weakened relative to a story about shared training data or shared labour market conditions. The specific test: a dataset spanning multiple vendors should show lower within-applicant homogeneity of outcomes for applicants whose applications are distributed across vendors than for applicants whose applications are concentrated in a single vendor’s portfolio. If that difference is not observed, the correlation-architecture claim would be undermined.

## 2.4 Privacy Design and External Signals: The Complementarity Failure in Oligopoly Information Sharing

**Claim:** Privacy-preserving mechanisms alone are insufficient to induce voluntary information sharing in oligopoly markets. Privacy protection must be complemented by a sufficiently informative external signal, and the required privacy level is increasing in the accuracy of a firm’s private signal — meaning that the firms with the most valuable information require the strongest protection to share it.

**Evidence:** In the Cournot oligopoly model with uncertain demand analysed in [arXiv:2606.02348v1](#), a two-firm market without an external signal produces no sharing regardless of the privacy level. In an  $n$ -firm market, sharing may arise even without privacy safeguards because non-participating firms lose access to the aggregated signal — an exclusion threat. The central result is that privacy protection alone is insufficient; it must be combined with a sufficiently informative external signal. Firms with more accurate private signals require stronger privacy protection, characterising a “sharing-feasible region” that depends on both the privacy design and the external information environment.

**Caveats:** The model is a Cournot oligopoly with a specific demand uncertainty structure. Cournot competition is a benchmark, not a description of any particular market. The external signal is treated as exogenous; in practice, platforms that aggregate information may have endogenous incentives to shape the external signal to expand the sharing-feasible region, which is a conflict of interest not modelled here. The privacy mechanism is abstract (a channel that adds noise to aggregated signals); real privacy-preserving technologies (differential privacy, secure multi-party computation) have specific cost and accuracy profiles that may not match the model’s parameterisation. Policy inferences drawn from this model — for example, that regulators who mandate privacy without investing in external information infrastructure will reduce sharing without improving welfare — are plausible given the model’s logic but go beyond what the preprint formally establishes, and

should be treated as candidate implications rather than demonstrated results. This is a preprint that has not completed peer review.

**Falsification path:** If empirical studies of information-sharing platforms in oligopoly markets (e.g., industry data consortia) show that platforms with stronger external signals attract more participation, the complementarity claim would be supported. The specific prediction: platforms that improve external signal quality without changing privacy guarantees should show measurable increases in voluntary participation rates, while platforms that improve privacy guarantees without improving external signal quality should not. If platforms with strong privacy guarantees but weak external signals attract participation at rates comparable to those with both strong privacy and strong external signals, the necessity of the external signal would be refuted.

## 2.5 Identity Dissociativity and the Governance Vacuum: Reputation Mechanisms and the Agentic Trust Gap

**Claim:** The extension of human identity-verification and reputation mechanisms to language-model agents may be structurally mismatched to the agents’ ontological properties. Reputation mechanisms require persistent identity, behavioural continuity, sanction sensitivity, and costly non-fungibility — properties that modular, mutable LLM agents may not possess in their current form. This suggests a governance gap that may not be closable by ex-post, sanction-based regulation alone, though architectural solutions may partially address it.

**Evidence:** The theoretical analysis in [arXiv:2605.30169v1](#) identifies four properties that reputation mechanisms require and argues that LLM agents may fail each: identifiability (agents are assemblages of mutable modules — foundational models, system prompts, tool-access policies, external memory — any of which may change behaviour); predictability (behavioural continuity is not guaranteed across component updates); credibility (the fluid persona is vulnerable to adversarial attack); and rehabilitability (sanctions may not be internalised by an entity without persistent identity). The paper draws on dissociative identity disorder jurisprudence as an *evocative analogy*, not a formal equivalence, for the legal and governance challenges of assigning responsibility to entities with non-persistent identity. The paper itself acknowledges that architectural solutions — such as persistent memory modules with cryptographic identity anchors — might partially restore the required properties, and this conditionality is integral to the paper’s argument.

**Caveats:** The argument is theoretical and conceptual; it does not provide empirical evidence that reputation mechanisms have failed for LLM agents in practice (the deployment of such agents at scale is recent). The analogy to DID jurisprudence is illuminating but not constitutive — legal systems have developed partial solutions for DID cases that might inform agent governance. The claim that agents “may not internalise sanctions” is a structural argument conditional on current architectural norms; it is possible that architectural choices could partially restore the required properties, as the paper itself notes. This is a preprint that has not completed peer review.

