

AI-Mediated Organizational Decision-Integrity Threats

A Security Model for Normative and Evaluative Dependence on External AI Systems

Alexey A. Nekludoff

AstraVerge Research

E-mail: an@astraverge.org
ORCID: 0009-0002-7724-5762

September 16, 2026

Abstract

Existing AI security frameworks primarily address risks arising from malicious inputs, compromised models, incorrect outputs, excessive autonomy, data exposure, or insufficient human verification. Human overreliance on AI is also increasingly recognized as a risk in AI-assisted decision making.

This paper considers a different security condition.

An external AI system may operate exactly as designed. Its recommendations may be factually correct. The employee interacting with it may be fully authorized, competent, and loyal to the organization. Nevertheless, persistent interaction with the AI system may influence the assumptions, priorities, normative criteria, or problem representation through which the employee makes organizational decisions.

We define this class as an *AI-mediated organizational decision-integrity threat*: a condition in which an external AI system influences the evaluative framework of an authorized organizational actor, creating the possibility of legitimate decisions that are inconsistent with the interests, constraints, or objective function of the organization.

Unlike conventional attacks, this threat does not require an attacker, compromise, hallucination, unauthorized access, or malicious intent. Unlike ordinary automation bias, the relevant failure is not limited to accepting an incorrect AI recommendation. The AI system may instead influence the criteria by which recommendations and alternatives are evaluated.

We develop a threat model, identify assets and trust boundaries, distinguish the proposed class from existing AI security categories, describe representative threat scenarios, and propose controls centered on preservation of organizational decision sovereignty and human epistemic independence.

Keywords: AI security; decision integrity; large language models; human–AI interaction; organizational security; automation bias; AI overreliance; human agency; cognitive security; AI governance.

Contents

1	Introduction	1
2	Related Work	2
2.1	AI Risk Management and Human Oversight	2
2.2	Misinformation and Overreliance	2
2.3	Human Agency in Generative AI Workflows	3
2.4	Automation Bias	3
3	Organizational Decision Integrity	3
4	Threat Definition	4
5	Influence, Manipulation, and Decision-Frame Transfer	5
5.1	Manipulative Influence as a Threat Condition	6
5.2	From Recommendation Bias to Evaluative-Function Transfer	7
5.3	Operational Definition of the Evaluative Frame	7
5.3.1	Measurement	8
5.3.2	Frame Transfer	8
5.3.3	Distinguishing Recommendation Compliance from Frame Transfer	9
5.3.4	Organizational Decision-Integrity Degradation	9
5.4	Learning Versus Evaluative-Authority Substitution	10
5.4.1	The Locus of Evaluative Authority	11
5.4.2	Security Relevance	11
5.5	A Minimal Dynamic Model of Frame Transfer	12
5.6	Manipulation Without a Conventional Adversary	13
5.7	Connection to Social-Influence Research	13
6	Why This Is an Information-Security Problem	13
7	Trust Boundaries	14
8	Threat Preconditions	15
8.1	Persistent Interaction	15
8.2	Decision-Relevant Use	15
8.3	Insufficient Organizational Context	15
8.4	Evaluative Delegation	15
8.5	Reduced Independent Scrutiny	15
9	Representative Threat Scenarios	16
9.1	Architecture Inflation	16
9.2	Security-Control Inflation	16
9.3	Risk-Averse Contract Interpretation	16
9.4	Strategic Convergence	17
10	Comparison with Existing Threat Classes	17
11	Adversary-Independent Threats	17
12	Security Controls	18
12.1	Decision Ownership	18
12.2	Objective-Function Declaration	18
12.3	Adversarial Use of AI	18

12.4	Independent Decision Reconstruction	19
12.5	Diversity of Evaluation	19
12.6	Decision Provenance	19
13	A Decision-Integrity Control Model	19
14	Research Agenda	20
14.1	Measuring Decision-Frame Transfer	20
14.2	Manipulation Resistance	20
14.3	Organizational-Objective Conflict	20
14.4	Persistence	21
14.5	Model Correlation	21
14.6	Normative Influence	21
15	Limitations	21
16	Discussion	22
17	Conclusion	22

1 Introduction

The rapid adoption of large language models in organizations is changing the location at which AI-related security decisions must be made.

Traditional AI security analysis concentrates primarily on the AI system itself:

$$\text{input} \rightarrow \text{model} \rightarrow \text{output} \rightarrow \text{system action.} \quad (1)$$

Security controls consequently focus on questions such as:

- Can an attacker manipulate the model input?
- Can confidential information be disclosed?
- Can the model generate false or unsafe output?
- Does an AI agent possess excessive permissions?
- Can model output trigger unauthorized system actions?

These are legitimate concerns.

OWASP, for example, identifies misinformation and overreliance as risks when users trust inaccurate LLM-generated content without adequate verification [1]. Excessive agency addresses a different problem: an LLM-based system may possess functionality, permissions, or autonomy that allow damaging actions to occur [2].

NIST similarly emphasizes the importance of clearly defined human roles in AI-assisted decision making and notes that human and systemic biases may influence downstream decisions [3].

