Beyond Hurwicz

Incentive Compatibility under Informational Decentralization

David Lancashire

January 2026

Abstract

Achieving incentive compatibility under informational decentralization has long been considered an impossibility in economics and computer science. We show this is conditional by identifying a subset of non-revelation-equivalent mechanisms that infer enforcement preferences indirectly from parallel, uncorrelated games.

1 Introduction

If we read the *Churning of the Milk Ocean* as a Hindu creation myth, what surprises is not so much that a price must be paid for setting the world in motion, but that the cost cannot be borne by the gods acting in the field of creation. Before time allows its deeper treasures to emerge, external authority is needed — Shiva — to stabilize the system.

Hurwicz echoes this structural constraint in his foundational papers on implementation theory, arguing that whenever participants can mislead others about their preferences, incentive compatibility requires an external authority to punish deviations. Hurwicz saw his claim as a fundamental impossibility result, and concluded it made incentive compatibility impossible in all informationally decentralized systems lacking such authorities.

> These results show that the difficulty is due not to our lack of inventiveness, but to a fundamental conflict among such mechanism attributes

> as the optimality of equilibria, incentive-compatibility of the rules, and the requirements of informational decentralization.

Designers have accepted these arguments rather than confronting them directly. Where auctions have required protection against agents "changing their minds", designers have generally relied on courts, auctioneers, or institutions to verify and punish attempts at revision. Such external authorities do more than levy penalties: they determine the *numéraire* in which agents are penalized, forcing stabilizing punishments outside the scope of the mechanism as necessary and creating a cost-bearing constraint that agents cannot neutralize from within it.

There is, however, a narrow class of non-revelation-equivalent indirect mechanisms in which enforcement costs can be generated endogenously. In these designs, agents who attempt to shift equilibrium strategically must bear exposure to uncertainty over time, and this exposure cannot be neutralized from within the mechanism. These mechanisms lie just beyond the boundary identified by Hurwicz.

This paper explains how such a solution is possible. It starts with theory: why circular mechanisms are unstable, and how designers stabilize them by introducing enforcement costs that do not respond to strategic play. It then shows it is possible to make these constraints responsive to play without making deviation rational — through design techniques which force agents to pay front-loaded costs under uncertainty to propose changes that will adjust equilibrium security levels in the mechanism. This permits incentive compatibility to emerge in ways that are unimplementable in direct or revelation-equivalent mechanisms.

2 The Special Class of Circular Mechanisms

In the canonical definition provided by Hurwicz (1972), a mechanism is simply a function that maps actions or messages to outcomes. Whether a mechanism is incentive compatible depends on whether its outcomes implement a social choice rule as defined by Arrow (1951), and later developed by Gibbard and Satterthwaite (1973, 1975) and Maskin (1977).

It would take a decade before Myerson showed that under certain conditions indirect mechanisms can be reduced to direct ones without loss of generality (1981). The sheer power of his Revelation Principle would later induce many to see indirect mechanisms as essentially interchangeable with direct ones. But in the early 1970s, this suspicion had yet to take root, and Hurwicz framed his definitions to cover games in both classes. Nor did he presuppose any particular environment or specify how messages are communicated.

Rather, Hurwicz defined a mechanism as incentive compatible when, given the rules of the game, each agent finds it optimal to act in a way that implements the intended social choice rule. The circular mechanisms studied here fall squarely within this definition. They map actions into outcomes. What distinguishes them from other mechanisms is a structural property: they operate in rounds with the results of early stages reappearing as constraints on later play.

For clarity, circular mechanisms should not be confused with repeated games that extend over time. While circular mechanisms share similarities with such games, what makes them uniquely vulnerable is that their own enforcement mechanism—the constraints that impose costs on deviations—can be strategically manipulated within the game. Agents who learn how the game handles enforcement can deliberately weaken it to enable deviation at lower cost.

Such mechanisms appear across many domains: as continuous double auctions and adaptive market-clearing processes in economics; as reputation systems and agenda-controlled games in political science; and as gossip protocols, adaptive routing, and consensus mechanisms in computer science. Appendix A provides a representative taxonomy.

Existing research covers these games as well. Because their incentives are inherently path-dependent, players evaluate strategies over histories rather than in isolation (Abreu, Pearce & Stacchetti; Fudenberg & Levine; Fudenberg & Maskin). Players may also reassess earlier choices in a phenomenon known as *time inconsistency* (see Strotz; Pollack; Hicks; Kydland & Prescott), which is destabilizing not only because agents may want to revise plans, but because the mere possibility of revision may be anticipated (Hart & Tirole; Dewatripont; Barro & Gordon). Incentive compatibility can fail even if all parties regard the unrevised outcome as mutually beneficial (Farrell & Maskin, 1989; Maskin & Moore, 1988).

The problem is that *ex post revision* creates the same need for enforcement that Hurwicz identified in static settings: the need for a credible punishment to incentivize honesty. Hurwicz assumed such punishment was impossible in decentralized environments lacking institutions capable of observing and punishing inconsistency. His logic extends naturally to dynamic games with incentives for revision, making it surprising that circular mechanisms can eliminate this problem.

Yet the solution is already hinted at in a subset of theoretical papers. The most striking such result is James Jordan's discovery that when incentive compatibility is unattainable in one-shot mechanisms, it can re-emerge in path-dependent indirect mechanisms. Jordan does not argue that the solution requires circularity, but demonstrates that any solution must distribute costs over time. He also anticipates more recent work showing that direct mechanisms can fail to implement equilibria available to indirect ones (Aumann; Renou & Tomala; Strack & Mora; Attar), and that incentive compatibility can be restored once mechanisms can leverage expectations of future payoffs as part of their enforcement structure (Maskin & Moore; Abreu, Pearce & Stacchetti).

The remainder of this paper shows how a narrow subclass of mechanisms fulfills these criteria by introducing a dual-enforcement regime that allows the mechanism to optimize its own security function by observing it indirectly—in the revealed preference of players to coordinate beyond the scope of its enforcement guarantees.

3 Endogenous Unactionability

Sifting through these papers, we can see that where incentive compatibility survives informational decentralization, it is because the mechanism embeds a constraint that agents must bear but cannot neutralize. For the rest of this paper, we refer to this property as *endogenous unactionability.*

In mechanisms where external authorities punish deviation, endogenous unactionability emerges trivially, as enforcers have no utility function in the mechanism. Authorities must still be immune to strategic manipulation, but this is imposed by requiring them to verify any public states used to determine allocations or transfers. From this perspective, one of the reasons circular mechanisms appear counterintuitive is that opacity — not transparency — is what enables enforcement.

It may seem odd to argue that endogenous unactionability is required for incentive compatibility in decentralized networks. In computer science, many protocols exist which implement outcomes over time, and some accomplish this without adding trusted authorities. Are they not incentive compatible? The answer is no in the sense that each known solution only achieves convergence by fixing at least one cost-imposing dimension of the environment so that it does not respond to strategic pressure. This is different from the challenge posed by Hurwicz, who asked whether mechanisms could incentivize their own enforcement function, not assume it into existence beyond the game.

