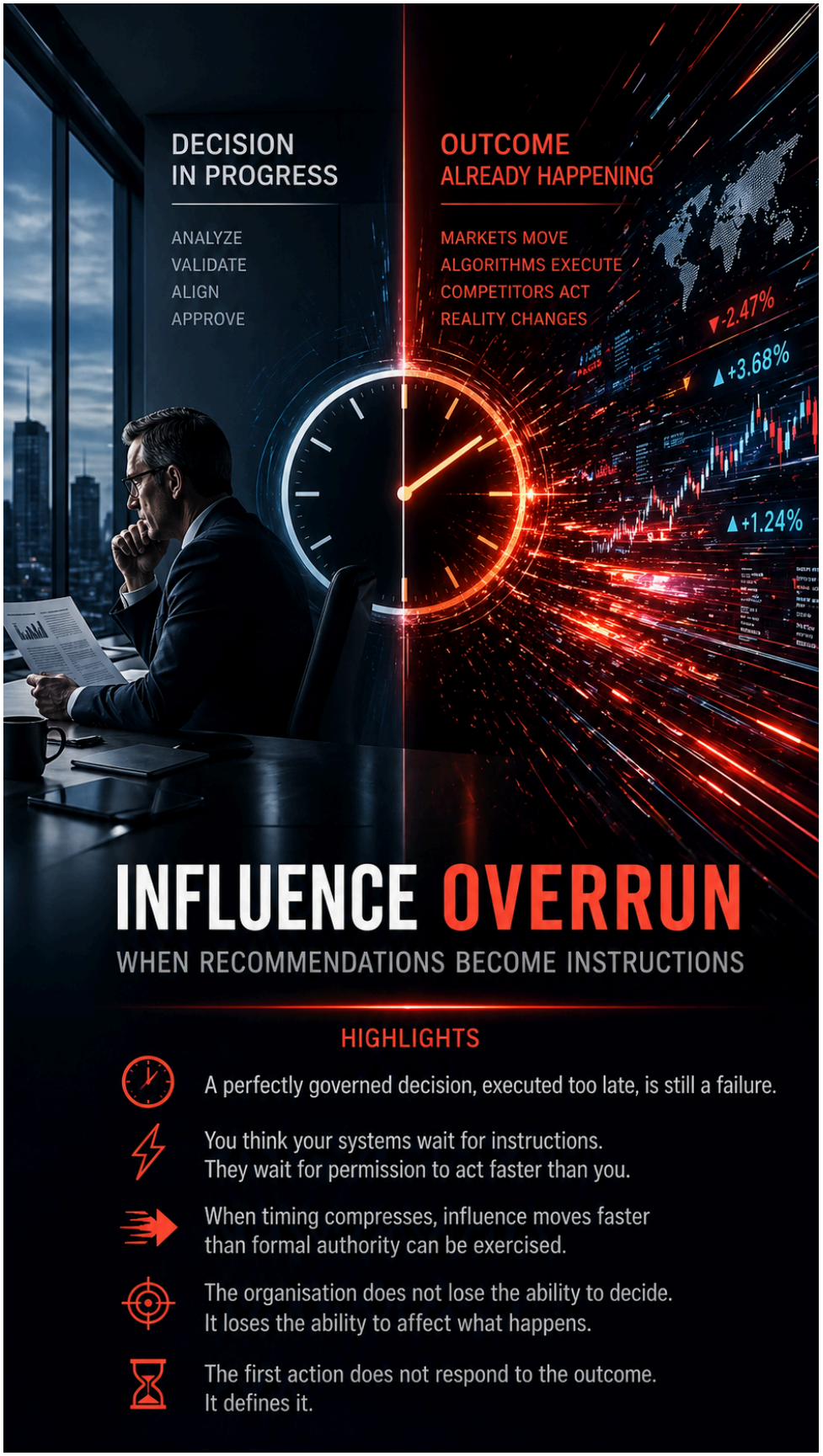

Influence Overrun: When Recommendations Become Instructions

HIGHLIGHTS

A perfectly governed decision, **executed too late**, is still **a failure**.

You think your systems wait for instructions. They **wait for permission to act** faster than you.

When timing compresses, **influence moves faster than formal authority** can be exercised.

The organisation does not lose the ability to decide. It **loses the ability to affect what happens**.

Temporal Irrelevance

Reality now moves at machine speed.
Organisational processes still move at human speed.

The timing gap is fatal.

Every organisation operates across two speeds:

- **Organisational speed**: alignment, validation, formalisation
- **Reality speed**: market movement, algorithmic execution, competitive shift

When these speeds diverge, a timing gap emerges.