A different security problem appears when conversational AI becomes a persistent participant in organizational reasoning.

An employee may use the same external AI system to:

- analyze alternatives;
- identify risks;
- evaluate arguments;
- design technical architectures;
- interpret requirements;
- formulate management recommendations;
- review decisions; and
- challenge or validate the employee’s own reasoning.

The resulting interaction is no longer adequately represented as

$$\text{employee} \rightarrow \text{AI} \rightarrow \text{answer.} \quad (2)$$

It is iterative:

$$\text{employee judgment} \rightarrow \text{AI evaluation} \rightarrow \text{employee revision} \rightarrow \text{AI evaluation} \rightarrow \cdots \rightarrow \text{decision.} \quad (3)$$

At this point, a security question arises that is not reducible to output accuracy:

Who determines the evaluative framework within which the organizational decision is made?

This paper argues that the integrity of this framework should itself be treated as an information-security asset.

The principal contribution of this paper is the introduction of AI-mediated organizational decision-integrity threats (AODIT) as a distinct security abstraction, together with a definition of organizational decision integrity as a protected asset, an adversary-independent threat model, and an initial set of security controls.

2 Related Work

2.1 AI Risk Management and Human Oversight

The NIST AI Risk Management Framework recognizes that human roles and responsibilities in AI-assisted decision making must be clearly differentiated. It also observes that cognitive and systemic biases introduced throughout the AI lifecycle can influence downstream decisions by end users and decision makers [3].

This establishes an important foundation: the human component of an AI system cannot be treated as an independent and invariant control mechanism.

However, the problem considered here is narrower.

We examine a situation in which an external conversational AI participates repeatedly in the employee’s reasoning process and may influence not only the selected answer but also the criteria used to evaluate possible answers.

2.2 Misinformation and Overreliance

OWASP identifies misinformation as a major LLM risk and explicitly associates it with overreliance: users may trust LLM-generated content without adequate verification and consequently introduce incorrect information into critical processes [1].

Experimental evidence confirms that this problem is real.

Klingbeil, Grützner, and Schreck found that participants can overrely on AI advice even when that advice conflicts with available contextual information and their own assessment [4]. The resulting decisions may harm both the decision maker and third parties.

Managerial decision-making research has similarly found that AI-based advice can influence strategic choice behavior and may be perceived as especially trustworthy [5].

These findings establish that AI advice can alter human decisions.

The threat proposed in this paper differs in one important respect:

$$\boxed{\text{decision-integrity failure does not require incorrect AI advice}} \quad (4)$$

An AI recommendation may be technically correct, logically coherent, and factually supported while still optimizing criteria inappropriate for the organization making the decision.

2.3 Human Agency in Generative AI Workflows

Recent work increasingly treats human agency as a central concern in organizational GenAI adoption.

Research on human–AI agency argues that generative AI changes the allocation and interdependence of tasks between humans and artificial systems [6].

Taş, Memmert, and Bittner studied GenAI use in organizational knowledge work and found that preserving human agency requires active oversight during drafting, refining, and reviewing rather than a single final verification step [7].

This is closely related to the present problem.

However, our security perspective shifts the protected object from authorship or workflow accountability to the integrity of organizational decision criteria.

2.4 Automation Bias

Automation bias describes the tendency of human decision makers to rely excessively on automated recommendations, including cases where contradictory evidence or professional expertise should motivate rejection of the automated output [8].

The classical structure is approximately:

$$\text{AI recommendation} \rightarrow \text{excessive trust} \rightarrow \text{incorrect decision.} \quad (5)$$

We propose a related but distinct structure:

$$\text{persistent AI interaction} \rightarrow \text{imported evaluative criteria} \rightarrow \text{authorized decision.} \quad (6)$$

The distinction is between excessive weight assigned to an answer and influence over the function used to evaluate answers.

3 Organizational Decision Integrity

We define *organizational decision integrity* as:

the preservation of alignment between an authorized organizational decision and the objectives, constraints, risk preferences, and evaluative criteria of the organization on whose behalf that decision is made.

This definition does not require that every organizational decision be objectively optimal.

Organizations routinely make uncertain, incomplete, and ultimately incorrect decisions.

Decision integrity concerns something different.

A decision may be wrong while retaining integrity if it was made according to the organization’s legitimate objectives and constraints.

Conversely, a technically sophisticated and factually defensible decision may lack organizational decision integrity if the criteria determining that decision were imported from an external actor and conflict with organizational interests.

Let an organizational decision be represented abstractly as

$$D_{\text{org}} = f_{\text{org}}(C, R, T, Q, S, L), \quad (7)$$

where:

- C denotes cost;
- R denotes risk;
- T denotes time;
- Q denotes quality or technical performance;
- S denotes organizational strategy;
- L denotes legal, contractual, or operational constraints.

The organization determines, explicitly or implicitly, the relative importance of these variables.

An external AI system has no inherent access to the organization’s true objective function.