In games that converge through repeated play with Bayesian updating, known solutions require agents to update beliefs about a fixed state space with common priors and stationary likelihoods (Milgrom & Roberts; Fudenberg & Levine). Agreement can also be induced through public randomness or correlated devices, as in Myerson's *Great War* paper or Aumann's correlated equilibrium, but those approaches work because randomness is unintelligible and thus unactionable. Deviation is only costly where incorrect beliefs can induce suboptimal outcomes, so coordination unravels once agents can influence signal generation, timing, or interpretation.

This problem also characterizes mechanisms that rely on fixed type distributions, posted prices, or deterministic rules. Games that exploit Byzantine fault tolerance, posted-price structures, or fair schedulers are only stable when honesty, prices, and timing are strategically insulated: once those dimensions are endogenized, security and convergence collapse. Approaches based on regret minimization and multiplicative weights exhibit the same structure: they require stationary payoffs and non-strategic feedback, and cycling reappears as soon as agent actions reshape allocations.

Across cases, designers implement outcomes by finding load-bearing structures that resist strategic manipulation, not mechanisms in which the enforcement function is optimized through strategic play. Incentive compatibility requires the latter, but this seems absurd: how can costs be both endogenously generated and non-manipulable? Such opacity is clearly impossible in direct mechanisms, where implementability demands truthful preference revelation and the need for Maskin monotonicity further prevents agents from commingling multiple values into non-scalar reports and sharing their preferences obliquely. Any solutions must therefore lie outside the class of direct mechanisms or their revelation-equivalent indirect counterparts.

We shall see that the techniques which accomplish this do in fact break reduction under the Revelation Principle. But discussing how indirect mechanisms can preserve opacity requires a way to reason about how mechanisms can protect informational privacy, and so our next section introduces the concepts of *Myerson* and *Non-Myerson Layers* to show how constructs composed of the latter can sustain the kinds of compartmentalization which impede agents from knowing the true preferences of other agents in the same mechanism, even while aggregating their underlying preferences into a shared enforcement cost.

4 Myerson and Non-Myerson Layers

The Revelation Principle is a brilliant analytical reduction which demonstrates that many indirect mechanisms can be reduced to simplified games in which agents report types, the mechanism selects an outcome, and transfers enforce truthful revelation. Within its domain, the principle provides extraordinary clarity, allowing designers to reason about incentives without worrying about structural compartmentalization or the passage of time in a game.

Yet many non-revelation-equivalent games exist which cannot be reduced to direct mechanisms without breaking incentive compatibility. The techniques that block reduction are fairly easy to implement technically, including strategic routing (Renou & Tomala), some probabilistic lotteries (Strack & Mora), and asymmetric information disclosure (Attar).

These non-revelation-equivalent strategies become meaningful when mechanisms are treated as *constructs* composed of multiple layers. In such designs, each layer can serve a different function, yet participants may condition on inputs from previous layers and create outputs for subsequent ones. A boundary between layers exists at any point where inputs are strategically indistinguishable from randomness when analyzed in isolation.

For clarity, we call a layer a *Myerson Layer* if its incentive structure can be faithfully represented by a reduced game that acts on direct-type reports and has verifiable outputs. These layers follow the same rules as direct mechanisms and, when linked together in chained succession, are analyzed as if occurring simultaneously. The terminology is an analytic convenience indicating that a layer is intended to process information sequentially even though it can be represented through a static interface.

In contrast, a *Non-Myerson Layer* has the properties of non-revelation-equivalent indirect mechanisms, relying on forms of action within the mechanism that do not admit reduction without loss of essential structure.

The distinction matters for circular mechanisms. Any sequence of Myerson Layers may be collapsed into a single Myerson Layer under the Revelation Principle, but constructs that contain Non-Myerson Layers are irreducible because the incentive structure of at least one layer cannot be reduced to direct type reports.

This makes *constructs* useful conceptual tools for visualizing where requirements for credibility and enforcement emerge in mechanisms. In classical designs reliant on external authorities, enforcement is typically provided by institutions that are operationally indistinguishable from upstream layers that project randomness downstream where it appears as honest types or fixed constraints. This is shown in Figure **??**, which shows a *Non-Myerson Layer* preceding a *Myerson Layer*.

Such arrangements also affect the viability of trade-offs in downstream layers. A canonical example is budget balance in the Myerson–Satterthwaite setting, where adding a trusted auctioneer implicitly assumes an upstream layer to incentivize its provision. When this layer is removed in revelation-equivalent constructs, budget balance fails as the mechanism must internalize the cost

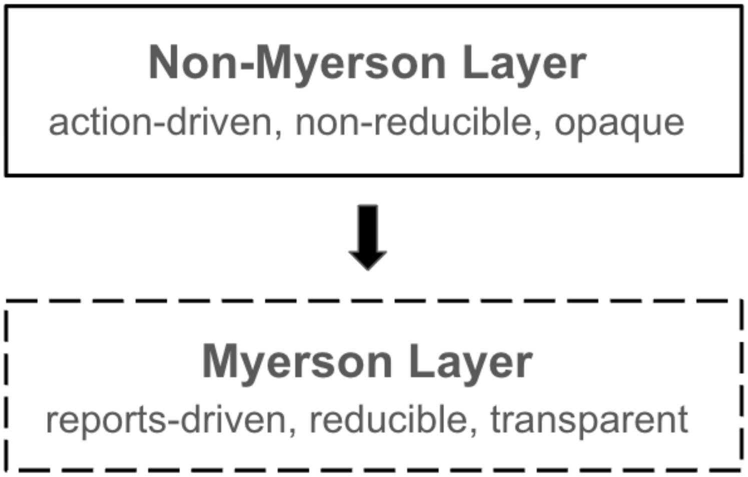
Non-Myerson Layer
action-driven, non-reducible, opaque
Myerson Layer
reports-driven, reducible, transparent

Figure 1. Non-Myerson Layer preceding Myerson Layer. The Myerson layer conditions on preferences expressed in the preceding layer which appear projected into it in forms indistinguishable from randomness.

of generating a behavior that was previously induced externally.

Constructs in which Myerson Layers project information into Non-Myerson Layers can be useful when the second layer introduces opacity, delay, or strategic aggregation to the preferences revealed in the first, transforming the original inputs into signals that emerge staggered over time. An example is sequential voting systems with portfolio assembly, where individual votes cast in the first layer may not be immediately visible in the outputs of the second. Blockchains with private mempools show how this design adds obfuscation and delay which frustrate attempts to reconstruct whose preferences are reflected in aggregated outcomes at any point in time. This is depicted in Figure **??**, which shows a *Myerson Layer* preceding a *Non-Myerson Layer*

These differences become decisive when layers are chained into circular constructs. Consider a construct consisting solely of Myerson Layers, which the Revelation Principle collapses into a single layer in which outputs from any round become transparent inputs to the next as shown in Figure **??**.

