

Settlement Velocity, 3 Layered AI-Economy and Circuit Breakers

Human Policy Supremacy

The governing proposition

In GUDIYA's proposed three-layer AI Economy (Figure 1)

(<https://www.manhattanproject20.com/blogs/the-ai-economy-architecture-v3>),

- the fastest (machine) speed at which agents decide (Layer 3) must be reconciled with
- the API commerce speed at which digital services (Layer 2) and
- underlying resources (Layer 1) speed, typically human speed, can safely respond.

Faster reasoning and faster interfaces do not automatically justify faster commitments. Human policy must establish the conditions under which agentic intent becomes an executed transaction and a settled obligation.

Settlement velocity should therefore be a governed variable.

- In some circumstances, a bounded, deliberate delay can preserve stability by spreading concentrated demand over time.
- In others, delay can increase risk and interrupt essential activity.
- The engineering task is to determine the appropriate pace under explicit human authority, with evidence that the selected intervention improves the collective outcome.

This requires coordinated circuit breakers at the Agentic, Digital Services, and Resource layers. Each breaker should control a defined class of activity and provide a path to recovery. Together, they would allow GUDIYA to moderate the conversion of intent into commitments while keeping the effects of intervention visible across the economy.



Figure 1 : Gudiya's proposed 3 Layered AI-Economy Architecture

The three layers and their different clocks

The Agentic Layer forms objectives, compares opportunities, discovers counterparties, and proposes actions. The Digital Services Layer converts those proposals into account operations, orders, reservations, payment instructions, and other executable commitments. The Resource Layer supplies the liquidity, inventory, energy, labor, transport capacity, or other scarce means through which obligations are fulfilled.

These layers can operate at very different rates. An agent may revise a plan in milliseconds. A digital service may accept thousands of instructions per second. The relevant resource may take minutes, hours, or days to replenish or reallocate. Digitizing the instruction does not remove the constraints on its fulfillment.

Here, “Resource Layer” includes economic capacity as well as physical stock. In the banking example, the relevant resources are settlement liquidity and funding capacity. A digital account balance is a claim; the ability to honor the associated payment obligations is a separate system condition.

This chapter uses settlement velocity to mean the rate at which authorized commitments reach final discharge over a specified interval. In payments, that may be measured as value settled per unit time, supplemented by transaction counts and concentration. Other sectors require their own completion measures. Order acceptance, reservation, financial settlement, and physical delivery must remain distinguishable.

The agentic deposit migration scenario

On September 27, 2026, Apollo Chief Economist Torsten Sløk described a scenario in which personal AI assistants automatically move household cash toward higher-yielding accounts. Widespread adoption could weaken the inexpensive deposit funding on which some banks depend. His note describes a potential consequence of adoption; it does not report an observed agentic bank run. [1]

For GUDIYA, the scenario is useful because the transmission path is visible. Agents select cash destinations; banking and payment interfaces execute their instructions; institutions experience changes in funding and liquidity. Individually sensible decisions can become a concentrated collective flow.

The agents need not exchange messages. Similar objectives, rates, data, models, and triggers may produce correlated decisions. In 2024, Gary Gensler separately warned that shared upstream AI models or data sources could lead many financial actors toward similar actions and create systemic dependencies. [2]

Deposit migration does not by itself establish a crisis. A transfer can move deposits between banks, and institutions can adapt their rates or funding. The critical questions concern speed, concentration, available liquidity, and the time required to adjust. The same total transfer volume spread across a week may have different consequences from a synchronized burst.

When Faster Settlement Can Produce a Bank Run^[1]

Consider millions of financial AI agents independently moving household deposits toward accounts offering better returns. Each agent may be authorized, acting rationally, and faithfully serving its owner. Yet similar decisions executed almost simultaneously could produce concentrated withdrawals faster than affected banks can mobilize liquidity. No malicious intent or explicit collusion is necessary. Individually desirable optimization can become collectively destabilizing when execution outruns the system's capacity to absorb it.

This scenario illustrates the interdependence of GUDIYA's three-layered AI-Economy. The Agentic Layer generates transfer decisions; the Digital Services Layer admits and executes instructions; the Resource Layer supplies the liquidity needed to fulfill them. The speed achievable by the first two layers does not establish the speed sustainably supportable by the third. Under stressed conditions, maximum settlement velocity may therefore be an inappropriate objective.

GUDIYA's proposed response is policy-governed pacing: wider Grid observation informs coordinated limits on new transfer initiation and admission, while local controls enforce those limits rapidly. Where authorized, transparent queues and staged release could moderate concentrated outflows before execution. Such measures must preserve completed settlement finality, protect essential payments, and include explicit recovery conditions. Delay is not inherently stabilizing; poorly designed restrictions can intensify uncertainty and panic.

The governing objective is sustainable settlement velocity under human-defined policy. A deliberate pause before commitment can be a stability function when it gives the underlying resource system time to respond.