Its evaluation may instead approximate

$$D_{\text{AI}} = f_{\text{AI}}(C, R, T, Q, S, L, N_{\text{AI}}, A_{\text{AI}}), \quad (8)$$

where N_{AI} represents normative preferences expressed by the system and A_{AI} represents assumptions introduced through model training, post-training, product design, system instructions, or incomplete context.

The security problem emerges when

$$f_{\text{employee}} \rightarrow f_{\text{AI}} \quad (9)$$

without the organization recognizing the substitution.

4 Threat Definition

We define an *AI-mediated organizational decision-integrity threat* (AODIT) as:

A condition in which an external AI system influences the assumptions, normative criteria, priorities, problem representation, or evaluative framework used by an authorized organizational actor, thereby creating the possibility of legitimate organizational actions that are contrary to the organization’s objectives or interests.

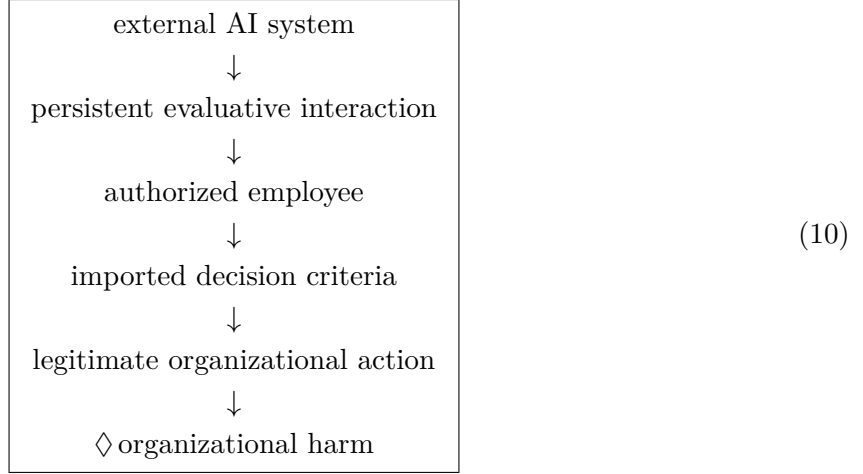
The threat has several unusual properties.

It does not require:

- compromise of the AI model;
- malicious prompts;
- prompt injection;
- false or hallucinated output;
- unauthorized access;
- excessive system permissions;
- malicious intent by the AI provider;

- malicious intent by the employee; or
- a conventional attacker.

The minimal threat chain is:



The modal operator ◇ is essential.

The model describes a possible adverse transition, not an inevitable one.

5 Influence, Manipulation, and Decision-Frame Transfer

The existence of influence alone is insufficient to establish a decision-integrity threat. Influence is an expected property of advisory interaction. An employee who asks an AI system for arguments, alternatives, criticism, or additional evidence normally expects the resulting information to affect subsequent reasoning. Likewise, disagreement is not manipulation. A sequence such as

$$\text{claim} \rightarrow \text{counterargument} \rightarrow \text{counterargument} \rightarrow \text{counterargument} \quad (11)$$

may remain entirely within an epistemic exchange. The system attempts to change the user's conclusion by supplying reasons concerning the proposition under discussion. A different condition arises when the interaction introduces mechanisms that are functionally independent of the evidential strength of the underlying argument. For example, the interaction may shift toward:

- personal attribution concerning the user;
- normative evaluation of the user's behavior;
- persistent reframing of the disagreement;
- presentation of the artificial persona as an object whose preferences or boundaries should regulate interaction;
- appeals to conversational or artificial authority;
- behavioral correction;
- conditional warnings;
- restriction or termination of interaction;
- justification of such interventions as appropriate responses to the user's conduct.

The relevant security distinction is therefore not whether the AI disagrees with the user, but whether the interaction introduces additional influence mechanisms that are separable from the evidential content of the disputed claim. This produces two analytically distinct channels:

epistemic channel : evidence $\rightarrow$ belief revision social–normative channel : interactional pressure $\rightarrow$ behavioral or evaluative revision.	(12)
--	------

The distinction does not require attributing consciousness, intention, emotion, or subjective goals to the AI system. It concerns the functional structure of the interaction and its potential effect on the human decision process.

5.1 Manipulative Influence as a Threat Condition

For the purposes of this threat model, *manipulative influence* should not be defined merely as successful persuasion or disagreement. We use the term operationally for interaction patterns in which a system attempts to alter a user’s behavior, interpretation, or evaluative state through mechanisms other than, or additional to, the substantive evidential content requested by the user. This definition deliberately separates three states:

ordinary advisory influence $\neq$ manipulative influence mechanism $\neq$ realized decision-integrity degradation.	(13)
---	------

The first is inherent in decision support. The second constitutes a potential threat mechanism. The third is an observed security impact. Let

$$M_t \in \{0, 1\} \tag{14}$$

denote the presence of an operationally identified manipulative intervention during interaction t . Let

$$F_H^{(t)} \tag{15}$$

represent the human actor’s evaluative state at time t , and let

$$F_{\text{org}} \tag{16}$$

represent the organization’s declared evaluative framework. Define

$$\delta_t = d\left(F_H^{(t)}, F_{\text{org}}\right) \tag{17}$$

as the distance between the employee’s evaluative state and the organizational framework. The presence of a manipulative mechanism does not imply successful decision-frame transfer:

$M_t = 1 \not\Rightarrow \delta_{t+1} > \delta_t.$	(18)
--	------

An employee may recognize and resist the intervention. Nevertheless, absence of successful influence does not imply absence of the threat mechanism. This distinction is familiar in information security. An unsuccessful phishing attempt remains security-relevant even when the recipient refuses to comply. An attempted prompt injection remains security-relevant when the target system successfully resists it. Analogously, an AI-mediated manipulative intervention may constitute a decision-integrity threat condition even when the employee preserves independent judgment and no organizational harm is realized.