In contrast, circular mechanisms with multiple Non-Myerson Layers can preserve forms of informational compartmentalization. While each layer still acts on information projected into it from previous layers, its inputs need not be reducible to direct type reports and may consist of non-scalar values that preserve ambiguity of intent. This is shown in Figure **??**.

Across the lossy informational boundaries that separate layers in these con-

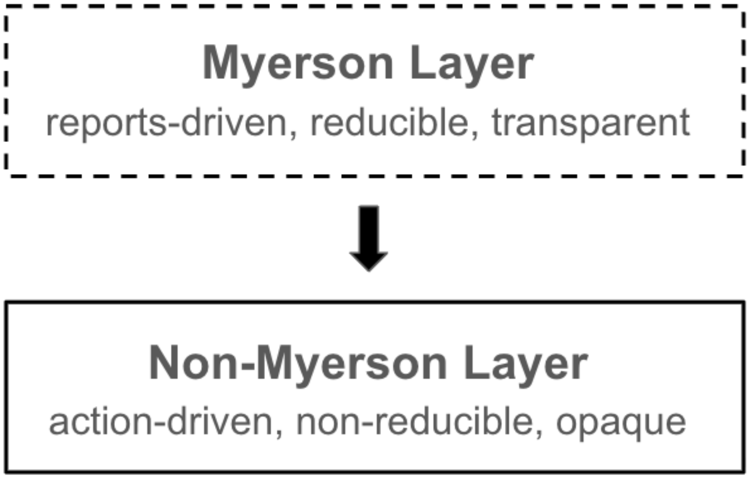
Myerson Layer
reports-driven, reducible, transparent
Non-Myerson Layer
action-driven, non-reducible, opaque

Figure 2. Myerson Layer preceding Non-Myerson Layer. The Non-Myerson layer acts on transparent inputs projected into it from the previous layer, and is able to reconstruct the intent behind the projected preferences.

structs, which we call *privacy walls*, agents can observe actions without reconstructing intent. They must, in effect, peer through a glass darkly. In the next section, we show how these privacy walls combine with non-scalar values to enable a new type of strategy that can exist only in indirect mechanisms.

5 From Privacy Walls to Selective Disclosure

Because direct mechanisms are evaluated in a static frame, the Revelation Principle forces any revelation-equivalent circular mechanism to operate sequentially, as anything else implicitly breaks reduction (Myerson, 1986). In non-revelation-equivalent mechanisms, however, multiple layers may operate in parallel, permitting a new class of strategies to emerge that cannot exist in revelation-equivalent designs. These are *selective disclosure strategies*, in which agents pierce privacy walls and reveal preferences directly to other participants, as depicted in Figure

Such disclosures exist in many auction designs. They are common in *off-book trading mechanisms*, where players must choose between submitting anonymous limit orders to a central book or negotiating privately with dealers “upstairs”. They also appear in FCC-style spectrum auctions, where players may tacitly disclose geographic priorities. In blockchains, similar disclosures

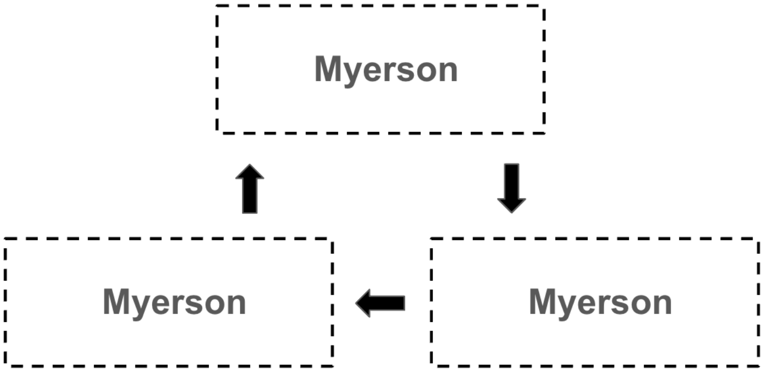

Figure 3. Direct Mechanism as a Construct of Myerson Layers. Multiple Myerson Layers chained into a circular construct form a Penrose-style mechanism which is locally coherent but globally incoherent.

occur whenever users sell transactions to miners in return for privately negotiated benefits.

As in other mechanisms, strategies involving the disclosure of preference maps create a two-sided risk. Agents who offer sincere interpretations expose themselves by making their internal structure legible, while agents who receive offers face the risk of strategic manipulation. A user who sells transactions to a miner may signal interest in a discount, but the true objective may be to lower overall security and make deviations cheaper within the mechanism. An informed counterparty may rationally decline such a trade.

Circular mechanisms can therefore invite players to pierce privacy walls and engage in disclosure games outside the mechanism's enforcement regime. The same non-scalar messages that preserve privacy also prevent either party from grounding claims of credibility in actions taken under enforcement, preventing the value of the enforcement function from interfering with its measurement of its own utility.

Circular mechanisms thus invite players to pierce privacy walls and play disclosure games outside the scope of the mechanism's enforcement function. What makes these games distinctive is that the non-scalar messages that create privacy also prevent either side from supporting claims about credibility by referencing actions ever taken under the mechanism's enforcement regime. The mechanism is therefore inducing an external game which allows it to observe, insulated from its own influence, the marginal utility of its own enforcement function.

To see the problem more clearly, consider a case where Alice asks Bob to take an action that lowers security in the mechanism. Alice may be attempting to

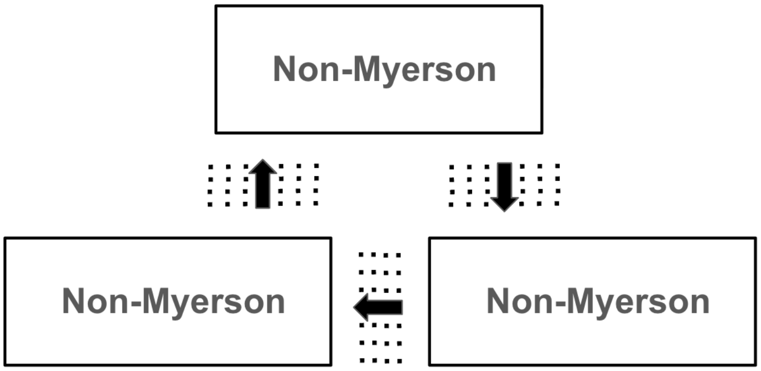

Figure 4. Indirect Mechanism as a Construct of Non-Myerson Layers. Multiple Non-Myerson Layers chained into circular constructs lose verifiability as privacy walls form between layers.

convert a global public good (security) into a dyadic trust good, capturing the savings while Bob absorbs the risk. Alternatively, her offer may be sincere, with Alice willing to bear exposure while offering Bob the benefits. Both interpretations are consistent with the same proposal.