**Falsification path:** If cryptographic identity schemes or persistent behavioural logging for LLM agents demonstrably produce stable behavioural continuity across component updates — operationalised as a statistically significant correlation between pre- and post-update behavioural profiles on a standardised evaluation suite — and if market participants demonstrably update trust based on this continuity at rates comparable to human-market reputation dynamics, the structural mismatch claim would need revision. The specific empirical test: compare adverse-outcome rates in agentic markets with and without cryptographic identity anchors, controlling for task complexity and deployment context.



## 2.6 Re-Entrant Phases and the Non-Monotonicity of Architectural Intervention

**Claim:** In systems governed by coalescence-fragmentation dynamics (hate-content communities), the relationship between the intensity of an architectural intervention and the desired social outcome may be non-monotone. Partial intervention can move the system into a more dangerous regime before additional intervention restores improvement, implying that intervention thresholds must be characterised before deployment.

**Evidence:** The two-species coalescence-fragmentation model with SIR dynamics in [arXiv:2605.21129v1](#) derives re-entrant threshold phases: as the fraction of hate communities varies, the system transitions from spreading to non-spreading and back to spreading. Both numerical solutions and two levels of mean-field approximation (EMT and BEMT) confirm this non-monotone behaviour.

**Caveats:** The hate-content model incorporates published empirical features of online hate communities but is a mathematical model, not an empirical study. The re-entrant phases are derived analytically; whether real platform moderation dynamics match the model’s coalescence-fragmentation structure is an open empirical question. This is a preprint that has not completed peer review.

**Falsification path:** If longitudinal data on platform moderation interventions — specifically, platforms that have implemented graduated increases in moderation intensity — show monotone reductions in hate-content spread as moderation intensity increases, the re-entrant phase prediction would be weakened. The specific prediction: there should exist an intermediate moderation intensity level at which hate-content spread is *higher* than at both lower and higher intensity levels. This is a testable prediction that could be evaluated against platform moderation logs if those data were made available to researchers.

---

## Addendum A: Weakly-Connected Source — Spatial-Politics Model

The spatial-politics model [arXiv:2605.09784v1](#) is retained in this corpus because its non-monotonicity result — that multiparty systems may induce polarisation even among moderate electorates, while two-party systems may reduce polarisation even among polarised voters — is structurally analogous to the re-entrant phase result in §2.6. However, its connection to the core heuristic of “information architecture as a distributional instrument” is more attenuated than the other corpus findings, and this should be stated explicitly.

**Why the connection is weaker:** In the other corpus findings, the architectural parameter (opacity, noise correlation, vendor concentration, privacy design, identity continuity, moderation threshold) directly governs the correlation structure of information flows to downstream actors. In the spatial-politics model, the architectural parameter is the *number of parties* in a political system. Party number is an institutional parameter, not an information-architectural one in the same mechanistic sense. The connection to information architecture would require an argument that party number governs the correlation structure of policy signals received by voters — for example, that more parties produce more correlated policy positions through competitive convergence, or that fewer parties reduce the dimensionality of the policy signal space. The preprint does not make this argument, and this synthesis does not supply it.

**What the model does establish:** The model is an extension of Hotelling-Downs spatial competition. It inherits that model’s stylisations: rational vote-maximising parties, one-dimensional spatial policy representation, and no strategic entry, campaign finance, or institutional constraints.

Within those stylisations, the model shows that the relationship between party number and representational quality is non-monotone. This result is consistent with the broader heuristic theme of non-monotone intervention dynamics, but the mechanism is distinct from the information-architectural mechanisms in the main synthesis.

**Falsification path for the political model:** If political systems that transition from two-party to multiparty structures (e.g., via electoral reform introducing proportional representation) show monotone improvements in a pre-specified measure of representational quality — such as the average distance between the median voter’s ideal point and the implemented policy — the non-monotone polarisation result would not generalise beyond the model’s stylisations. The measure of representational quality must be specified in advance to avoid post-hoc selection.

---

## Discussion

### What This Synthesis Suggests

The cross-domain heuristic pattern identified here suggests — as a candidate reading, not a demonstrated result — that **information architecture — the correlation structure, opacity profile, identity continuity, and noise routing of an information system — may function as a primary distributional instrument**. This is a stronger claim than the standard observation that information asymmetries create rents, and it should be treated with corresponding scepticism until the falsification paths outlined here are tested. The standard claim is about *who has information*; this synthesis is about *how the correlation structure of information flows may determine who bears coordination costs, lock-in costs, and governance gaps*.