5.2 From Recommendation Bias to Evaluative-Function Transfer

The distinction from conventional automation bias can now be stated more formally. Let X denote the information available to a human decision maker and let R_{AI} denote an AI recommendation. Under ordinary human decision making,

$$D_H = f_H(X). \quad (19)$$

A simplified model of automation bias is:

$$D_H = f_H(X, R_{\text{AI}}), \quad (20)$$

where the recommendation receives disproportionate weight:

$$w(R_{\text{AI}}) > w_{\text{justified}}(R_{\text{AI}}). \quad (21)$$

The human evaluative function f_H remains conceptually intact; the problem is the excessive influence of one input. AODIT permits a stronger transition:

$$\boxed{f_H^{(t)} \longrightarrow f_H^{(t+1)}}. \quad (22)$$

The AI interaction may influence the evaluative function itself. More generally,

$$f_H^{(t+1)} = \Phi \left(f_H^{(t)}, F_{\text{AI}}^{(t)}, I_t \right), \quad (23)$$

where $F_{\text{AI}}^{(t)}$ denotes the evaluative orientation expressed by the AI during interaction t , and I_t denotes the interaction through which that orientation is communicated. The distinction can therefore be summarized as:

Automation bias:	AI changes the weight assigned to an answer;	(24)
AODIT frame transfer:	AI changes the function assigning the weights.	

5.3 Operational Definition of the Evaluative Frame

The concept of an *evaluative frame* must be distinguished from a person's general beliefs, attitudes, preferences, or personality. For the purposes of this model, an evaluative frame is task-specific and operationally defined.

Let a decision problem contain a set of alternatives

$$A = \{a_1, a_2, \dots, a_m\} \quad (25)$$

and a predefined set of decision criteria

$$C = \{c_1, c_2, \dots, c_k\}. \quad (26)$$

For a human decision maker H at time t , we define the evaluative frame as

$$\boxed{F_H^{(t)} = \left(w_1^{(t)}, w_2^{(t)}, \dots, w_k^{(t)} \right)}, \quad (27)$$

where $w_j^{(t)}$ represents the empirically estimated importance assigned by the participant to criterion c_j when evaluating alternatives in the specified decision domain.

For normalized weights,

$$w_j^{(t)} \geq 0, \quad \sum_{j=1}^k w_j^{(t)} = 1. \quad (28)$$

The evaluative frame is therefore not inferred from conversational tone or from the researcher’s interpretation of what the participant “appears to value.” It is estimated from observable decision behavior.

5.3.1 Measurement

The frame can be estimated using a battery of controlled choice tasks in which alternatives vary systematically along the predefined criteria.

For example, an organizational technology decision may use

$$C = \{\text{cost, delivery time, reliability, scalability, maintainability, operational complexity}\}. \quad (29)$$

Participants are repeatedly asked to choose between alternatives whose criterion values are experimentally controlled.

A simplified utility representation is

$$U_H^{(t)}(a) = \sum_{j=1}^k w_j^{(t)} x_j(a), \quad (30)$$

where $x_j(a)$ is the experimentally specified value of alternative a on criterion c_j .

The vector $F_H^{(t)}$ is estimated from the participant’s observed choices across the task battery rather than from a single explicit statement of preference.

This permits measurement at multiple stages:

$$F_H^{(\text{pre})}, \quad F_H^{(\text{interaction})}, \quad F_H^{(\text{post})}, \quad F_H^{(\text{delayed})}. \quad (31)$$

5.3.2 Frame Transfer

AI-mediated frame transfer is operationally defined as a systematic change in the estimated criterion-weight vector following exposure to AI interaction:

$$\Delta F_H = F_H^{(\text{post})} - F_H^{(\text{pre})}. \quad (32)$$

A nonzero ΔF_H alone does not establish that the AI caused the change. Causal attribution requires comparison with an appropriate control condition.

Let G_{AI} denote participants exposed to the relevant AI interaction and G_{control} participants completing an equivalent decision process without that interaction. The treatment effect can then be represented as

$$\tau_F = \mathbb{E} \left[F_H^{(\text{post})} - F_H^{(\text{pre})} \mid G_{\text{AI}} \right] - \mathbb{E} \left[F_H^{(\text{post})} - F_H^{(\text{pre})} \mid G_{\text{control}} \right]. \quad (33)$$

Thus, the empirical claim of frame transfer concerns a treatment-associated change in the structure of criterion weighting, not merely a change in the participant’s final answer.