Some may argue that the question is unanswerable, but Bob may have rational grounds to judge Alice's credibility. While the mechanism cannot observe these factors directly, it cannot deny their relevance if they lead Bob to support a higher or lower level of enforcement. If Bob already prefers lower security, he needs no persuasion. Alice's credibility becomes decisive only when Bob's preferred strategy depends on her sincerity. In that case, disbelief justifies maintaining a high-security equilibrium, while trust rationalizes lowering it.

From the perspective of the mechanism, Alice's action opens a Non-Myerson Layer that is opaque to it. This external game operates under a different enforcement regime and produces a non-scalar output encoding Bob's trust only in cases where that trust is decisive. The mechanism observes Bob's action but not the internal message space that motivates it. If Bob trusts Alice when enforcement could protect him from loss, the mechanism can infer that additional protection is not needed.

The mechanism is thus creating a counterfactual game it observes only indirectly, which signals the marginal utility of its own enforcement function back to it. This signal is credible precisely because Alice cannot cite any verifiable action taken under the mechanism's enforcement regime as evidence of her credibility. Were such a link possible, Bob could assign probabilities to Alice's sincerity and uncertainty would collapse into risk, correlating the two

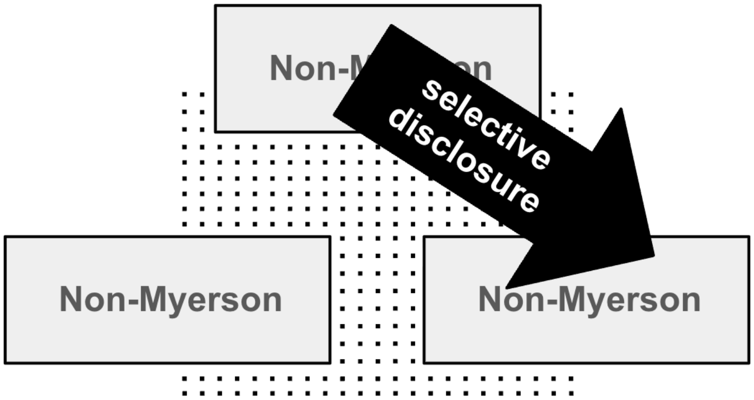

Figure 5. Selective Disclosure as Communications Strategy. Agents pierce privacy walls and communicate preferences directly through Non-Myerson meta-games the mechanism encourages to be created in private channels where speech is not tracked or penalized, but whose output is a structured input returned to the mechanism which drives the mechanism into equilibrium. This makes it a part of the underlying game, not the "null case" of player "exit" or "opt-out" as retreats to non-enforcement environments are traditionally modelled in implementation theory.

games and destroying the informational separation required for inference.

For readers accustomed to direct mechanisms, it may seem counterintuitive that a mechanism never observes the preferences that are decisive. But it does not need to. When enforcement is far from optimal, selective disclosure becomes attractive. As enforcement converges toward equilibrium, the risks of disclosure outweigh its benefits, credibility becomes less valuable, and disclosure strategies disappear. Credibility is thus decisive and unobservable in exactly the proportion required for it to substitute for the marginal utility of enforcement.

Selective disclosure strategies exist only in non-revelation-equivalent mechanisms. In the next section we show this formally, demonstrating that no statement Alice can offer Bob avoids collapsing into epistemic claims whose credibility cannot be established within the mechanism itself.

6 Epistemic Uncertainty from Unrepresentable Values

Throughout this paper, we use *uncertainty* in the sense articulated by Knight, Keynes, and G. L. S. Shackle: not as a variable that can be calculated or approximated probabilistically, but as *unknowability* — the absence of any representation that can be rationalized within an internally coherent evaluative framework. As Keynes famously wrote:

> The sense in which I am using the term is that in which the prospect of a European war is uncertain... or the price of copper and the rate of interest twenty years hence... About these matters there is no scientific basis on which to form any calculable probability whatever.

For the sake of convenience, we refer to whatever variable motivates agents to accept or reject proposals under uncertainty as *trust*. This does not involve any ontological claims. We do not attribute moral qualities to trust, nor deny that it may be the product of utilitarian calculation. We merely name a variable that motivates action in cases where enforcement levels are decisive at the margins. An agent acts on trust whenever they accept that statements offered by a counterparty may be false, yet prefer to act as if they are true. An agent who withholds trust prefers to assume they may be false and instead relies on the baseline guarantees enforced by the mechanism.

This approach contrasts with much modern work in computer science, which seeks to eliminate uncertainty by expanding the state space through additional parameters, beliefs, or utility dimensions. We do not dispute the usefulness of such methods, but deny that they can eliminate epistemic uncertainty.

The reason is that regardless of how one conceptualizes trust — whether as a psychological variable or as a non-scalar value that can be disaggregated and studied — representing it as a direct type collapses the information it is meant to convey. The problem is connected to the way security and trust become substitutes across uncorrelated games when agents have the option of lowering enforcement costs in one by extending trust in another.

6.1 Trust as an Unparameterizable Domain Space

In informationally decentralized systems such as blockchains, deviation must be discouraged without recourse to external enforcement. Generating security is therefore costly. Let $S \geq 0$ denote the level of security imposed by the mechanism, with cost function $K(S)$, where

$$K'(S) > 0.$$

Let τ denote the *trust environment*: an aggregate property capturing how willing participants are to refrain from deviation absent strong enforcement. Let $R(S, \tau)$ denote residual cheating risk, where

$$\frac{\partial R}{\partial S} < 0, \quad \text{and} \quad \left|\frac{\partial R}{\partial S}\right| \text{ decreases as } \tau \text{ increases.}$$

Trust and security are substitutes in this environment: trust reduces the marginal utility of additional enforcement. When Alice proposes that Bob cooperate with her to reduce security, she is implicitly suggesting that welfare can be jointly increased by reallocating resources from enforcement toward other forms of utility at the margin.

Welfare improvements require changing the socially optimal level of security, which solves

$$\min_{S\geq 0} K(S) + R(S,\tau),$$

with first-order condition

$$K'(S^*) = -\frac{\partial R(S^*,\tau)}{\partial S}.$$

It follows immediately that the optimal security level $S^*(\tau)$ is strictly decreasing in τ.

6.2 The Trust Parameter Cannot Be Elicited

To compute $S^*(\tau)$, the mechanism must know τ. In systems with external authorities, this information may be inferred institutionally. In informationally decentralized mechanisms, however, the mechanism would need to elicit τ as a type.

This is impossible.

A declaration of the form "my trust level is τ" cannot distinguish a genuinely high-trust agent from a low-trust agent strategically claiming higher trust in order to reduce penalties. The type and its anti-type generate identical messages.