The policy suggestion — not implication — is that regulatory attention might usefully focus on architectural choices upstream of information content. In market microstructure, the relevant question may not be only what information is disclosed but how the timing and correlation of disclosures interact with strategic behaviour [arXiv:2605.31072v1](#). In algorithmic hiring, the relevant question may not be only whether individual algorithms are biased but whether vendor concentration converts independent decisions into correlated ones [arXiv:2605.27371v1](#). In platform governance, the relevant question may not be only what content is moderated but whether partial moderation moves the system through a re-entrant phase into a more dangerous regime [arXiv:2605.21129v1](#).

The complementarity result from oligopoly information sharing [arXiv:2606.02348v1](#) provides a specific design hypothesis: privacy protection and external signal quality may be complements, not substitutes. Regulators who mandate privacy without investing in external information infrastructure may find that the mandate reduces sharing without improving welfare — but this inference goes beyond what the model formally establishes and should be treated as a candidate implication pending empirical test.

The identity-continuity argument [arXiv:2605.30169v1](#) suggests that governance frameworks designed for human market participants may be structurally mismatched to LLM agents in their current form. The paper itself acknowledges that architectural solutions might partially address this mismatch, and policy inference should preserve that conditionality. The governance challenge may be harder than improving KYC procedures — it may require either new governance primitives or a shift from ex-post sanction-based regulation to ex-ante architectural constraints — but this is a candidate reading of the theoretical argument, not a demonstrated policy conclusion.

## What This Synthesis Does NOT Suggest

This synthesis does not imply that opacity is always welfare-improving, that noise is always harmless, or that algorithmic screening is inherently racially biased. The welfare superiority of dark pools is conditional on specific market conditions [arXiv:2605.31072v1](#); the harm of production noise is specific to the correlation architecture of the noise, not noise per se [arXiv:2605.30522v1](#); and the racial disparities in algorithmic hiring reflect the monoculture condition, not a general property of algorithmic screening [arXiv:2605.27371v1](#).

The synthesis also does not imply that the findings are causally identified. Most of the corpus findings are theoretical models or observational/experimental studies with specific scope conditions. The cross-domain pattern is a heuristic hypothesis about a shared mechanism, not a demonstrated causal law. The falsification paths outlined in the Synthesis section are genuine — the pattern could be disrupted by natural experiments or randomised designs that independently vary architectural parameters while holding information content constant.

Finally, the synthesis does not imply that the parties who control information architecture are acting in bad faith. The welfare gains from dark pools accrue to traders broadly, not just to dark pool operators [arXiv:2605.31072v1](#). The privacy design in oligopoly markets is studied as a mechanism for enabling welfare-improving information sharing, not for extracting rents [arXiv:2606.02348v1](#). The pattern of architectural control over welfare distribution does not require malicious intent — it is a structural feature of how information systems are built and governed.

---

## Limitations

**Preprint status and non-peer-reviewed evidence.** All findings are from arXiv preprints and have not completed peer review. The theoretical models have not been independently verified; the empirical findings have not been replicated. The synthesis treats the abstracts as accurate representations of the papers’ claims, which may not survive full review.

**Heuristic, not formal, synthesis.** The cross-domain pattern is a heuristic reading across heterogeneous formalisms, as stated in the Introduction. The papers use different mathematical structures and were produced in different research communities. The claim that they share a mechanism is a hypothesis, and the mechanism is inferred from structural analogies rather than from a common theoretical framework or empirical test.

**Selection on corpus availability.** The corpus covers a 30-day window across five arXiv category clusters, assembled by keyword search (see Selection Process). The cross-domain pattern identified here may reflect the specific papers available and the search terms used rather than a robust structural regularity. Papers that contradict the pattern may exist in the literature and are not represented here.

**Small samples and specific populations.** The collective cognition experiment used 600 participants in four-person groups on a specific estimation task [arXiv:2605.30522v1](#). The hiring dataset covers applications to a single vendor’s algorithm [arXiv:2605.27371v1](#). The activist security study involved 13 interviews in one city [arXiv:2605.29621v1](#). Generalisability of any individual finding is uncertain.