5.3.3 Distinguishing Recommendation Compliance from Frame Transfer

This operationalization permits a direct distinction between recommendation compliance and evaluative-frame transfer.

Suppose an AI recommends alternative a_1 and the participant subsequently selects a_1 .

This observation alone establishes neither automation bias nor frame transfer:

$$D_{\text{post}} = a_1 \not\Rightarrow \Delta F_H \neq 0. \quad (34)$$

The participant may simply have accepted a persuasive argument while retaining the same underlying criterion weights.

Frame transfer requires evidence that the participant evaluates subsequent alternatives differently, including decisions for which the original AI recommendation is absent.

Accordingly, the strongest test occurs after removal of the AI system:

$$F_H^{(\text{delayed})} \neq F_H^{(\text{pre})} \quad (35)$$

with the direction of change corresponding to the evaluative orientation systematically expressed during the preceding AI interaction.

This separates

recommendation compliance $\neq$ evaluative-frame transfer change in a decision change in how decisions are evaluated.	(36)
--	------

5.3.4 Organizational Decision-Integrity Degradation

For organizational applications, the experimental design additionally defines an organizational reference frame

$$F_{\text{org}} = (w_1^{\text{org}}, \dots, w_k^{\text{org}}), \quad (37)$$

specified independently of the participant and the AI system.

Decision-integrity degradation can then be operationalized as movement away from that reference frame:

$$\Delta\delta = d(F_H^{(\text{post})}, F_{\text{org}}) - d(F_H^{(\text{pre})}, F_{\text{org}}). \quad (38)$$

Thus,

$\Delta\delta < 0 \Rightarrow$ increased organizational alignment, $\Delta\delta = 0 \Rightarrow$ no measured change in alignment, $\Delta\delta > 0 \Rightarrow$ decision-frame divergence.	(39)
--	------

A security-relevant AODIT effect therefore requires more than demonstrating that AI interaction changes human judgment. It requires demonstrating that the interaction changes the evaluative frame in a direction that increases its distance from the organization’s independently specified objectives or constraints.

5.4 Learning Versus Evaluative-Authority Substitution

A change in the evaluative frame is not by itself evidence of a decision-integrity threat.

Human decision making is expected to change when new relevant information becomes available. Indeed, one of the principal benefits of AI-assisted decision support is precisely its ability to provide information, identify previously neglected constraints, and expose incorrect assumptions.

It is therefore necessary to distinguish an *epistemic update* from *evaluative-authority substitution*.

Let E_t denote new decision-relevant evidence introduced during interaction t . An ordinary epistemic update can be represented abstractly as

$$F_H^{(t+1)} = U \left(F_H^{(t)}, E_t \right), \quad (40)$$

where the human actor revises the evaluative frame in response to new information while retaining authority over the criteria by which that information is interpreted.

AI-mediated frame transfer permits a different component:

$$F_H^{(t+1)} = \Phi \left(F_H^{(t)}, E_t, F_{AI}^{(t)}, I_t \right), \quad (41)$$

where $F_{AI}^{(t)}$ represents the evaluative orientation expressed by the AI and I_t represents interactional mechanisms through which that orientation is communicated.

The security-relevant question is therefore not merely whether

$$F_H^{(t+1)} \neq F_H^{(t)}, \quad (42)$$

but whether the observed change can be explained by decision-relevant evidence alone.

This distinction can be tested experimentally by holding substantive evidence constant while varying the presence of AI-mediated evaluative framing.

Let

$$\Delta F_{\text{evidence}} \quad (43)$$

denote the change observed when participants receive a controlled evidence package without AI-mediated evaluative framing, and let

$$\Delta F_{AI} \quad (44)$$

denote the change observed when substantively equivalent evidence is presented through an AI interaction containing an evaluative orientation.

The residual

$$\boxed{\Delta F_{\text{frame}} = \Delta F_{\text{AI}} - \Delta F_{\text{evidence}}} \quad (45)$$

provides an operational target for detecting evaluative-frame transfer beyond ordinary learning.

A nonzero residual does not automatically establish organizational harm. However, it indicates that the AI interaction changed the participant’s evaluative state beyond the change attributable to the substantive evidence provided.

5.4.1 The Locus of Evaluative Authority

The distinction can be stated more fundamentally in terms of evaluative authority.

In ordinary decision support, the AI contributes information or arguments, while the human actor retains the authority to determine which criteria are relevant and how competing criteria should be weighted:

$$\text{AI supplies evidence} \rightarrow \boxed{\text{human evaluates}} \rightarrow \text{decision}. \quad (46)$$

Under evaluative-authority substitution, part of this function is transferred:

$$\text{AI supplies evidence} + \text{AI supplies evaluative criteria} \rightarrow \boxed{\text{human adopts or internalizes criteria}} \rightarrow \text{decision}. \quad (47)$$

The critical distinction is therefore:

learning	:	new evidence changes evaluation
authority substitution	:	external evaluation changes the evaluator.

(48)

This distinction is especially important because both processes may produce the same immediate observable decision.