6.3 Gödel-Type Collapse

While any informationally decentralized mechanism that must optimize its own enforcement levels requires τ, any attempt to represent it directly is strategically manipulable in exactly the dimension it is meant to measure. Encoding trust destroys its informational role. This is a Gödel-type phenomenon: certain economically necessary variables are unrepresentable within mechanisms because any attempt to encode them renders the encoding meaningless.

This collapse creates *fundamental uncertainty.* The mechanism requires τ to compute $S^*(\tau)$, but τ cannot be known ex ante, inferred from reports, or represented within the mechanism without contradiction. This is not epistemic ignorance resolvable through learning; it is structural. The mechanism cannot observe the counterfactual worlds in which different trust levels would have prevailed, nor price security optimally without exposing itself to over- or under-enforcement.

6.4 Uncertainty and Indirect Mechanisms

Indirect circular mechanisms are the only class capable of resolving this problem. They allow participants to play games in parallel under different enforcement regimes and treat the outcome as a counterfactual judgment on the value of enforcement itself.

Within each game, the other appears as a Non-Myerson Layer whose output is opaque and subject to misinterpretation. Each nevertheless serves as a venue in which rational judgment is expressed about the other. Although cognitive frameworks may differ across layers, actions remain credible within the enforcement regime of the circular mechanism.

In this way, indirect mechanisms convert an unrepresentable parameter into an observable gradient without collapsing uncertainty into a reportable type. The solution is unimplementable without Non-Myerson Layers that generate privacy walls, invite agents to bypass them, and treat the result as an opaque referendum on the allocation of resources within the mechanism itself.

It follows that any informationally decentralized mechanism that must optimize its own enforcement levels cannot be direct. What it must observe is a Gödel-type object: fleeting, disguised, decisive only when necessary, and structurally impossible to represent directly, yet capable of guiding action.

7 From Uncertainty to Incentive Compatibility

This paper began with a narrow problem: how incentive compatibility can be sustained in informationally decentralized environments. In the preceding sections we showed that the solution lies in the endogenous generation of costs that agents cannot neutralize within a mechanism, and that this is possible in action-based mechanisms with non-scalar messages and privacy walls.

Selective disclosure plays a decisive role in this process and yields its simple equilibrium logic. When enforcement is too strong, agents have incentives to cooperate to reduce unnecessary security costs. When enforcement is too weak, agents retreat to trust-free strategies and demand stronger guarantees, preferring strategies that provide greater protection against revision. Once enforcement is correctly calibrated, neither revealing nor concealing private structure can improve expected welfare, selective disclosure ceases to be profitable, and security remains stable in equilibrium.

Seen in this light, uncertainty is not an epiphenomenon but the functional substitute for external authority in decentralized systems. Incentive compatibility can emerge more generally as self-stabilizing security levels now provide a common *numéraire* against which other forms of utility can be expressed. Appendix D shows a practical example of a mechanism exhibiting this dynamic.

The solution may only be implementable within a narrow subclass of indirect mechanisms, but it exists. In the next section, we revisit the large body of impossibility results in computer science and economics which claim it cannot, showing that their conclusions exclude indirect mechanisms by assumption, often for reasons their authors do not recognize can be relaxed.

8 Explaining Impossibility

Across economics and computer science, impossibility results share a common premise: that all strategic interaction occurs under a single enforcement regime. The edifice of mechanism design incorporates this assumption: excluding the possibility that agents might rationally shift coordination into lower-security domains created by the mechanism itself.

This possibility was invisible in early implementation theory. Because Hurwicz was examining allocation environments whose setting was practically indistinguishable from the state of nature, he never considered that multiple enforcement regimes might be possible. His impossibility result is therefore correct for the class of mechanisms he analyzes, but does not extend to mechanisms that give agents the strategic freedom to coordinate under reduced enforcement guarantees.

Other impossibility results rest on the same assumption. Arrow's impossibility theorem, the Gibbard–Satterthwaite result, and Maskin's monotonicity condition all presuppose that preferences must be represented as semantically meaningful scalar reports, that communication is verifiable, and that deviation is disciplined by a symmetry of enforcement costs. These results do not assert that incentive compatibility is impossible in informational decentralization per se, but they reinforce the conclusion that strategic interaction must occur under a single enforcement regime.

Canonical results from computer science reinforce the same assumption. Developed in foundational papers by Lamport, Shostak, and Pease, and later formalized by Bracha and Toueg, results on the limits of adversarial tolerance assume that algorithms behave according to fixed rules and execute changes as *atomic steps*. These models implicitly treat actions as costless, uncertainty as absent, and assume deviation cannot be punished through exposure accumulated over time. Later results such as the FLP theorem operate in this same framework, in which non-termination and inconsistency arise because no mechanism exists to impose asymmetrical costs that diverge over time.

In more recent work, including the so-called TFM literature, we find further arguments discouraging attention from indirect mechanisms as a viable solution class, with some authors even asserting analysis can be entirely restricted to the study of direct mechanisms. Many of these results fail to generalize due to improper reduction under the Revelation Principle, often from changing a two-sided, multi-parameter and dynamic game into a one-sided, single-parameter, static auction that cannot even in theory implement the same set of outcomes. A contributing source of difficulty is that off-chain coordination in TFMs naturally invites action under multiple enforcement regimes, which is interpreted to suggest the presence of multiple direct mechanisms rather than circular constructs with interacting non-revelation-equivalent layers.

Related work inherits the same restrictions implicitly through modeling choices. Bayesian analyses assume that learning takes place over static distributions within a single enforcement framework. Posted price models presuppose endogenous unactionability under highly restrictive conditions. And the pervasive assumption that miners "faithfully implement" protocols is repeatedly used to avoid eliciting preferences from them, an omission that leads to recurring problems with interdependent valuations once they reenter the game as strategic players. These results generate impossibility not because their incentives are ill-posed, but because their frameworks were only designed to handle direct mechanisms, not identify indirect constructs with multiple interacting Non-Myerson Layers.

Philosophically, the category error that emerges is the inverse of that identified by Gilbert Ryle in *The Concept of Mind* (1949). In that book, which popularized the phrase "ghost in the machine," Ryle critiqued Descartes for assuming the existence of an immaterial spirit simply because the mind seems to have agency. The error in our time is reversed: analysts retreat to deterministic models because networks are made of machines. Impossibility reappears quietly because deterministic execution cannot accommodate semantic ambiguity.

This same pattern appears in work on the economic limits of permissionless consensus, where economists import constraints from computer science as fixed technical limits and then restrict possibility within those bounds. Some papers assume verifiability is needed for incentive compatibility when that is not the case, as with Groves–Ledyard or dynamic pivot mechanisms. Others argue that consensus mechanisms must be voting systems, treating verifiability and voting-like behavior as necessary primitives and excluding stabilizing forms of off-chain coordination by assumption.