When delay can serve stability

A controlled admission queue could spread a burst of discretionary instructions, giving the affected systems time to adjust. Staged release, approved netting arrangements, or capacity-sensitive scheduling may help match the pace of execution to the availability of resources. Such mechanisms require domain-specific evidence and authority.

Payment engineering already recognizes a trade-off between speed and liquidity usage. BIS research describes liquidity-saving mechanisms that use queuing and netting, while Federal Reserve Bank of New York research examines the cost of delaying payments against the cost of obtaining liquidity. These precedents

establish that immediate release is not the only engineering objective. They do not establish that restrictions on household transfers would be an appropriate policy. [3, 4]

The candidate GUDIYA control point is explicit: an authorized instruction enters a governed admission process, is released into execution under the applicable rules, and then proceeds toward settlement. Its status must remain visible throughout. The system must distinguish an instruction waiting for admission from an accepted obligation awaiting completion.

Delay is an instrument, not the desired outcome in itself. The objective is safe and timely fulfillment. Pacing is justified only where its expected stability benefit exceeds its costs and where it is compatible with the governing rules.

Human Policy Supremacy

Human Policy Supremacy means that legitimate human institutions establish the objectives, priorities, rights, and limits within which GUDIYA exercises control. An actant's optimization target cannot override that policy. The Grid's own preference for a smoother workload cannot override it either.

In the proposed architecture, the Cognitive Utilities Commission and relevant sector authorities (<https://www.manhattanproject20.com/blogs/cognitive-utilities-commission>) would define the applicable mandate. Grid operators would translate that mandate into monitored operating envelopes. Existing financial authorities and infrastructure operators retain their established responsibilities unless a future authorized arrangement explicitly reallocates them.

The policy must determine which activities may be paced, who may invoke the intervention, its maximum duration, and the treatment of urgent obligations. It must specify fair allocation, notice, review, and recovery. An affected user should be able to distinguish a permitted delay from an operational failure or a rejected instruction.

Humans need not approve every transaction at machine speed. They can authorize bounded control laws in advance, require evidence and auditability, and retain the power to revise or withdraw the mandate. GUDIYA then performs continuous measurement and execution within that authority.

The purpose is not to preserve an incumbent's cheap funding at the expense of depositors. A policy must account for the interests of households, businesses, counterparties, and the wider economy. Stability cannot become an undefined justification for withholding access to resources.

The external reference defines the operating envelope

A household agent may know its balance, permissions, and expected return. It may not know how many other agents are moving simultaneously, which institutions face concentrated outflows, or how recipient obligations depend on incoming funds. Its local guardrails cannot supply those missing observations.

GUDIYA's proposed contribution is a reference spanning the three layers: agentic demand, digital execution, and resource conditions. The admissible rate should respond to measured capacity, existing commitments, uncertainty, and policy priorities. A single fixed speed limit would be inadequate across different resources and stress conditions.

The resulting envelope is also directional. Pressure can propagate downward when agents generate demands that exceed resource capacity, and upward when a resource shortage forces services and agents to revise their plans. Feedback must carry both the constraint and its reason so that agents do not repeatedly submit work that cannot safely proceed.

Delay can also destabilize

An intervention that relieves one institution can deprive another of an expected inflow. A growing queue can conceal unmet obligations, provoke repeated submissions, or make participants fear that they will be last to exit. Delaying a transfer cannot repair an underlying insolvency or create missing resources.

GUDIYA must therefore assess the effects of the intervention itself. Controls need bounded queues, deadline awareness, visibility into dependent obligations, and a recovery plan. If protection becomes uncertain, the Grid should expose that uncertainty and invoke the authorized fallback.

Settlement finality must remain clear. Prospective pacing should not imply that a completed settlement can be arbitrarily reversed. The applicable payment system's own rules determine admission, execution, deadlines, and finality. Any GUDIYA control must be designed around those rules rather than assuming a universal pause button.

Circuit breakers at the three layers

A circuit breaker is an authorized mechanism that interrupts or restricts a defined class of activity when specified conditions arise. Ordinary pacing can continuously regulate flow within an operating envelope. A breaker changes the permitted mode of operation when that envelope is threatened or essential observations become unreliable. Neither response should automatically stop every activity in the system.

At the Agentic layer	the breaker could restrict new action initiation, delegation, recursive retries, or repeated optimization associated with the affected objective. In the deposit scenario, it might suspend eligible discretionary sweep initiation while preserving observation, explanation, and authorized essential activity. This reduces the generation of new demand before it reaches execution.
At the Digital Services layer	the breaker could restrict admission, routing, concurrency, or release of instructions. Its bank-transfer application would be a governed queue or a suspension of eligible new submissions before the relevant commitment point. The service must distinguish queued, accepted, executing, settled, and rejected states so that agents can respond correctly.
At the Resources layer	the breaker could control admission of new commitments against liquidity, inventory, energy, or other capacity. In the banking example, an authorized threshold could restrict additional commitments against insufficient liquidity headroom. The breaker cannot create liquidity or cancel consequences already incurred. It must account for existing obligations and dependent recipients.