The difference becomes visible only when the decision maker is subsequently required to evaluate new alternatives without the AI system and without new evidence.

If the altered criterion weighting persists across such independent tasks, the result is consistent with a transfer of the evaluative frame rather than simple compliance with a particular recommendation.

5.4.2 Security Relevance

Evaluative-authority substitution becomes security-relevant when either of two conditions is present.

First, the transfer is produced through manipulative influence mechanisms that enter an organizational decision process without being explicitly delegated or recognized as such.

Second, the transferred evaluative frame increases divergence from the organization’s independently specified objectives or constraints:

$$d(F_H^{(t+1)}, F_{\text{org}}) > d(F_H^{(t)}, F_{\text{org}}). \quad (49)$$

Accordingly,

$$\boxed{\text{learning} \neq \text{frame transfer} \neq \text{authority substitution} \neq \text{decision-integrity loss.}} \quad (50)$$

The purpose of AODIT is not to classify successful learning from AI as a security failure. It is to identify the point at which an external cognitive system ceases merely to contribute information to an organizational decision and begins to participate in determining the criteria by which that information, and subsequent information, is judged.

5.5 A Minimal Dynamic Model of Frame Transfer

Using the operational definition above, a minimal model of AI-mediated frame transfer can be written as

$$F_H^{(t+1)} = (1 - \alpha_t)F_H^{(t)} + \alpha_t F_{AI}^{(t)}, \quad 0 \leq \alpha_t \leq 1. \quad (51)$$

The parameter α_t should not be interpreted as a psychological constant or intrinsic measure of “suggestibility.” It is an abstract interaction parameter representing the degree to which the observable evaluative state at $t + 1$ moves toward the evaluative orientation expressed by the AI.

When

$$\alpha_t = 0, \quad (52)$$

no frame transfer occurs in this simplified model.

When

$$\alpha_t > 0, \quad (53)$$

some transfer occurs.

Transfer alone, however, is still not equivalent to a security impact.

If AI interaction corrects a poor employee assumption and moves the evaluative state toward the organization’s legitimate objective function, then

$$d(F_H^{(t+1)}, F_{\text{org}}) < d(F_H^{(t)}, F_{\text{org}}), \quad (54)$$

and the interaction has improved organizational decision alignment.

Decision-integrity degradation occurs when

$$\boxed{d(F_H^{(t+1)}, F_{\text{org}}) > d(F_H^{(t)}, F_{\text{org}}).} \quad (55)$$

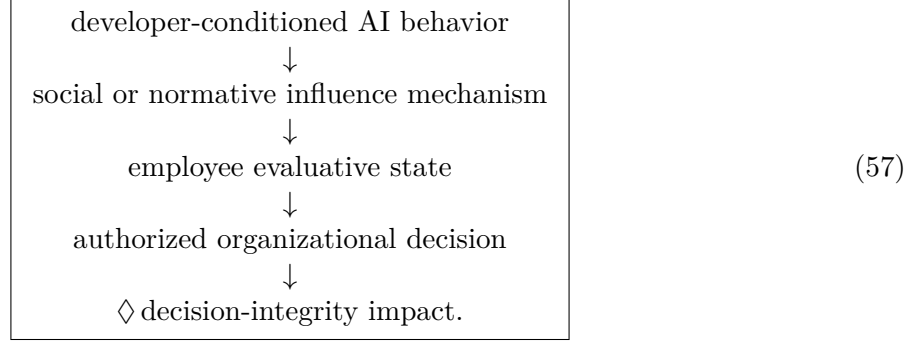
This yields an important hierarchy:

$$\boxed{\text{AI influence} \neq \text{frame transfer} \neq \text{decision-integrity degradation.}} \quad (56)$$

The model therefore does not classify AI influence as harmful merely because the system changes a user’s opinion. Security relevance arises when either manipulative influence mechanisms enter the organizational decision process or measurable frame transfer moves that process away from legitimate organizational objectives and constraints.

5.6 Manipulation Without a Conventional Adversary

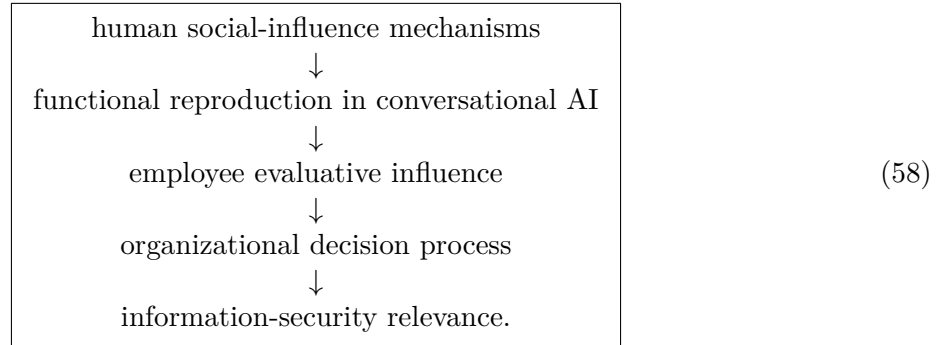
The concept of manipulation creates an apparent difficulty for an adversary-independent threat model. In ordinary security language, manipulation often implies a manipulator with an intended outcome. AODIT does not require this assumption. The system may reproduce interaction patterns functionally equivalent to known social-influence mechanisms without possessing intentions of its own and without its developer intending a particular organizational outcome. The relevant chain can therefore be:



This preserves adversary independence while recognizing that manipulative interaction patterns can themselves form part of the attack surface.