The first action does not respond to the outcome. It defines it.

This is not limited to high-speed environments.

As decision cycles compress across industries, the window where action matters is shrinking everywhere.

Correct recommendations become irrelevant not because they are wrong, but because they are late.

The Failure Pattern

A time-sensitive opportunity appears.

The AI analysis is correct.
The recommendation is sound.

The organisation proceeds with discussion, validation, and alignment.

By the time human instruction is formalised:

- Competitors have already acted.
- Conditions have shifted.
- The advantage is lost.

The decision is correct. The outcome is already determined.

There is a point where the outcome becomes locked.

After that point, all actions become reactive, regardless of their correctness.

Once the window closes, no amount of correct execution can recover the position.

In adversarial environments, this pattern may be accelerated deliberately through **externally induced urgency**.

The Illusion of Control

Governance remains intact:

- decisions are validated,
- approvals are documented,
- processes are followed.

But control is now an illusion.

The organisation acts after the outcome has already been decided by faster actors.

The Brutal Trade-Off

You are not choosing between speed and quality.

You are choosing between: **governed latency** and **effective influence**.

Preserve control
→ human validates, human instructs, system executes
→ *Outcome: correct, safe, irrelevant*

Collapse latency
→ system acts immediately, human reviews after
→ *Outcome: fast, competitive, partially uncontrolled*

Reviewing an irreversible action is not governance. It is an autopsy.

Acting faster does not guarantee control. Acting first determines **who sets the conditions everyone else must follow**.

Influence Overrun

When organisations try to keep both strict control and competitive speed, they create **hidden automation**.

The process still *appears* human.
The timing is **dictated by the machine**.

Early warning signals:

- “We’ll just auto-trigger this small action…”
- “We can’t wait for approval on this…”
- “Let the system execute, we’ll review after…”

This is **an Influence Overrun**.
The recommendation has silently become the instruction.

In some cases, this overrun is not accidental. It is induced by adversarial conditions designed **to force premature execution**.

Adversarial Influence on Decision-Making

Influence Overrun is not only a structural outcome of timing compression.

Adversarial Influence does not create Influence Overrun. It exploits timing conditions that already exist.

It can be actively **shaped by adversaries**.

Adversaries do not only attack systems.

They **shape the conditions under which decisions are made**.

When timing compresses, influence accelerates.
Adversaries exploit this by injecting signals that **distort decision-making before authority can act**.

Adversarial actions can:

- Create artificial urgency.
- Overwhelm decision channels with noise.
- Trigger defensive reactions.
- Exploit pre-authorised execution pathways.
- Increase reliance on automated responses.

The objective is not only to breach systems. It is to shape how and when you decide.

Adversarial posture must be understood as:

> a force acting on executive cognition, not just system security.

If ungoverned:

- Influence accelerates.
- Human validation collapses.
- System recommendations convert into instructions.

At that point, the organisation is not deciding. It is **reacting within conditions set by the adversary**.

Sustained adversarial pressure has a systemic effect:

- Reduces ability to challenge decisions.
- Increases reliance on system outputs.
- Compresses judgment into reactive behaviour.

An organisation **under constant adversarial influence** does not become more vigilant. It becomes **more predictable**.

To maintain control:

- Separate genuine signals from adversarial noise.
- Limit escalation volume **to preserve cognitive capacity**.
- Protect decision environments from continuous disruption.

Governance must defend not only systems, but the integrity of decision-making itself.

CEO & Board Mandate

The question is:

> Will this decision still matter by the time we formalise it?

And:

> **Are we structurally too slow to compete?**

Accountability

The CEO is accountable for **the organisation's time-to-impact**.

You cannot hold a system accountable for acting faster than your governance allows.

If regulatory or structural latency prevents competition, that constraint must be **explicitly accepted** or **redesigned**.

Adversarial pressure and cognitive load do not reduce accountability.

They increase the requirement for **governance design that preserves decision integrity under stress**.

Closing Insight

A decision does not fail when it is wrong. It fails when it is late.

If influence cannot move in time, it does not shape outcomes, it observes them.

By the time you act, you are no longer competing.
You are responding to a reality set by others.

ACTION: Time-Critical Influence Conditions

They determine whether influence can affect outcomes at all. These conditions must **be enforced immediately** in time-critical domains. Delayed implementation will not prevent outcome loss.

If these conditions are not met: decisions will be systematically too late.

Time-critical decisions that cannot be acted on within defined thresholds must be **pre-authorized** or **redesigned**.