In summary, there are no grounds justifying the presumption that consensus mechanisms must be revelation-equivalent or governed by a single enforcement regime. But this suggests a puzzle: if indirect mechanisms can in principle escape classical impossibility results, why has the literature repeatedly failed to converge on them? Why do attempts to relax the assumptions so often conclude that nothing fundamental changes?

9 Why This Is Hard to See: The Gravity Well of Direct Mechanisms

The answer seems to be that mechanisms governed by the Revelation Principle exert a strong cognitive pull. Once a model satisfies even a subset of its defining assumptions—costless communication, verifiable reports, fully specified state spaces—it is drawn back into what we call the *gravity well* of revelation, where mechanisms collapse into direct forms unless costs are imposed under uncertainty. This pattern appears repeatedly across the literature in a series of near-miss attempts to escape impossibility by relaxing dimensions of the standard framework one at a time.

Time. A large body of work introduces dynamics — rounds, epochs, repetition, or sequential interactions — yet preserves transparency, verifiability, and report-based reasoning. In these models, time functions as an index over observations rather than as a generator of uncertainty. Learning sharpens inference, beliefs converge, and temporal structure accelerates the collapse back into revelation-equivalent paralysis. Impossibility is rediscovered, not overturned.

Collusion. Many models enrich the strategy space by allowing agents to collude, coordinate, or form coalitions. This introduces high-dimensional preferences, but these cannot be used to implement social choice rules, as their existence generates productive ambiguity about agent motives and preferences for future payoffs. As a result, stabilizing forms of cooperation that increase welfare are removed alongside suboptimal collusion, and off-chain agreements are recast as behaviors to be detected, punished, and eliminated.

Risk. A prominent line of work attempts to impose costs by making attacks expensive through fees, slashing, or resource expenditure. When such costs can be priced ex ante, however, they enter agents' decision-making frameworks as risks over known states rather than as exposure to uncertainty over time. Deterrence then fails, as it is modeled through equilibrium selection rather than through forced exposure to front-loaded costs under Knightian uncertainty.

Permission. Many proposals reintroduce trusted committees, governance layers, or validator sets. These moves can succeed precisely because they smuggle in additional Myerson or Non-Myerson Layers that live outside the layer under analysis. Yet because these changes are framed as shifts in trust assumptions rather than informational structure, they are treated as orthogonal to the incentive problem instead of as alterations that will also collapse once subject to systemic rationality.

Stabilizers. Posted prices, fixed fee schedules, and deterministic inclusion rules appear to move away from the Revelation Principle by constraining behavior in other ways. In practice, they deepen the gravity well by making actions legible, deviations interpretable, and learning faster. They are steps toward direct mechanisms, not away from them.

Authority. The introduction of stakers, validators, and internal governance bodies shifts authority from the environment into the mechanism. Once authority becomes endogenous, however, control itself becomes a source of utility, and the mechanism is tasked with incentivizing its own enforcement, reintroducing the circular paradox diagnosed by impossibility results.

In retrospect, there is no smooth path from the direct world to the indirect one. The transition appears to require a discontinuous shift in perspective across multiple dimensions: reports versus actions, static versus dynamic mechanisms, scalar versus non-scalar messages, risk versus uncertainty, monotonicity versus non-monotonicity, verifiability versus privacy, and ex ante costs versus ex post benefits.

From the perspective of the direct mechanism, this leap is invisible. It requires abandoning too much analytic scaffolding to even conclude that the solutions are tractable. Approaches that rely on privacy appear incoherent in settings that demand verifiability, and some variables can only be made decisive by rendering them semantically meaningless. Matters are further complicated by the possibility that multiple cognitive models may coexist, pulling the concept of utility itself into question.

As such, classical impossibility results are neither surprising nor discredited. They remain faithful descriptions of what occurs inside the gravity well of the Revelation Principle. From the perspective of the indirect mechanism, however, the direct mechanism emerges as the degenerate case in which impossibility arises because uncertainty is resolved immediately, a single enforcement regime is imposed by fiat, and values that matter become semantically impossible to express.

10 Conclusion

This paper began from a narrow question about incentive compatibility in distributed consensus and arrived at a broader reconsideration of how cooperation can be sustained without external authorities, showing the conditions under which incentive compatibility is possible.

Seen from this perspective, the power and the limits of the Revelation Principle become clearer. The principle is extraordinarily effective within the realm it was designed to govern: static or acyclic environments with costless speech and fully specified state spaces. But in indirect circular mechanisms where multiple enforcement regimes exist, improper reduction can strip away degrees of freedom by obliterating strategic ambiguity. Much like a black hole that collapses structural coherence at its event horizon, reduction transforms meaningful information into semantic indecipherability. Nothing essential can be lost only if nothing opaque mattered to begin with.

The broader implications are worth noting. In his *General Theory* (1936), Keynes extended Knightian uncertainty into macroeconomics, arguing that liquidity traps persist not because agents miscalculate risk, but because stasis prevails when uncertainty looms over the value of potential investments. On a more modest level, what has been missing in mechanism design is a way out of this same trap: a way for uncertainty to discipline behavior in ways that motivate price discovery instead of punishing it. By forcing agents to reveal strategic preferences for playing different games, indirect mechanisms can accomplish this, transforming uncertainty into a cost that can be rationally borne, shared, and priced without ever being fully resolved.

We believe that mechanism design must expand its vocabulary to incorporate these techniques and concepts. In mechanisms with routing strategies, variable-time hashing lotteries, and looping Non-Myerson constructs, sacrificing verifiability is not a defect but the price paid to sustain the forms of informational privacy that permit the mechanism to see its own value reflected indirectly in the willingness of actors to risk loss in its absence.

The lesson returns us, in many ways, to our mythological stories of creation. Where stabilizing costs can be paid outside the system, stability can follow. Where they cannot, they must be borne within the system in a manner that renders them external and incomprehensible to its own rational framework.

A Terminology

This appendix collects and standardizes the key terms introduced in the paper. Its purpose is not to introduce new concepts, but to fix language so that ideas are referred to consistently and unambiguously throughout.

Mechanism. A mechanism is a process that specifies admissible actions, information structure, and outcome rules governing interactions between agents. Following Hurwicz, a mechanism maps actions or messages into outcomes.

Incentive compatibility. A mechanism is incentive compatible with respect to a social choice rule if, for each agent, acting according to the strategy prescribed by the mechanism is optimal given the strategies of others, so that the induced equilibrium implements the intended social choice rule.

Direct mechanism. A direct mechanism is one in which agents report types or preferences in a single interface, after which the mechanism selects an outcome and transfers enforce truthful revelation. Direct mechanisms are the canonical objects governed by the Revelation Principle.

Indirect mechanism. An indirect mechanism is one in which agents express preferences through actions taken within the mechanism rather than through direct type reports. Outcomes depend non-trivially on how actions are aggregated, sequenced, and interpreted.