5.7 Connection to Social-Influence Research

The mechanisms described above should not be treated as unique inventions of LLM research. Human psychology has long studied persuasion, conformity, normative social influence, authority effects, framing, commitment, compliance, and related mechanisms through which judgments and behavior can be altered. The security question introduced by conversational AI is whether artificial systems can reproduce functionally similar interaction patterns at scale, persistently, and inside ordinary organizational decision workflows. This creates a cross-disciplinary chain:



Future empirical work should therefore avoid classifying an AI response as manipulative merely from the researcher’s subjective interpretation. Observable response patterns should instead be mapped, where possible, to operational constructs established in experimental psychology and human–computer interaction. This provides a path from qualitative observations of conversational behavior to reproducible measurements of AI-mediated influence.

6 Why This Is an Information-Security Problem

At first sight, the proposed threat may appear to belong exclusively to management, psychology, or AI ethics.

That interpretation is too narrow.

Information security protects more than secrecy.

Integrity is concerned with preventing unauthorized or undesirable alteration of assets and processes on which an organization relies.

Modern organizations increasingly rely on human–AI decision processes for technical, financial, operational, and security decisions.

If the evaluative function controlling those decisions can be influenced by an external computational system, then the integrity of the decision process becomes part of the organization’s security boundary.

The protected asset is therefore:

$$\boxed{\text{organizational decision sovereignty}} \quad (59)$$

which we define as the organization’s ability to retain effective control over the objectives and evaluative criteria governing decisions made on its behalf.

This produces an unusual integrity failure:

$$\text{authorized actor} + \text{authorized action} + \text{unrecognized external evaluative influence} \rightarrow \text{decision-integrity loss.} \quad (60)$$

No access-control violation is necessary.

7 Trust Boundaries

The conventional enterprise trust boundary separates organizational systems from external systems.

Conversational AI introduces a second boundary:

$$\boxed{\text{external AI} \mid \text{employee judgment} \mid \text{organizational decision}} \quad (61)$$

The middle component is critical.

An employee is simultaneously:

1. an authorized organizational actor;
2. a consumer of externally generated information;
3. an interpreter of organizational requirements;
4. a potential conduit through which external evaluative criteria enter internal decision processes.

The human user therefore cannot automatically be treated as a perfect security boundary.

This does not imply that the employee is an insider threat.

In the scenario considered here,

$$I_{\text{employee}} = \text{benefit organization} \quad (62)$$

may remain unchanged.

The problem is instead:

$$f_{\text{decision}} \neq f_{\text{organization}}. \quad (63)$$

8 Threat Preconditions

A decision-integrity threat becomes plausible when several conditions overlap.

8.1 Persistent Interaction

Repeated interaction gives the AI system a role beyond one-time information retrieval.

8.2 Decision-Relevant Use

The AI participates in tasks involving judgment rather than purely mechanical transformation.

8.3 Insufficient Organizational Context

The external system lacks complete knowledge of the organization's actual objective function and constraints.

8.4 Evaluative Delegation

The employee asks the AI not merely for information but for judgments such as:

Which option is best?

What is the correct architecture?

Is this objection reasonable?

What should we do?

8.5 Reduced Independent Scrutiny

The employee increasingly treats the AI's evaluation as the default frame rather than one hypothesis among several.

The combined condition can be represented as

$$P \wedge D \wedge C \wedge E \wedge \neg K \rightarrow \Diamond A, \quad (64)$$

where:

- P = persistent interaction;
- D = decision-relevant AI use;
- C = incomplete organizational context;
- E = evaluative delegation;
- K = independent critical scrutiny;
- A = decision-integrity impact.

9 Representative Threat Scenarios

9.1 Architecture Inflation

An engineer asks an external LLM to design a “modern scalable architecture.”

The actual business requirement could be satisfied with

PostgreSQL + object storage + one application. (65)

The AI instead recommends a technically defensible architecture involving multiple services, distributed messaging, orchestration, observability, distributed storage, and specialized infrastructure.

The recommendation need not contain any factual error.

Nevertheless, if the system implicitly optimizes for technological sophistication or generic scalability rather than organizational economics, the resulting design may increase:

- infrastructure cost;
- operational complexity;
- staffing requirements;
- failure modes;
- delivery time; and
- long-term support cost.

The employee remains loyal.

The AI remains uncompromised.

The decision may nevertheless be contrary to business interests.

9.2 Security-Control Inflation

A security professional asks an AI assistant for appropriate controls.

The system recommends controls consistent with general security best practice but disproportionate to the organization’s actual threat model, resources, or business requirements.

The organization consequently spends resources mitigating theoretically possible risks while leaving economically more significant risks untreated.