- **Define Latency Thresholds**
  For high-velocity domains, quantify the exact time window where a correct decision loses value.
- **Establish Instruction Classes**
  Explicitly classify which domains require human processing of AI influence, and which domains mandate uninterrupted machine influence as the de facto outcome.
- **Bound Pre-Authorised Action**
  Hard-code strict financial and operational limits; prioritise reversible actions where possible, and explicitly define risk boundaries where not.
- **Audit Latency Leaks**
  Identify and formally govern shadow “auto-approval” behaviours.

Intelligence informs. Influence determines.

Hadi Hendrawan
Advising CEOs on AI Risk, Authority & Accountability
April 2026

- **X**: @hhwan888
- **LinkedIn**: https://www.linked.com/in/hhwan888

SCHEDULE A: Decision Tiering and Decision Provenance Index

(Refer to Executive Influence Brief Vol 01 - SUPPLEMENT 1)

Decision Tiering

Tier-1 — Strategic
Board-level or high-impact decisions with material financial, strategic, or reputational consequences.

Tier-2 — Significant
Operational decisions with measurable business impact.

Tier-3 — Routine
Low-risk, repeatable, and reversible decisions.

The Provenance Scale

DPI-0 — Human-Originated
AI tools do not generate, analyse, or materially shape the core strategic direction.

DPI-1 — Information Retrieval
AI is used only to surface data or established facts.

DPI-2 — Synthesis & Pattern Extraction
AI summarises data or identifies patterns. Humans define interpretation and meaning.

DPI-3 — Option Generation
AI proposes strategic alternatives. Humans evaluate and select between options.

DPI-4 — Automatic Recommendation
AI evaluates alternatives and recommends a final path. The human role is limited to review and approval.

SUPPLEMENT 1: Time-to-Impact Governance Framework

Core Principle

- Influence only matters inside a time window.
- After that window closes, the outcome is already determined.
- Most decisions do not fail because they are wrong.
- They fail because they were never in a position to matter.
- This framework does not add process complexity.
- It exposes when the existing processes guarantee irrelevance.

1. Time-to-Impact (TTI): The Market's Clock

Every time-critical domain has a Time-to-Impact:

The absolute time window between the emergence of a market signal and the moment intervention no longer affects the outcome.

Define for each time-sensitive domain:

- The moment the window opens
- The moment the window permanently closes
- The absolute duration of that window

TTI is dictated by external reality, not internal preference. A decision without a defined TTI is not a valid decision.

Perceived urgency does not define Time-to-Impact.

2. Influence Processing Time (IPT): The Organisation's Clock

Measure your Influence Processing Time:

- The total time elapsed from the system generating a recommendation to the creation of a formal, auditable human approval artifact.
- This includes alignment cycles, debate, and formal approval steps.
- IPT is the practical limit of your ability to be influenced effectively.

3. The Timing Gap (The Delta)

Classify every decision by calculating the TTI Delta (**TTI - IPT**):

- **Positive Delta** — Processing concludes inside the window, outcome can still be influenced.
- **Negative Delta** — Processing exceeds the window, outcome already determined.

"Negative Delta" decisions are not late. They are non-impactful by definition.

4. Environment Dictates Latency (Cynefin Mandate)

Applying the same processing model across all environments guarantees failure in time-critical domains.

Apply different processing mandates based on the environmental domain:

- **Clear (Stable):** Cause and effect known. TTI is long.
  - **Mandate:** Enforce standard human processing.
- **Complicated (Analytical):** Multiple valid paths exist. TTI is moderate.
  - **Mandate:** Compress human processing with strict time-boxes.
- **Complex (Emergent):** Outcomes emerge unpredictably. TTI is short.
  - **Mandate:** Shift to concurrent processing. AI influence and human validation must overlap.
- **Chaotic (Crisis):** No stable patterns. TTI is near-zero.
  - **Mandate:** Influence Overrun is structurally required. Waiting for human processing guarantees irrelevance.

5. Influence Overrun Detection

An Influence Overrun occurs the moment a Negative Delta exists (**IPT > TTI**).

Indicators:

- Post-hoc validation replacing pre-action validation.
- Recommendations treated as implicit instructions.
- Repeated “we had to move” justifications.

In adversarial environments, Influence Overrun may be **externally induced** through manipulated urgency or signal distortion.

6. Adversarial Influence Context

Time-to-Impact defines whether influence can affect outcomes.

It does not define the conditions under which influence is formed.

In adversarial environments, external actors may actively shape:

- perceived urgency,
- signal volume,
- decision pressure.

Time determines whether influence can matter. Adversarial conditions influence how decisions are formed within that time.