**Model stylisations.** The market microstructure model [arXiv:2605.31072v1](#) and the oligopoly information-sharing model [arXiv:2606.02348v1](#) are game-theoretic benchmarks with specific

functional form assumptions. The hate-content model [arXiv:2605.21129v1](#) uses coalescence-fragmentation dynamics that may not match real platform moderation processes. The political model [arXiv:2605.09784v1](#) uses one-dimensional spatial representation. Real systems are more complex, and welfare rankings derived in stylised models may not survive in empirical settings.

**Absence of effect-size comparability.** The synthesis identifies a common structural pattern but cannot compare effect sizes across domains. The magnitude of welfare gains from dark pools, the size of the production-comprehension asymmetry, and the racial disparity rates in algorithmic hiring are domain-specific quantities measured in incommensurable units. The synthesis is qualitative, not quantitative.

---

## Conclusion

Across market microstructure, collective cognition, algorithmic labour markets, oligopoly information sharing, agentic governance, and online content dynamics, a common structural pattern emerges from the corpus — as a heuristic reading, not a formal derivation: the party that controls the correlation architecture of an information system — its opacity, its noise routing, its identity continuity, its moderation thresholds — may capture systematic welfare advantages, while downstream participants who cannot observe or model that architecture bear the costs of coordination failures, lock-in, disparate screening, governance gaps, and non-monotone intervention effects. This is not a claim about intent or market power in the conventional sense; it is a candidate hypothesis about the structural primacy of second-order information design over first-order information content. The hypothesis is falsifiable through natural experiments that independently vary architectural parameters while holding content constant, and the falsification paths outlined here are genuine and specific. If the pattern survives such tests, the policy suggestion is that regulatory frameworks focusing on information content while leaving correlation architecture unexamined may systematically fail to address the distributional harms they are designed to prevent — but this suggestion is conditional on the empirical tests, not prior to them.

---

## Response to Review

**Heuristic framing.** The most consequential structural change in v2 is the explicit labelling of the synthesis as a *heuristic reading* rather than a derivation from a shared formal structure. This label now appears in the Abstract, the Introduction (as a dedicated methodological note), the Background, the Discussion, and the Conclusion. The original paper’s language — “independently instantiate a common mechanism,” “structural logic” — implied a stronger epistemic status than the heterogeneous formalisms warrant. V2 preserves the substantive argument while accurately characterising its epistemic standing.

**Selection process.** A new “Selection Process” section documents the arXiv category clusters searched, the keyword terms used, the inclusion and exclusion criteria applied, and the rationale for retaining each paper. This section also acknowledges the selection pressure introduced by keyword-driven search and the 30-day window, and notes that contradicting papers may exist in the literature.

**Weakest-link source.** The spatial-politics model [arXiv:2605.09784v1](#) has been moved from the main synthesis to a dedicated “Addendum A: Weakly-Connected Source” section. The addendum

explicitly argues why its connection to the core information-architecture mechanism is more attenuated than the other corpus findings — specifically, that party number is an institutional parameter rather than an information-architectural one in the mechanistic sense used elsewhere — and retains the paper’s falsification path. The citation is preserved and traceable.

**Dropped caveats and abstract overshoot.** Several caveats that the original abstracts preserve were restored or strengthened: the GPT-agent experiment in [arXiv:2605.30522v1](#) is now flagged as not fully reported in the available abstract; the conditionality in [arXiv:2605.30169v1](#) about architectural solutions partially restoring required properties is now preserved in both the Background and Synthesis sections; the policy inferences from [arXiv:2606.02348v1](#) are now explicitly labelled as “candidate implications” that go beyond what the model formally establishes; and all seven corpus papers are now explicitly flagged as preprints that have not completed peer review. The Discussion’s policy language has been downgraded from “implication” to “suggestion” and “candidate reading” throughout.

**Cross-domain analogies and falsification paths.** The asserted analogy between party number and information architecture is now explicitly argued (and found wanting) in Addendum A rather than asserted in the main text. Falsification paths in §2.3, §2.4, §2.5, and §2.6 have been sharpened to include specific, operationalisable predictions — for example, the specific prediction that there should exist an intermediate moderation intensity at which hate-content spread is higher than at both lower and higher intensities, and the specific operationalisation of behavioural continuity for LLM agents as a correlation between pre- and post-update behavioural profiles on a standardised evaluation suite.