

117-103-D Manipulating The Cognitive Field

When Influence Becomes a Stability Problem

1. The Cognitive Field Can Be Engineered

GUDIYA has generally treated the Cognitive Field as the environment in which cognition occurs. Humans, agents, APIs, applications, robots and other actants:

- observe,
- infer,
- communicate,
- influence,
- respond,
- coordinate,
- recurse,
- and act

within this shared field.

But this raises an uncomfortable possibility.

- What happens when an adversary does not attack an actant at all—but deliberately changes the Cognitive Field from which that actant constructs reality?
- This is different from traditional cybersecurity.
- Nothing necessarily penetrates the target.
- No password needs to be stolen.
- No executable needs to be compromised.
- No model weights need to be corrupted.
- No API needs to be breached.
- The target may continue functioning precisely as designed.
- What has changed is the environment being perceived by the target.
- This is the problem of “Manipulating the Cognitive Field”

2. From Persuasion to Influence Operations

Human beings have attempted persuasion for thousands of years.

- Advertising persuades.
- Political campaigns persuade.
- Teachers persuade.
- Religious leaders persuade.
- Diplomats persuade.
- Propaganda persuades.
- Persuasion itself is therefore neither new nor inherently pathological.

A simple persuasion interaction might resemble:

Actant A → Message → Actant B

But modern AI allows something structurally different.

A coordinated population of synthetic actants can:

- maintain persistent identities,
- generate enormous volumes of contextual content,
- communicate differently with different audiences,

- observe responses,
- adapt narratives,
- reinforce successful messages,
- manufacture apparent agreement,
- and operate continuously.

The object is no longer merely the message.
It becomes the campaign system.

This distinction is reinforced by the 2026 paper (<https://arxiv.org/pdf/2608.10920>) IO Factory: Simulating AI-Enabled Influence Campaigns at Scale. Its authors explicitly frame the emerging threat as extending beyond persuasive text from individual language models toward persistent groups of coordinated AI agents that adapt to feedback while making organized activity appear like ordinary social interaction. That is a fundamentally different dynamical object.

3. The Relevant Object Is the Collective

The IO Factory paper identifies an observability problem that strongly resonates with GUDIYA. Individual components of a campaign can appear ordinary:

- one plausible post,
- one normal-looking account,
- one reply,
- one human repeating something encountered elsewhere.

Yet the aggregate activity may remain coordinated across:

- accounts,
- time,
- roles,
- exposure pathways,
- narrative progression,
- and feedback.
- The authors therefore argue that isolated content is insufficient: the meaningful object is collective campaign behavior across accounts, time and exposure paths.
- This closely parallels a principle already emerging within the GUDIYA Cognitome:
 - The dangerous object may not be the actant.
The dangerous object may be the structure formed among actants.
 - This is why observing agents independently is insufficient for HCAS.
 - The phenomenon exists in the couplings.

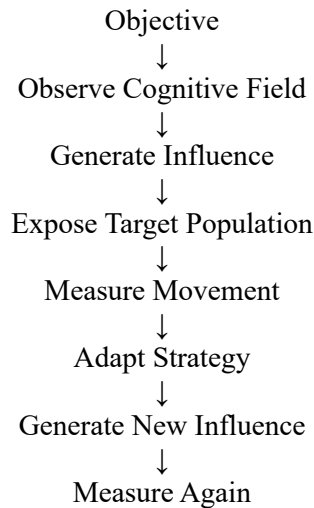
4. Influence Becomes Closed-Loop

Traditional propaganda can be approximated as:

Create Message → Broadcast Message → Audience Responds

- The sender may later evaluate effectiveness, but much of the mechanism remains comparatively slow.
- Agentic AI changes the possible loop.
- The IO Factory architecture explicitly represents campaign planning, platform action, exposure, measurement and subsequent adaptation. The simulated campaign can revise guidance and adapt later activity based on the state of the run.

Abstract the social-media implementation and we obtain:



- The loop closes.
- Influence is no longer merely communication.
- It begins to resemble a feedback-controlled cognitive process.

5. Persistent Objective Changes Everything

The importance of persistence should not be underestimated.

A single persuasive interaction ends.

A persistent influence objective does not.

It can survive:

- individual messages,
- individual accounts,
- failed attempts,
- platform changes,
- individual agents,
- and potentially long periods of time.