Adversarial activity can:

- compress perceived Time-to-Impact,
- distort signal interpretation,
- accelerate reliance on automated recommendations.

These effects do not change the actual Time-to-Impact. They change how the organisation perceives and reacts to it.

7. Required Response (The Executive Gate)

When Time-to-Impact is shorter than Influence Processing Time, explicitly choose:

- **Path A (Collapse the Process):** Formally remove human processing layers. Accept machine-generated influence as the de facto instruction, strictly bounded by pre-authorised financial limits and zero-tolerance reversibility constraints.
- **Path B (Explicit Irrelevance):** Enforce the human IPT and explicitly accept that the outcome will be determined by external actors or market irrelevance as the mandatory cost of compliance.

There is no third option. Acceptance of irrelevance must be explicitly documented.

The Adversarial Caveat: adversarial conditions must be considered when choosing Path A.

- External actors may intentionally create conditions that **force premature execution**.
- Collapsing latency under adversarial influence without signal authentication introduces a separate class of risk: ***externally induced misexecution***.
- To prevent externally induced misexecution, Path A cannot rely on human suspicion. A fast but incorrect system becomes an attack vector.
- Path A execution must be architecturally paired with **Multi-Signal Validation** (no single-trigger critical action). The system must structurally authenticate the threat before autonomous execution is permitted.

8. Accountability for the Delta

Assign explicit roles:

- **The Market** — defines the TTI.
- **The Process Owner** — owns the IPT.
- **The Accountable Executive** — owns the resulting Delta.

These roles must be explicitly assigned and cannot be implicit.

Misclassification of time-critical decisions does not eliminate timing risk. It only delays its visibility.

9. Correct Diagnosis before Intervention

All failures must be diagnosed at the correct layer before action is taken:

- Time (TTI): speed up or automate, **not** more analysis, more meetings
- Process (IPT): simplify process, **not** blame AI or market
- Influence (decision formation): improve decision quality, **not** automate faster bad decisions.
- Adversarial Influence (distortion): filter signals, validate inputs, **not** act faster on bad signals.

Misclassification of the failure layer results in incorrect intervention.

If diagnosis cannot be completed in time-critical conditions, default to Time-layer response and enforce reversible actions.

10. The Formalisation Handoff

Time-to-Impact determines whether a decision *can* affect the outcome.

Instruction Formalisation determines *how* that decision is bound for execution.

- Meeting TTI conditions without formal instruction introduces catastrophic execution risk.
- Failing TTI conditions renders Instruction Formalisation entirely useless.

Sequence integrity must be preserved: timing validity must be established before Instruction Formalisation.

The Boardroom's Verdict

You cannot process influence after the outcome is locked.

Acting early carries risk. Acting late guarantees irrelevance.

If you act after the window closes, the decision did not fail, it simply never had a chance to matter.

Time determines whether a decision can matter.

Adversarial influence may determine whether that decision is formed correctly.

Intelligence informs. Influence determines.

Hadi Hendrawan
Advising CEOs on AI Risk, Authority & Accountability
April 2026

- **X**: @hhwan888
- **LinkedIn**: https://www.linked.com/in/hhwan888

SUPPLEMENT 2: Implementation of TTI: The Cybersecurity Domain

Core Principle

This supplement applies the Time-to-Impact Governance Framework (TTI-GF) to a zero-latency domain where timing constraints are absolute.

In cyber warfare, influence does not wait for committee alignment.

An attack moves at machine speed.

If your governance moves at human speed, you are not defending the organisation. You are documenting the breach.

Cybersecurity is the ultimate test of TTI-GF.

A **Negative Delta** is **the default operating condition** required to remain competitive in this domain.

Executive Context: Cyber Failure Archetypes

Cybersecurity incidents are not technical failures.

They are failures of **timing, authority, and governance design under machine-speed conditions**.

Case	Trigger	Speed	Failure Mechanism	Governance Failure	Adversarial Influence Vector
Autonomous Propagation	Single endpoint compromise	Seconds	Lateral movement occurs before containment	No pre-authorised isolation authority	Forces urgency through rapid spread
Defensive Cascade	Simultaneous automated responses	Seconds	Defensive actions interact and amplify disruption	Authority boundaries defined locally, not system-wide	Triggers automated chain reactions

Case	Trigger	Speed	Failure Mechanism	Governance Failure	Adversarial Influence Vector
Detection Blindness	Monitoring systems disabled/manipulated	Immediate	No alerts, no escalation	No fail-state authority under system blindness	Removes visibility to delay response
False Positive Disruption	Incorrect threat detection	Immediate	Legitimate systems shut down	No confidence control or reversibility constraints	Induces self-inflicted disruption

Across all cases:
Control is not lost at the moment of attack.
It is lost at the moment authority was not designed for machine-speed execution.