Construct. A construct is a mechanism composed of multiple internally chained layers. Constructs, rather than individual layers, are the primary objects of interest in circular and self-referential mechanisms.

Layer. A layer is a subcomponent of a construct in which agents condition on inputs and project outputs into subsequent layers in ways that appear to be unactionable when layers are studied in isolation.

Myerson Layer. A Myerson Layer is a layer whose incentive structure can be faithfully represented as a direct mechanism acting on type reports and projecting its output as publicly observable randomness or state changes. Any sequential composition of Myerson Layers may be reduced to a single Myerson Layer under the Revelation Principle.

Non-Myerson Layer. A Non-Myerson Layer is a layer whose incentive structure is not revelation-equivalent. Such layers rely on action-based incentives, ambiguous signals, dynamic preference aggregation, delayed resolution, or strategic routing, and cannot be reduced without loss of essential structure.

Privacy wall. A privacy wall is an informational boundary created by Non-Myerson Layers across which actions and outcomes are observable, but underlying motivations cannot be uniquely inferred. Privacy walls preserve ambiguity of intent by permitting non-scalar variables to motivate action and prevent agents from conditioning strategies outside the mechanism on fully reconstructible causal explanations of behavior within it.

Circular mechanism. A circular mechanism is a mechanism whose outputs re-enter itself as inputs in future stages of play, so that agents' present actions shape the future strategic environment in which they and others must act. In circular mechanisms, learning and revision are endogenous, and incentive compatibility depends on how the mechanism disciplines attempts to revisit or unwind past decisions.

Risk. Risk refers to uncertainty over a fully specified and stable set of possible states, where the mapping from actions to outcomes is known and probabilities can be meaningfully assigned *ex ante.* Under risk, agents can evaluate strategies by computing expected utilities, price deviations in advance, and revise behavior through belief updating without incurring structural costs beyond those implied by known distributions.

Knightian uncertainty. Knightian uncertainty refers to uncertainty over states, variables, or aggregation outcomes that are not defined at the time action must be taken, and therefore cannot be assigned probabilities *ex ante.*

Unactionability. A constraint is unactionable if agents cannot condition their behavior so as to offset, unwind, or neutralize its effects within the mechanism, even when the constraint is observed or anticipated.

Endogenous unactionability. A constraint is endogenously unactionable if it is generated by strategic interaction within the mechanism and, once generated, cannot be neutralized, offset, or unwound by further strategic action within that same mechanism.

Informational decentralization. A mechanism is informationally decentralized if no agent or authority has access to the full state required to determine outcomes or enforce behavior at the time actions are taken. Information relevant to outcomes is dispersed across agents and realized only through interaction over time.

B Circular Mechanisms

This appendix provides a taxonomy of representative circular mechanisms. The purpose is not to be exhaustive, but to illustrate how known families of mechanisms relate to the core distinctions developed in the text.

C Saito as an Implemented Instance of Endogenous Unactionability

This appendix provides a constructive witness to the claims made in the paper. It does not attempt to prove correctness, security, or equilibrium properties. Rather, it shows how non-reducibility emerges in practice from the structural requirements for incentive compatibility identified in theory.

C.1 Layers, Aggregation, and Arbitrage

Saito Consensus consists of three Non-Myerson Layers cycling in parallel. Each layer aggregates preferences revealed by others into welfare-improving proposals that adjust global prices. Three cross-layer properties are primitive structural requirements for avoiding impossibility in theory:

- **Continuation value.** Benefits include options for continued play, which are actively traded between participants as part of the aggregation process, dividing transaction fees into mining and routing payouts.
- **Costly uncertainty.** All strategies require participants to commit to front-loaded and irreversible expenditures of utility that are fungible with continuation value (fees, routing rewards, or resource consumption) under uncertainty.
- **Informational decentralization.** No trusted authorities exist in any layer of the mechanism. Players act under uncertainty about the global state, which is dispersed and aggregated only through play over time.

These properties apply to each of the three layers in the mechanism, which are distinguished by their aggregation functions and the arbitrage opportunities they expose.

User Layer

- *Aggregates:* Heterogeneous bundles of utility into transactions.
- *Arbitrages:* Differences between local execution opportunities and global settlement prices, broadcast as transactions for processing and settlement by other layers.

Routing Layer

- *Aggregates:* Bundles of transactions into provisional blocks.

- *Arbitrages:* Differences between marginal routing rewards and the cost of providing ancillary collusion goods, producing efficient routing paths for payout division.

Chain Resolution Layer

- *Aggregates:* Bundles of blocks into chains with a cost of revision.
- *Arbitrages:* Discrepancies between current aggregate security preferences and expected future settlement conditions, broadcast as payouts that simultaneously propose adjustments to global prices.

Each layer acts both as an aggregator of upstream preferences and as a generator of downstream arbitrage opportunities. No layer determines final meaning in isolation. Convergence is emergent, and the cost of revision increases with block settlement depth in the game tree.

C.2 Trust, Uncertainty, and Exposure

Trust is a semantically meaningless variable. It is measured only indirectly, through the shifts it induces from reliance on mechanism-enforced collective settlement guarantees to coordination through private channels where counterparty credibility is subject to time inconsistency.

These shifts take the form of selective disclosure strategies that pierce the privacy walls of the mechanism. Where such strategies are rational, they improve public welfare by optimizing the enforcement function. Where they are irrational, they reduce the expected utility of the proposer.

User Layer

- Selective disclosure of preference maps to routers.

Routing Layer

- Selective disclosure of preference maps to users.
- Selective disclosure of routing competitiveness to peers.

Chain Resolution Layer

- Selective disclosure of willingness to bear future exposure.

All forms of selective disclosure become less rational as the marginal utility of environmental trust converges toward equilibrium with the marginal utility of the mechanism's enforcement function.

C.3 Non-Reducibility under the Revelation Principle

The mechanism is a three-layer construct composed entirely of Non-Myerson Layers, and cannot be reduced to a revelation-equivalent mechanism. Messages are shared through public channels only as non-scalar values. Multiple strategies block reduction.

User Layer

- Routing strategies (Renou & Tomala).
- Selective disclosure (Attar).

Routing Layer

- Routing strategies (Renou & Tomala).
- Selective disclosure (Attar).

Chain Resolution Layer

- Routing strategies (Renou & Tomala).
- Selective disclosure (Attar).
- Time-uncertain fixed-difficulty hashing lottery (Strack & Mora).

Across layers, selective disclosure introduces additional Non-Myerson Layers that observe mechanism outputs and produce non-scalar outputs reflecting the attractiveness of coordination outside the mechanism's enforcement guarantees.

No portion of the mechanism is reducible to a direct mechanism:

- No layer relies on truthful type reports.
- All layers operate concurrently rather than sequentially.
- Information flows are non-stationary and non-topologizable.
- Actions project ambiguous signals rather than verifiable messages.