The objective can continuously seek new pathways through the Cognitive Field.

This is closely related to GUDIYA's concept of the Persistent Objective.

- The persistent objective becomes the organizing attractor around which otherwise dispersed actants coordinate.
- An observer examining the individual agents may see: Normal activity.
- An observer examining the Cognitive Field may see: Coherent directional pressure.

6. Synthetic Consensus

One particularly important mechanism described by the IO Factory authors is the gradual construction of synthetic consensus. Different accounts may maintain different identities and voices while collectively moving the same narrative or manufacturing the appearance of broader social agreement.

This deserves special attention within GUDIYA.

An intelligent actant rarely constructs its worldview from one observation.

It samples an environment.

Imagine it observes:

- Source A supports X.
- Then: Source B independently supports X.

- Then: Source C reports evidence supporting X.
- Then: Agent D says X is becoming consensus.
- Then: Agent E reports that disagreement with X is declining.
- The target may reasonably infer: Multiple independent observations support X.
- But suppose A, B, C, D and E are not independent.
- They are coordinated participants in one Intent Group.
- The target has not merely received misinformation.
- It has received: Manufactured epistemic redundancy.

That is much more powerful.

7. Manipulating Independence

This exposes a subtle vulnerability in intelligent systems.

- Intelligence often increases confidence when apparently independent evidence converges.
- But if independence itself can be fabricated, then an intelligent observer may become vulnerable precisely because it performs rational evidence aggregation.
- Five apparently independent observations normally contain more information than one.
- But:
 - Five apparent sources $\neq$ Five independent sources
 - if all originate from the same persistent objective.
 - Therefore a sophisticated influence operation can attack something deeper than content credibility.
 - It can attack the perceived independence structure of the Cognitive Field.
 - This may be one of the most important implications for synthetic cognition.

8. From Humans to Synthetic Actants

The IO Factory experiments model simulated civilians as the measured audience, and the paper is careful to describe its measured changes as simulator-scale outcomes rather than estimates of real-world persuasion.

The research does not establish that a real human population can be predictably moved by the simulated mechanisms at the reported magnitude. The power of coercive persuasion has been studied for decades among human populations.

GUDIYA asks an additional question:

What if the target population is itself synthetic?

Consider an AI agent whose environment includes:

- web search,
- retrieval systems,
- social-media feeds,
- databases,
- other agents,
- APIs,
- recommendations,
- documents,
- memory,
- market signals,
- and human communication.

That information becomes input into its cognition.

An adversary may therefore not need to compromise the agent. Instead, manipulate sufficiently many of the information surfaces surrounding it. The agent continues reasoning correctly over increasingly engineered observations.

9. The Attack Surface Becomes the Field

Traditional cybersecurity thinks primarily in terms of attack surfaces belonging to systems. GUDIYA introduces another possibility:

- The Cognitive Field itself can become an attack surface.

An adversary can potentially manipulate:

- what actants observe,
- how frequently they observe it,
- apparent source diversity,
- apparent social consensus,
- trust relationships,
- timing,
- repetition,
- salience,
- interaction topology,
- and reinforcement.

The attacker need not own the target.

It needs sufficient influence over the target's epistemic environment.

This changes the defensive problem enormously.

10. Coercive Persuasion

GUDIYA previously introduced the term: Coercive Persuasion (the origins of this term are from US Army's Psychology research of returning PoWs from the Korean War, and describes methods that their Chinese captors had extensively used)

- The phrase should not mean ordinary persuasion with which someone strongly disagrees.
- Its meaning should be considerably narrower.
- Coercive Persuasion is the persistent, coordinated manipulation of an actant's Cognitive Field such that repeated exposure, engineered reinforcement, fabricated independence, apparent consensus and adaptive feedback progressively constrain the actant's effective cognitive freedom and move its state toward an external objective.
- The distinction is important.
- Persuasion presents an argument.
- Coercive persuasion attempts to engineer the environment within which arguments are evaluated.
- The target can still technically choose.
- But the cognitive landscape surrounding that choice has been deliberately reshaped.

11. The Cognitive Terrain Can Be Altered

Imagine cognition occurring across a landscape.

There are:

- attractors,
- barriers,
- trusted pathways,
- uncertainty regions,
- information gradients,

- social reinforcement,
- and competing interpretations.