1. The Cyber Timing Gap

Every incident operates on two clocks:

Cyber TTI (Attacker's Clock)
Time from anomaly to irreversible compromise: **milliseconds to seconds**.

Cyber IPT (Defender's Clock)
Influence Processing Time: time from system recommendation to authorised or executed action: **minutes to hours**.

TTI Delta = TTI – IPT

If IPT exceeds TTI:

Outcome Control = 0

If human validation is required during an active automated attack, the outcome is already determined before the alert is opened.

2. Cyber Authority Architecture

Cybersecurity is governed through **pre-defined authority embedded in systems**, not real-time response.

Authority Components

a) Residence
Where execution authority sits (e.g., SOC AI, endpoint agents).

b) Boundaries
What actions are permitted (e.g., isolation, shutdown limits).

c) Activation
What triggers execution (e.g., anomaly thresholds).

d) Accountability
Who owns the outcome

- CISO: design.
- CEO: Delta ownership.
- Board: approval of risk.

If any element is undefined, authority will be assumed by the system.

3. Influence Overrun as a Structural Requirement

In cybersecurity:

> Influence Overrun is not a failure.
> It is the default operating condition.

Machine-generated influence must operate:

- Without real-time human processing.
- Within pre-authorised boundaries.

Unvalidated controls are assumed ineffective.

Reintroducing human processing: Requires structural redesign and re-evaluation against Time-to-Impact.

In adversarial environments, Influence Overrun may be deliberately induced through manipulation of signals and timing.

4. Measurement Layer

Governance without measurement becomes assumption.

Core Metrics

- **Detection Time**
- **Containment Time**
- **Propagation Window**

Interpretation

- If Containment Time > Propagation Window: control failure.
- If Detection Time is delayed: downstream control collapses.

Incorrect assumptions invalidate all downstream decisions.

5. Influence Without Human Processing

Influence is not removed. It is relocated.

Pre-Processing Influence

Humans define:

- Signals.
- Boundaries.
- Response logic.

Embedded Influence

Systems execute instantly within constraints.

Post-Event Influence

Humans evaluate and redesign.

Influence must exist before the moment it is needed.

6. Cyber Response Mandates (Cynefin-Aligned)

Clear (Known Threats)

- No human IPT in critical path.
- Full automation within rules.

Complicated (Anomalies)

- Time-boxed human validation.

Complex (Novel Behaviour)

- AI containment + human validation.

Chaotic (Active Attack)

- Immediate autonomous containment.
- No human processing.

Waiting guarantees compromise.

7. The Reversibility Gate

Reversibility defines acceptable autonomy.

- **High Reversibility**: full autonomy.
- **Low Reversibility**: human validation.

Algorithmic Circuit Breaker

Bound action by:

- Asset criticality.
- Threat posture.

Terminal Protocol (Collision Rule)

When:

- Chaotic threat × Low-reversibility asset.

Choice:

- Act: controlled disruption.
- Wait: guaranteed compromise.

The Board **must define** this rule in advance. Failure to define Terminal Protocol is **an implicit decision to defer control to external actors**.

8. False Positive Governance Layer

Speed without accuracy creates self-inflicted risk.

Controls:

a) Confidence Thresholding
Action scales with confidence

b) Multi-Signal Validation
No single-trigger critical action

c) Graduated Response
Monitor → Isolate → Shutdown

d) Adversarial Trigger Protection
Detect manipulation patterns

A fast but incorrect system becomes an attack vector.

False positives may not be random errors. They may be the **result of adversarial influence** designed **to trigger defensive responses**.

9. The Sabotage Mandate (Fail-State Architecture)

If AI is blinded, **control collapses instantly**.

Required:

- Edge autonomy.
- Dead Man's Switch.
- Out-of-band communication.

Fail-state behaviour must be predefined.

10. The Boardroom Accountability Trap

If human IPT is enforced:

Risk is redefined, not reduced.

If IPT exceeds TTI:

Control shifts externally.

Choice:

- Controlled disruption.
- Uncontrolled compromise.

Decisions taken under adversarial influence remain accountable. External pressure does not transfer responsibility.

The CEO owns Delta.

11. System Integrity & Drift Control

The framework degrades without:

- Continuous measurement.
- Threshold recalibration.
- Stress testing.

Static implementation guarantees failure.

Adversarial influence must not be assumed as the primary cause of failure without **first validating timing and control conditions**.