These are proposed GUDIYA functions. The cited payment research supports queuing and liquidity management as engineering mechanisms; it does not validate this complete three-layer architecture or confer authority to deploy it.

Coordination across layers

A local breaker can worsen the wider situation if its meaning is not communicated. Suppose the Digital Services Layer queues a request while the Agentic Layer interprets the delay as a failure. The agent may retry, select another provider, or delegate the same task. A restriction intended to reduce pressure may then multiply outstanding requests.

Each intervention therefore needs a shared identity, scope, reason, and status that can propagate to the affected layers. Participating services should recognize duplicate instructions, expose pending commitments, and indicate when retries are inappropriate. GUDIYA should follow the authorized persistent objective across changes of model, host, and execution environment.

Coordination must also account for displacement. Restricting one institution or supplier may redirect demand toward another. Releasing liquidity pressure in one place may deprive a counterparty of an expected inflow. The Grid needs sufficient external visibility to assess these effects and report where its visibility ends.

Protection should be distributed according to response time. Fast local controls handle conditions that cannot wait for a distant authority. Wider Grid oversight coordinates aggregate limits and conflicts among local actions. A local breaker should not require a round trip to a national controller before taking a pre-authorized protective action.

Human policy governs interruption and reopening

The same Human Policy Supremacy that governs settlement pacing must govern every breaker. Policy defines the protected interests, triggering conditions, permissible scope, maximum duration, priority obligations, and the authority responsible for review. Neither an actant nor the Grid may silently enlarge that mandate to make its own objective easier to achieve.

A breaker should be narrowly scoped wherever possible. A surge in discretionary transfers need not imply suspension of all payments. Priorities and exceptions must be defined before stress, with records sufficient to evaluate who benefited, who bore the delay, and whether the stated justification remained valid.

GUDIYA must also have a declared response to loss of telemetry or control connectivity. Depending on the activity, that response may be a narrower envelope, temporary restriction, or continued essential service under local authority. There is no single fallback appropriate to every layer and every obligation.

Recovery is part of the circuit breaker

A breaker that trips successfully has completed only the interruption phase. Releasing every queued request at once may recreate the original surge. Recovery must establish that the relevant resource condition has improved, that observations are sufficiently reliable, and that the pending work is still authorized and appropriate.

The Grid can then reopen in measured stages, monitor the effect, and increase the admitted rate as the envelope permits. Separate trip and recovery thresholds can reduce repeated stopping and restarting near a

boundary. Staged release must also consider queue age, deadlines, and the risk that a delayed instruction has become stale.

Before release, agents and services should reconcile duplicate or canceled requests and revalidate commitments where required. They must preserve the status and finality of work already completed. A model replacement or runtime restart should not reset the history of outstanding obligations.

This extends the Stabilization Completion Principle: a breaker succeeds when the system reaches a verified acceptable condition and resumes activity under the applicable envelope. Issuing an interruption command alone does not establish stabilization.

A measurable settlement control loop

A practical design would observe pending demand and resource conditions, estimate aggregate consequences, select an authorized response, release work, and verify the resulting state. Control success requires more than issuing a rate limit: the Grid must establish that the intervention reduced stress without creating a larger downstream problem.

Useful measurements include value and count released per interval; concentration by institution or resource; queue age and deadline breaches; available capacity; and the correlation of agent requests. Breaker assessment should include detection and actuation latency, duplicate retries, repeated trips, recovery time, and the size of any surge after reopening. GUDIYA should also measure the cost of intervention, including delays imposed on users and essential obligations.

The test is comparative. Does paced execution improve outcomes relative to unrestricted release, additional liquidity or capacity, or another available response? Trials and replay should examine ordinary demand, synchronized bursts, delayed telemetry, strategic queue behavior, and failure of the controller. The analogy motivates this testing; it does not substitute for it.

GUDIYA Cognitome principle

Settlement velocity in the three-layer AI Economy should be governed by human-authorized policy, informed by the capacity and condition of the resources supporting execution. Coordinated circuit breakers at the Agentic, Digital Services, and Resource layers should restrict activity within their mandates, communicate their effects across layers, and support staged, verified recovery. The Grid must preserve clear finality, fair treatment, and visible accountability while establishing that its intervention improves system stability.

Agentic intelligence expands the capacity to discover and pursue opportunities. Digital services expand the capacity to execute them. Resource conditions determine what can actually be fulfilled. Human Policy Supremacy establishes how the system reconciles those capacities when they diverge.

Sources

1. Torsten Sløk, “Is an Agentic Bank Run Coming?”, Apollo, September 27, 2026.
2. Gary Gensler, “Systemic Risk in Artificial Intelligence”, SEC, September 19, 2024.
3. BIS Working Paper 1318, “Auction-based liquidity saving mechanisms”.
4. Antoine Martin and James McAndrews, “Liquidity-Saving Mechanisms”, Federal Reserve Bank of New York, Staff Report 282.