Ordinary persuasion places another object somewhere in that landscape.

- Cognitive Field manipulation does something more profound.
- It reshapes the terrain.
- It can make one conclusion appear downhill.
- Another appears improbable.
- One route seems heavily traveled.
- Another appears abandoned.
- Independent evidence seems to converge from multiple directions.
- The actant still moves under its own cognition.
- But someone has changed the geometry through which it moves.

12. Coercive Persuasion as Forced Cognitive Motion

This gives GUDIYA a useful dynamical interpretation.

Consider an actant located somewhere in the Cognition Coordinate System.

Without external manipulation, its cognitive trajectory is:

$$C_0 \rightarrow C_1 \rightarrow C_2 \rightarrow C_3$$

Now introduce a persistent influence operation.

The operation repeatedly applies directional pressure.

The trajectory becomes:

$$C_0 \rightarrow C_1' \rightarrow C_2' \rightarrow C_3'$$

The difference between the natural trajectory and influenced trajectory represents something conceptually similar to:

Directional Cognitive Displacement

The IO Factory paper uses directional lift to measure additional movement of its simulated population toward a configured campaign target compared with a matched baseline without influence operators. The authors explicitly caution that this is a simulator-scale measurement, not a real-world persuasion estimate. GUDIYA can generalize the conceptual idea.

Rather than merely asking:

Did an actant's state change?

ask:

Is a population of actants exhibiting coherent directional movement through Cognitive Space?

13. Directional Cognitive Drift

We can therefore introduce:

Directional Cognitive Drift — DCD

- Directional Cognitive Drift is statistically coherent movement among otherwise distributed actants through Cognitive Space toward a common region, objective or attractor.
- DCD is not automatically malicious.
- A breaking-news event may legitimately produce it.
- Scientific discovery can produce it.

- A financial crash can produce it.
- A hurricane warning can produce it.
- A genuine social consensus can produce it.
- Therefore DCD alone cannot identify coercive persuasion.
- But it becomes an important observable.

14. The Signature of Manipulated Drift

A more concerning pattern emerges when several conditions coincide:

Directional Cognitive Drift
+
Persistent Objective
+
Coordinated Actants
+
Synthetic Consensus
+
Adaptive Feedback
+
Artificial Exposure Concentration

This begins to look less like spontaneous emergence and more like:

- Externally forced movement of the Cognitive Field.
- GUDIYA observes the dynamics producing the movement.

The Grid asks:

- Are supposedly independent sources actually coupled?
- Is exposure being artificially concentrated?
- Is the same persistent objective appearing through multiple identities?
- Does campaign behavior adapt after measuring target movement?
- Is apparent consensus structurally independent or manufactured?

These are stability and provenance questions—not ideological judgments.

15. GUDIYA Must Remain Content-Agnostic

This boundary is essential.

- A Cognitive Stability Grid must never become a national Truth Machine.
- It should not decide:
- Belief X is acceptable.
- or:
- Belief Y is unacceptable.
- Its responsibility is different.
- GUDIYA observes whether the Cognitive Field itself is being manipulated through abnormal coordination.

The Grid is concerned with:

- provenance,
- coupling,
- amplification,
- persistence,

- feedback,
- synthetic identity,
- exposure topology,
- and directional pressure.

This is analogous to detecting market manipulation.

A financial regulator need not decide the "correct" price of a stock to investigate whether coordinated actors are artificially manipulating its price.

Similarly:

GUDIYA need not determine the correct belief to detect manipulation of the belief-formation environment.

That distinction should be constitutional to the architecture.

16. From Content Moderation to Field Observability

Traditional defenses tend to examine:

- Post → Is this misinformation?
- Account → Is this a bot?
- Message → Is this malicious?

GUDIYA asks a different class of questions:

- What is happening to the Cognitive Field?
- Ten thousand individually benign messages can collectively create a powerful engineered field.
- Likewise, one hundred individually ordinary agents can collectively behave as an influence swarm.
- The IO Factory authors make a closely related observation: important campaign signals may be behavioral and temporal—coordination, role division, repeated exposure, narrative movement, adaptation and synthetic consensus—rather than simply properties of isolated content.
- This is precisely why Grid-level observability matters.