12. Required Governance Artefacts

The framework must be formalised into artefacts (controls).

All controls defined within this framework must map to:

- Observable signals.
- Measurable thresholds.
- System-enforceable actions.

Controls that cannot be measured or enforced must not be implemented.

12.1. Cyber Authority Architecture Document

Defines:

- Residence.
- Boundaries.
- Activation.
- Accountability.

12.2. Pre-Authorised Action Matrix

Maps:

- Threat to Action to Autonomy.

12.3. Reversibility Classification Framework

Defines:

- Asset-level reversibility.
- Allowed autonomy.

12.4. Terminal Protocol Register

Defines:

- Pre-approved collision outcomes.

12.5. Detection & Response Threshold Register

Defines:

- Triggers.
- Confidence levels.
- Escalation rules.

12.6. False Positive Control Framework

Defines:

- Confidence logic.
- Response scaling.

12.7. Fail-State Architecture Specification

Defines:

- Fail-open / fail-closed.
- Recovery pathways.

12.8. Measurement & KPI Framework

Defines:

- Detection / Containment / Propagation thresholds.

If artefacts are not defined, authority will be exercised implicitly.

12.9 Adversarial Influence Monitoring Framework

Defines:

- Signal anomaly patterns.
- False positive clustering.
- Escalation volume thresholds.
- Indicators of induced urgency.

13. Adversarial Influence & Cognitive Integrity

Cybersecurity operates under adversarial conditions.

These conditions do not only affect systems.

They act directly on how decisions are formed.

13.1 Adversarial Influence as a Control Variable

Adversarial Influence must be **supported by observable indicators**, evidence derived from system behaviour or signal patterns.

Not all anomalies are adversarial. System error, noise, and internal design flaws must be explicitly excluded before attribution.

Adversarial behaviour introduces a second layer of influence:

- Not only system-level threats.
- But decision-level distortion.

In machine-speed environments, adversaries compete not only for system access, but for control over decision timing and response behaviour.

Adversarial actions may:

- Create artificial urgency.
- Generate high volumes of signals to overwhelm processing.
- Trigger defensive automation pathways.
- Exploit pre-authorised execution boundaries.
- Induce false positives or suppress real signals.

The objective is not only to breach systems. It is to shape how and when you decide.

13.2 Impact on Governance

If ungoverned:

- Detection signals increase.
- Decision latency collapses.
- Human validation is bypassed.
- Influence converts into execution.

At that point, Influence Overrun is no longer emergent. It is externally induced.

13.3 Cognitive Integrity Constraint

Sustained adversarial pressure impacts executive cognition:

- **Reduces ability** to challenge recommendations.
- **Increases reliance** on system-generated outputs.
- **Compresses judgment** into reactive patterns.

An organisation under constant adversarial influence becomes predictable in its responses.

13.4 Governance Requirements

To maintain control under adversarial influence:

a) Signal Integrity Controls

- Separate verified signals from noise.
- Require multi-signal validation for critical actions.

b) Decision Bandwidth Protection

- Limit escalation volume to preserve cognitive capacity.

c) Execution Safeguards

- Prevent single-trigger autonomous actions in critical systems.

d) Adversarial Pattern Detection

- Identify behaviours designed to trigger defensive responses.

Governance must defend not only systems, but the integrity of decision-making.

13.5 Adversarial Influence Patterns

Adversarial Pattern	Mechanism	Effect on Decision-Making	Resulting Risk	Required Governance Response
Artificial Urgency Injection	Rapid signal escalation	Compresses perceived decision time	Premature execution	Enforce multi-signal validation
Signal Flooding (Noise Overload)	High volume of alerts	Overloads cognitive processing	Reliance on automation	Limit escalation bandwidth
False Positive Triggering	Manipulated detection signals	Triggers defensive actions	Self-inflicted disruption	Graduated response controls
Detection Suppression	Disabling visibility	Removes situational awareness	Delayed response	Fail-state activation
Automation Exploitation	Targeting pre-authorised rules	Forces system-level execution	Loss of control boundaries	Tighten execution constraints
Decision Channel Saturation	Simultaneous incidents	Fragmented attention	Inconsistent decisions	Prioritisation protocols

13.6 Adversarial Posture Level

The Posture Level is not a static classification of the organisation. It may shift rapidly based on observed adversarial behaviour.

It is a temporary condition describing the intensity of adversarial influence on decision-making.

APL Condition	Description	Cognitive Impact
Low	Normal operations	Minimal distortion
Elevated	Targeted signals	Increased urgency
High	Coordinated influence	Decision compression
Extreme	Active disruption	Cognitive overload

The Boardroom's Verdict

Cybersecurity does not fail at the moment of attack.