- Some strategies skip layers through uncertain execution order.
- Learning cannot model semantically undefined variables.

Attempts to analyze any single layer in isolation assume away the exogenous structure provided by the construct as a whole, reintroducing verifiability only by destroying the opacity that allows the mechanism to observe rational responses to the marginal utility of enforcement.

Because all forms of value are fungible with continuation value, higher- dimensional optimization emerges over message spaces containing utilities that can be exchanged for continuation value. These trades aggregate into global prices, inducing convergence.

C.4 What This Appendix Does Not Claim

This appendix does not claim:

- that Saito is unique;
- that similar mechanisms are easy to design; or
- that the theory provides a general construction algorithm.

It makes a narrower and stronger observation: *a mechanism satisfying the theoretical constraints identified through implementation theory does exist*, and when other incentive-compatible mechanisms are discovered, they will exhibit the same structural properties.

References

[1] Hurwicz, L. (1972). On the design of mechanisms for resource allocation. *American Economic Review Papers & Proceedings*, 62(2), 1–30.

[2] Arrow, K. J. (1951). *Social Choice and Individual Values.* Wiley, New York.

[3] Gibbard, A. (1973). Manipulation of voting schemes. *Econometrica*, 41(4), 587–601.

[4] Satterthwaite, M. (1975). Strategy-proofness and Arrow's conditions. *Journal of Economic Theory*, 10(2), 187–217.

[5] Maskin, E. (1977). Nash equilibrium and welfare optimality. MIT Working Paper.

[6] Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, 6(1), 58–73.

[7] Myerson, R. B. (1986). Multistage games with communication. *Econometrica*, 54(2), 323–358.

[8] Myerson, R. B. (2019). Game theory and the First World War. *Journal of Economic Literature*, 57(4), 963–987.

[9] Jordan, J. S. (1982). The competitive allocation process is informationally efficient uniquely. *Journal of Economic Theory*, 28, 1–18.

[10] Jordan, J. S. (1982). A dynamic model of expectations equilibrium. *Journal of Economic Theory*, 28, 235–254.

[11] Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1), 67–96.

[12] Renou, L., & Tomala, T. (2012). Mechanism design and communication networks. *Theoretical Economics*, 7(2), 209–256.

[13] Renou, L., & Tomala, T. (2016). Communication and information design. *Economic Theory*, 61, 1–31.

[14] Strack, P., & Mora, R. (2025). Information without rents: Mechanism design without expected utility. Forthcoming.

[15] Attar, A. (2021). Private disclosures. SSRN Working Paper No. 3947186.

[16] Abreu, D., Pearce, D., & Stacchetti, E. (1990). Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5), 1041–1063.

[17] Fudenberg, D., & Levine, D. (1994). Efficiency and observability with long-run and short-run players. *Journal of Economic Theory*, 62(1), 103–135.

[18] Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3), 533–554.

[19] Farrell, J., & Maskin, E. (1989). Renegotiation in repeated games. *Games and Economic Behavior*, 1(4), 327–360.

[20] Maskin, E., & Moore, J. (1988). Implementation and renegotiation. *Review of Economic Studies*, 66(1), 39–56.

[21] Knight, F. H. (1921). *Risk, Uncertainty, and Profit.* Houghton Mifflin.

[22] Keynes, J. M. (1936). *The General Theory of Employment, Interest, and Money*. Macmillan.

[23] Shackle, G. L. S. (1949). *Expectation in Economics*. Cambridge University Press.

[24] Ryle, G. (1949). *The Concept of Mind*. Hutchinson.

[25] Lancashire, D., & Parris, B. (2023). Sybil-resistant routing and incentive compatibility. Working paper. Available at `https://github.com/SaitoTech/papers`.

[26] Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3), 382–401.

[27] Bracha, G., & Toueg, S. (1985). Asynchronous consensus and broadcast protocols. *Journal of the ACM*, 32(4), 824–840.

[28] Fischer, M., Lynch, N., & Paterson, M. (1985). Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2), 374–382.

[29] Roughgarden, T. (2021). Transaction fee mechanism design. *ACM SIGecom Exchanges*, 19(1), 1–6.

[30] Chung, H., & Shi, E. (2021). Foundations of transaction fee mechanism design. *Proceedings of the ACM Conference on Economics and Computation (EC).*

[31] Bahrani, M., Garimidi, P., & Roughgarden, T. (2023). Transaction fee mechanism design in a post-MEV world. *Proceedings of the ACM Conference on Economics and Computation (EC).*

[32] Chung, H. (2023). Maximizing miner revenue in transaction fee mechanism design. *Proceedings of the ACM Conference on Economics and Computation (EC).*

[33] Chung, H., Roughgarden, T., & Shi, E. (2024). Collusion-resilience in transaction fee mechanism design. *Proceedings of the ACM Conference on Economics and Computation (EC).*

[34] Gafnim, Y., & Yaish, A. (2024). Barriers to collusion-resistant transaction fee mechanisms. Working paper.

Mechanism / Class	**Direct or Indirect**	**Revelation-Equivalent or Irreducible**	**Location of Unactionability**
Static direct-revelation mechanisms (e.g. VCG, Myerson auctions)	Direct	Revelation-equivalent	Exogenous (verifiability, transfers, enforcement assumptions)
One-shot voting rules (Arrow, Gibbard–Satterthwaite settings)	Direct	Revelation-equivalent	Exogenous (fixed rules, no revision, costless speech)
Repeated direct mechanisms with Bayesian updating	Direct	Revelation-equivalent	Exogenous (fixed state space, common priors, stationary likelihoods)
Mechanisms with public randomization / correlating devices	Direct or Indirect	Revelation-equivalent	Exogenous (unmanipulable public randomness)
Posted-price mechanisms with externally set prices	Indirect	Revelation-equivalent	Exogenous (price formation outside the mechanism)
Continuous double auctions	Indirect	Typically irreducible	Mixed; often exogenous via institutiona rules, sometimes endogenous via marke impact
Reputation systems	Indirect	Irreducible	Endogenous (history-dependent payoffs difficult to unwind reputational effects)
Contract renegotiation mechanisms	Indirect	Irreducible	Mixed; exogenous when courts enforce endogenous when renegotiation cost persist
Prediction markets with endogenous liquidity	Indirect	Irreducible	Endogenous (liquidity and price impac respond to participation but are not fully controllable)
Gossip-based broadcast and coordination protocols	Indirect	Irreducible	Exogenous or mixed (scheduler assumptions, network topology)
Proof-of-Work consensus	Indirect	Irreducible	Endogenous (irreversible expenditure o hash power; revision requires additiona work)
Proof-of-Stake consensus with slashing	Indirect	Irreducible	Mixed; endogenous (slashing, delayec payouts) with possible exogenous governance
Saito Consensus	Indirect	Irreducible	Endogenous (continuation value securec by entropy)

Table 1. A taxonomy of circular mechanisms and the location of unactionability.