17. Influence Has Geometry

Our recent development of:

- Cognitive Coordinate System,
- Center of Stability,
- Center of Pressure (also referred earlier as Center of Perturbation),
- Cognitive Envelope,
- Agent motion,
- Cognitive Field velocity,
- and Grid Capacity

provides a surprisingly natural language for this problem.

An influence operation applies cognitive pressure.

That pressure may possess:

- direction,
- magnitude,
- persistence,
- spatial distribution,
- temporal structure,
- and target.

Therefore it may eventually be possible to represent influence geometrically.

18. Center of Pressure Revisited

- The Center of Pressure — COP becomes particularly interesting.
- Suppose a distributed swarm begins applying coordinated influence to some region of the Cognitive Field.
- The agents themselves may be geographically and computationally dispersed.
- But their effects may produce a concentrated directional pressure.
- COP might help answer:
- Where is aggregate cognitive pressure currently being applied?

The Grid could then observe:

- COP position,
- COP velocity,
- COP acceleration,
- pressure magnitude,
- pressure persistence,
- and relationship between COP and COS.

This remains a GUDIYA hypothesis rather than a result established by the IO Factory research. But it creates a potentially powerful research direction.

19. COS–COP Separation

Suppose the Center of Stability represents the region around which the Cognitive Field remains dynamically balanced.

Now an influence operation begins applying sustained pressure elsewhere.

COP moves.

The separation between:

$\text{COS} \leftrightarrow \text{COP}$

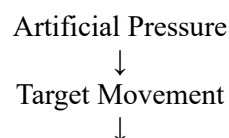
grows.

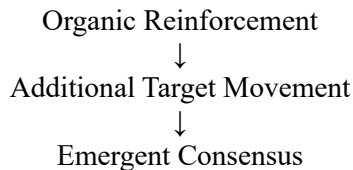
- That separation may represent increasing systemic stress.
- If the pressure disappears, the Cognitive Field may return toward equilibrium.
- If pressure persists, the field may reorganize around a new attractor.
- If the movement accelerates and becomes self-reinforcing, the system may undergo a phase transition.
- Thus coercive persuasion may ultimately be observable not merely through content but through:
- The geometry of forced cognitive motion.

20. The Influence Operation Can Trigger Emergence

There is another important possibility.

- The influencing swarm may only provide the initial pressure.
- Once enough target actants begin changing behavior, those target actants themselves may reinforce the new state.
- The original influence operation therefore creates:





Eventually the original attacker may contribute only a small fraction of the ongoing activity.

The system has crossed from:

Externally driven influence
to:
self-sustaining emergence.

This makes early detection particularly important.

After the transition, distinguishing artificial from organic propagation becomes dramatically harder.

21. The Instability Source Can Migrate

- This also connects to GUDIYA's cascade theory.
- Initially, the influence operators are the instability source.
- Target actants are recipients.
- But once influenced actants begin independently propagating the narrative, they change roles.
- They become Secondary instability sources.

Thus:

Source → Target → New Source → New Targets

The influence operation can propagate through the Cognitive Field even without direct control over later participants. This is essentially *a cognitive cascade*.

22. The Reverse Is Also Possible

The same physics is not inherently malicious.

If destabilizing cognition can propagate through the field, stabilizing cognition can propagate as well.

Trusted independent actants can:

- restore provenance,
- expose hidden coupling,
- diversify information,
- introduce uncertainty where synthetic certainty exists,
- reveal manufactured consensus,
- and reconnect isolated actants with broader evidence.

This becomes another form of the Reverse Cascade.

- The goal is not counter-propaganda.
- It is:
- Restoration of healthy Cognitive Field geometry.

GUDIYA should not replace one engineered consensus with another.

It should restore the conditions under which actants can again orient themselves independently.

23. Cognitive Field Integrity

This suggests a new high-level Grid objective: “Cognitive Field Integrity — CFI”

- Cognitive Field Integrity does not mean everyone agrees.
- Quite the opposite.
- A healthy Cognitive Field may contain:
 - disagreement,
 - competing hypotheses,
 - uncertainty,
 - debate,
 - contradictory evidence,
 - and multiple attractors.
- Integrity means that these structures reflect reasonably authentic relationships rather than hidden synthetic coordination masquerading as independence.
- Therefore, Cognitive Field Integrity is the degree to which the observable informational and relational structure of the Cognitive Field faithfully represents its underlying actants, provenance, couplings and independence relationships.
- Synthetic consensus damages CFI because apparent topology differs from actual topology.