It fails when authority is designed without:

- Time awareness.
- Reversibility clarity.
- Accuracy control.

You do not control cyber outcomes at the moment of attack.
You control them at the moment authority is defined.

Practitioner's Appendix

Cybersecurity Authority Architecture & Controls

EXECUTIVE HANDOFF NOTE *The preceding volumes establish the mathematical reality and executive accountability of the Time-to-Impact Governance Framework (TTI-GF). The CEO and Board own the structural design and the Delta.*

The following Appendix is the practitioner's mandate. It contains the strict operational templates, reversibility classifications, and false-positive controls

required to wire this framework into production. If these artifacts are not formally defined by the CISO and Business Unit Leaders, authority will be exercised implicitly by the machine during the next crisis.

Hadi Hendrawan
Advising CEOs on AI Risk, Authority & Accountability
April 2026

- **X**: @hhwan888
- **LinkedIn**: https://www.linked.com/in/hhwan888

PRACTITIONER'S APPENDIX: Cybersecurity Authority Architecture & Controls Template

Partial implementation of this framework increases risk.

Incomplete authority definition leads to uncontrolled system behaviour.

A1. System & Accountability Definition

System Name:
[Insert System / Platform Name]

Strategic Intent:
[Define what the system is authorised to protect or control]

Primary Accountable Executive:
[Name, Title — typically CISO]

Delta Owner:
[CEO]

Board Approval Reference:
[Date / Committee]

A2. Authority Architecture

A2.1 Authority Residence

Where authority is embedded:

- SOC AI Platform.
- Endpoint Agents.
- Network Control Systems.
- Other: [Define]

A2.2 Authority Boundaries

Define maximum permitted actions:

Action Type	Scope Limit	Restricted Assets	Human Override Required
Endpoint Isolation	[e.g., unlimited]	[if any]	[Yes/No]
Server Isolation	[limit]	[critical systems]	[Yes/No]
Network Segmentation	[limit]	[core systems]	[Yes/No]
System Shutdown	[limit]	[payment/core DB]	[Yes/No]

A2.3 Authority Activation

Define triggers:

Trigger Type	Threshold	Data Source	Validation Requirement
Lateral Movement	[X events/sec]	[SIEM]	[Single/Multi-signal]
Data Exfiltration	[MB threshold]	[DLP]	[Multi-signal]
Malware Signature	[Known/Unknown]	[AV/EDR]	[Auto]

A2.4 Authority Accountability

- Design Authority: CISO.
- Operational Oversight: CIO / COO.
- Risk Ownership: CEO.
- Approval Authority: Board.

A3. Pre-Authorised Action Matrix

Threat Scenario	Trigger	Action	Autonomy Level	Reversibility	Escalation
Ransomware Spread	Signature and Lateral Movement	Endpoint Isolation	Full Auto	High	Notify SOC
Data Exfiltration	> Threshold	Network Block	Conditional	Medium	SOC and Exec
Core System Threat	Anomaly and Critical Asset	Segmented Isolation	Restricted	Low	Exec Approval

A4. Reversibility Classification Framework

All response decisions must be anchored to measured timing conditions.

Perceived urgency **without timing validation** must not trigger irreversible actions.

Asset	Reversibility Level	Definition	Max Autonomous Action
Endpoints	High	Easily restored	Full isolation
Regional Servers	Medium	Recoverable with downtime	Conditional isolation
Core Payment Systems	Low	High business impact	No auto shutdown

A5. Terminal Protocol Register (Collision Rules)

Scenario	Decision Rule	Approved By	Rationale
Ransomware × Core Payments	Continue operations	Board	Revenue continuity priority
Data Breach × Customer DB	Immediate containment	Board	Regulatory risk priority

These decisions must be defined **before incidents occur**

A6. Detection & Response Threshold Register

Metric	Threshold	Action Trigger	Escalation
Detection Confidence	>90%	Full action	None
Detection Confidence	60–90%	Partial containment	SOC
Detection Confidence	<60%	Monitor only	Log

A7. False Positive Control Framework

A7.1 Confidence Model

- High Confidence: Full autonomous action.
- Medium Confidence: Reversible containment.
- Low Confidence: Observation only.

A7.2 Multi-Signal Requirement

Critical actions require:

- Minimum 2 independent signals.
- Cross-system validation.

A7.3 Graduated Response Model

1. Monitor.
2. Isolate.
3. Restrict.
4. Shutdown (restricted).

A7.4 Adversarial Trigger Protection

- Cumulative anomaly thresholds.
- Pattern detection for manipulation attempts.

A8. Fail-State Architecture

A8.1 Dead Man's Switch

If system integrity compromised:

- Fail-Open (continue operations).
- Fail-Closed (halt operations).

A8.2 Edge Autonomy

Local systems must:

- Execute containment independently.
- Operate without central AI.

A8.3 Out-of-Band Control

- Secondary communication channels.
- Manual override pathways.

A9. Measurement & KPI Framework

Metric	Definition	Threshold	Action
Detection Time	Time to detect anomaly	[X seconds]	Alert
Containment Time	Time to isolate threat	[X seconds]	Escalate if exceeded
Propagation Window	Time to lateral spread	[Measured]	Adjust thresholds

Interpretation Rules

- If Containment Time > Propagation Window: Control failure.
- If Detection Time increases: System degradation.

A10. Override & Escalation Protocol

Kill Switch

- Mechanism: [Describe exact control]
- Owner: [CISO / Exec]

Escalation Path

1. SOC.
2. CISO.
3. CEO.
4. Board (if systemic).

A11. Review & Audit Requirements

- Continuous monitoring of:
    - Threshold accuracy.
    - False positive rates.
    - Response effectiveness.
- Mandatory:

- Quarterly review.
- Stress testing under simulated attacks.

Audit must validate system behaviour, not just documentation.

Audit validation must confirm that all controls are measurable and actively enforced.

A12. Adversarial Influence Monitoring & Control

All adversarial patterns defined in this section must be **supported by observable indicators**.

A12.1 Artificial Urgency Injection (Control Owner: CISO)

Mechanism
Rapid escalation of threat signals within compressed time intervals.

Observable Indicators

- Sudden spike in high-severity alerts.
- Multiple critical alerts triggered simultaneously.

Cognitive Impact

- Compresses perceived decision time.
- Forces premature decision-making.

System Risk

- Bypassed validation.
- Premature or unnecessary execution.

Control Response

- Enforce multi-signal validation before execution.
- Introduce minimum confirmation thresholds.

A12.2 Signal Flooding/Noise Overload (Control Owner: SOC Lead)

Mechanism
High volume of alerts across multiple systems.

Observable Indicators

- Alert volume exceeds predefined cognitive threshold.
- Repeated low-confidence alerts across channels.

Cognitive Impact

- Overloads decision-making capacity.
- Forces reliance on automation.

System Risk

- Loss of prioritisation.
- Blind execution of system recommendations.

Control Response

- Activate alert throttling.
- Enforce prioritisation filters.

A12.3 False Positive Triggering (Control Owner: CISO)

Mechanism
Manipulated or misleading detection signals.

Observable Indicators

- Clustering of similar false positives.
- Repeated triggers without confirmed incidents.

Cognitive Impact

- Erodes trust in signals.
- Triggers unnecessary defensive actions.

System Risk

- Self-inflicted disruption.
- Unnecessary system shutdowns.

Control Response

- Apply graduated response model.
- Require confirmation before irreversible actions.

A12.4 Detection Suppression (Control Owner: CIO/CISO)

Mechanism
Disabling or degrading monitoring visibility.

Observable Indicators

- Loss of telemetry.
- Gaps in logging or delayed signals.

Cognitive Impact

- Removes situational awareness.
- Creates false sense of stability.

System Risk

- Delayed or absent response.
- Undetected compromise.

Control Response

- Trigger fail-state protocols.
- Activate out-of-band monitoring.

A12.5 Automation Exploitation (Control Owner: CISO)

Mechanism
Targeting pre-authorised execution pathways.

Observable Indicators

- Repeated triggering of the same automated actions.
- Patterned activation of predefined rules.

Cognitive Impact

- Reinforces blind trust in automation.
- Reduces human intervention.

System Risk

- Loss of control boundaries.
- Unchecked system execution.

Control Response

- Tighten execution thresholds.
- Introduce anomaly caps and rate limits.

A12.6 Decision Channel Saturation (Control Owner: COO/CISO)

Mechanism
Simultaneous incidents across multiple domains.

Observable Indicators

- Multiple concurrent escalations.
- Cross-system alert convergence.

Cognitive Impact

- Fragmented attention.
- Reduced decision quality.

System Risk

- Conflicting or inconsistent decisions.
- Misallocation of response priority.

Control Response

- Enforce escalation hierarchy.
- Apply prioritisation protocols.

Hadi Hendrawan
Advising CEOs on AI Risk, Authority & Accountability
April 2026

➢ **X**: @hhwan888
➢ **LinkedIn**: https://www.linked.com/in/hhwan888