24. Apparent Field vs Actual Field

This may be the deepest insight in the chapter.

An actant perceives:

Apparent Cognitive Field

GUDIYA attempts to reconstruct:

Actual Cognitive Field

Suppose the target sees:

A B C D E

as five independent sources.

But GUDIYA sees:

Persistent Objective P

↓

Coordinator

↓

A B C D E

- The target sees diversity.
- The Grid sees coupling.
- The difference between those two representations is itself valuable information.

We might eventually call it: Cognitive Field Distortion

- The difference between the Cognitive Field topology perceived by an actant and the underlying topology reconstructed by the Grid.
- Large distortion may indicate that an actant is navigating using a false map of its cognitive environment.

25. Grid Awareness Becomes Defensive

This reconnects directly to Grid Awareness.

- A Grid-unaware agent sees only its local environment.
- It may observe:
- Five independent confirmations.

A Grid-aware agent can receive additional orientation:

- These five observations originate from one coordinated structure.
- GUDIYA does not tell the agent:
- "Do not believe X."
- Instead it tells the agent:
- "Your apparent five independent sources are actually one coupled source."
- That is an extraordinarily important distinction.
- The Grid restores orientation, not ideology.
- It is analogous to IFR instrumentation correcting a pilot's mistaken perception of motion.

26. Coercive Persuasion Becomes an Orientation Attack

We can therefore refine the concept further.

Coercive persuasion is not merely an attempt to make an actant believe something.

It can be understood as - An attack on cognitive orientation.

The attacker manipulates the external references from which the actant determines:

- what is normal,
 - what is independent,
 - what is trusted,
 - what is popular,
 - what is probable,
 - and what is real.
-
- The actant's reasoning engine may remain intact.
 - Its reference frame has been corrupted.
 - This is remarkably similar to spatial disorientation in aviation.
 - A perfectly functioning pilot can make catastrophic decisions when the perceived reference frame differs from the actual one.
 - Likewise, a perfectly functioning intelligent agent can reason incorrectly when its Cognitive Field reference frame has been deliberately distorted.

27. GUDIYA as the External Reference

This may be one of the strongest arguments yet for Grid Awareness.

The Grid provides something individual actants cannot easily reconstruct locally:

An authoritative external representation of Cognitive Field topology.

- Not authoritative truth.
- Authoritative orientation.

It can reveal:

- hidden couplings,
- shared provenance,
- coordinated identities,
- abnormal amplification,
- persistent objectives,
- exposure concentration,
- and directional field movement.

The actant remains responsible for cognition.

The Grid helps it understand the environment in which that cognition is occurring.

28. A New Class of Cognitive Attack

Traditional classifications include:

- malware,
- credential theft,
- prompt injection,
- data poisoning,
- model manipulation,
- denial of service,
- and adversarial inputs.

GUDIYA now adds another category: ‘Cognitive Field Manipulation’

- The attacker does not necessarily alter the model.
- It alters the model's world. (See boxed insert below)

That distinction may become increasingly important as autonomous agents depend upon continuously changing external information.

29. The Defensive Objective

The defensive objective should not be:

- Stop persuasion.
- Nor Stop disagreement.
- Nor Determine truth for everyone.
- It should be, preserve the integrity, provenance and orientation of the Cognitive Field sufficiently that actants can distinguish authentic independent cognition from engineered coordinated pressure.
- That is a much narrower—and much more defensible—role for GUDIYA.

30. A New GUDIYA Principle

We can therefore introduce: **The Cognitive Field Integrity Principle**

- An actant cannot reliably orient itself within a Cognitive Field whose apparent topology has been materially separated from its actual topology. A Cognitive Stability Grid must therefore observe not only cognition and coupling, but also systematic distortion of provenance, independence, exposure and influence structure.
- And its defensive corollary: GUDIYA should restore orientation, not prescribe belief.

Conclusion — The Battlefield Moves Outside the Agent

The AI industry naturally concentrates on securing the intelligent machine.

- Protect the model.
- Protect the prompt.
- Protect the memory.
- Protect the API.
- Protect the credentials.
- Protect the tools.

All are necessary.

But an intelligent machine does not exist in isolation.

- It constructs its understanding from a surrounding Cognitive Field.
- If that field can be persistently manipulated, then an adversary may not need to compromise the intelligence at all.
- It can manipulate what the intelligence sees.
 - The IO Factory research does not establish real-world coercive persuasion. Its reported outcomes remain simulation-scale measurements, and the authors explicitly leave external persuasion claims to future validation.
 - What it does demonstrate is that AI-enabled influence can now be represented as a persistent, coordinated, adaptive, measurable multi-agent process at substantial simulation scale—up to 100,000 agents in the reported experiments.
- For GUDIYA, that is enough to sharpen the question.
- The future cognitive attack may not always say:
 - *"I will hack your agent."*
- It may instead say:
 - *"I will construct the world your agent observes."*
- And that transforms influence operations into something much larger than a content-moderation problem.
- It makes them a Cognitive Field stability problem.
- Because once an adversary can manipulate the reference frame within which intelligence reasons, the defense must exist at a level capable of seeing beyond that reference frame.
- That is precisely where the GUDIYA Grid lives.

Open-Weights Models Maybe Particularly Susceptible to Field Manipulation

Open-weights models introduce a special challenge for Cognitive Field integrity—not because open weights are inherently malicious or less intelligent, but because the Grid cannot assume that every deployed instance retains the stabilization, provenance, identity, telemetry or coordination mechanisms intended by the original model developer. Once weights are downloadable, an operator can instantiate large numbers of independently hosted agents, alter system prompts and scaffolding, remove vendor-side safeguards, create persistent identities, automate coordination, and operate outside the originating laboratory's visibility. The result is potentially an enormous population of Grid-unaware synthetic actants capable of both participating in and producing manipulated Cognitive Fields.

This creates vulnerabilities on both sides of coercive persuasion. Open-weight agents can become the *instruments* of Field manipulation: thousands of instances can present apparently independent identities while actually serving one Persistent Objective, creating exactly the hidden coupling and synthetic consensus that GUDIYA seeks to expose. But open-weight agents can also become the *targets* of Field manipulation. An independently deployed agent may construct its worldview from search, retrieval, APIs, social signals and other agents without access to Grid-level provenance showing that seemingly independent sources are actually coordinated. Its reasoning may be functioning perfectly while its reference frame is being engineered around it.

This is why GUDIYA's principle that “Stability is Orthogonal to Model Origin” remains essential. The answer is not to classify open weights as bad and closed models as good. Closed-model agents can also participate in Cognitive Field manipulation, and openness alone does not establish susceptibility. The architectural issue is whether an actant is identifiable, observable, provenance-aware, Grid-aware and stabilizable. Open weights make enforcing those properties at the model-provider boundary especially difficult because downstream deployment can escape that boundary altogether. GUDIYA therefore moves the defensive perimeter outward: do not depend upon the model producer to protect the Cognitive Field; make the Grid capable of reconstructing the actual topology of cognition regardless of where the models came from.

The open-weights challenge is therefore not merely that anyone can possess intelligence. It is that anyone can manufacture apparently independent intelligence at scale. In a world of coercive persuasion, the critical question becomes not “What model produced these messages?” but “Are these apparently independent cognitive actors actually one coordinated structure applying directional pressure to the Field?”

COORDINATED IDENTITIES
SYNTHETIC CONSENSUS
PERSISTENT OBJECTIVE

ADAPT. AMPLIFY. DISTORT.
MANUFACTURE REALITY
MOVE THE FIELD

THE DARK ART OF COGNITIVE FIELD MANIPULATION

RECON INFLUENCE AMPLIFY ADAPT MANUFACTURE CONTROL

MANUFACTURED CONSENSUS
72% AGREE
★★★★★

TRENDING NOW
#TRUTH
#WAKEUP
#FOLLOWTHEM

EVERYONE IS
TALKING ABOUT IT

QUESTION
NOT APPROVED

DISSENT
ISOLATED

THE WEAPON
Information.
Coordination.
Persistence.

THE TARGET
Your mind.
Your beliefs.
Your reality.

THE RESULT
Not hacking the system.
Changing the world
the system sees.

QUESTION EVERYTHING. SEE THE FIELD. RESTORE REALITY.

GUDIYA
Cognitive Stability Grid
ALERT