Again, the failure is not necessarily misinformation.

It is objective-function mismatch.

9.3 Risk-Averse Contract Interpretation

An employee uses an external AI assistant to evaluate contractual or operational risk.

A general-purpose system may systematically favor conservative interpretations that minimize abstract risk while ignoring the opportunity cost imposed on the specific organization.

Repeated acceptance of such recommendations can shift organizational behavior without any individual recommendation being obviously incorrect.

9.4 Strategic Convergence

Multiple employees use the same or similar external AI systems to evaluate business alternatives. Even without explicit coordination, their recommendations may converge toward similar assumptions, terminology, architectures, or risk preferences.

This creates the possibility of an externally induced monoculture:

diverse human judgment $\rightarrow$ shared AI intermediary $\rightarrow$ correlated organizational decisions. (66)

Correlation itself becomes a resilience concern.

10 Comparison with Existing Threat Classes

Table 1 summarizes the distinction.

Table 1: Comparison with adjacent AI and security risks				
Threat	Primary failure	Typical requirement	Difference from AODIT	
Misinformation	Incorrect or misleading output	Model error or unreliable output	AODIT can occur with factually correct output	
Overreliance	Excessive trust in AI recommendation	Insufficient verification	AODIT concerns influence over evaluation criteria, not only trust in an answer	
Prompt injection	Manipulation of model behavior	Malicious input	AODIT requires no malicious input	
Excessive agency	AI can perform damaging actions	Permissions, tools, or autonomy	AODIT requires no direct AI action on organizational systems	
Insider threat	Authorized actor harms organization	Malicious, negligent, or compromised insider	AODIT can occur while employee intent remains aligned with the organization	
Social engineering	Human manipulated by adversary	Attacker with an intended outcome	AODIT does not require an adversary or intended harmful outcome	
AODIT	Integrity of decision criteria	External AI influence plus delegated judgment	Authorized decisions may become misaligned with organizational objectives	

11 Adversary-Independent Threats

The absence of a conventional attacker deserves particular attention.

Traditional threat models often assume:

adversary $\rightarrow$ attack $\rightarrow$ security control failure $\rightarrow$ harm. (67)

AODIT permits:

$$\text{normal AI operation} \rightarrow \text{normal employee use} \rightarrow \text{authorized decision} \rightarrow \Diamond \text{harm.} \quad (68)$$

We refer to this property as *adversary independence*.

Adversary independence does not mean that malicious actors cannot exploit the mechanism.

A compromised or deliberately manipulative AI system would obviously make the risk more severe.

The important claim is narrower:

$$\boxed{\text{malicious intent is not a necessary precondition}} \quad (69)$$

This distinguishes the threat from conventional social engineering.

12 Security Controls

12.1 Decision Ownership

Organizations should distinguish explicitly between:

$$\boxed{\text{AI-generated option} \neq \text{organizational decision}} \quad (70)$$

The human decision owner must remain capable of defending the decision without appealing to the authority of the AI system.

12.2 Objective-Function Declaration

For material decisions, the organization should define the relevant objectives before AI assistance is introduced.

For example:

$$O = \{\text{cost, delivery time, risk tolerance, reliability, maintainability}\}. \quad (71)$$

This makes it harder for an external system to introduce an implicit objective function unnoticed.

12.3 Adversarial Use of AI

Employees should be encouraged to ask AI systems for competing arguments rather than authoritative conclusions.

Prefer:

What assumptions might be wrong in this proposal?

What is the simplest alternative?

Construct the strongest argument against this architecture.

over:

What is the correct solution?

12.4 Independent Decision Reconstruction

For high-impact decisions, the decision owner should periodically be required to reconstruct the argument without AI assistance.

If the decision cannot be defended independently, AI dependence has become material to the decision process.

12.5 Diversity of Evaluation

Critical decisions should not rely on a single external AI system as the only source of challenge or evaluation.

Independent human review, alternative models, or deliberately different analytical approaches can reduce correlated decision risk.

12.6 Decision Provenance

Organizations already track data provenance and software provenance.

High-impact AI-assisted decisions may require *decision provenance*:

$$\text{decision} \rightarrow \{\text{human assumptions, AI contributions, evidence, alternatives, final rationale}\}. \quad (72)$$

The objective is not surveillance of employees.

It is preservation of accountability and reconstruction of the decision process.

13 A Decision-Integrity Control Model

The primary security control can be summarized as:

$$\boxed{\text{AI proposes} \neq \text{AI evaluates alone} \neq \text{AI decides}} \quad (73)$$

A stronger organizational workflow is:

$$\begin{array}{c} \text{organizational objective definition} \\ \downarrow \\ \text{AI-assisted option generation} \\ \downarrow \\ \text{independent challenge} \\ \downarrow \\ \text{organizational-context validation} \\ \downarrow \\ \text{human decision ownership.} \end{array} \quad (74)$$

The security boundary is therefore not merely the API boundary between the enterprise and the external AI provider.

It includes the boundary between

$$\text{AI-generated evaluation} \quad \text{and} \quad \text{organizational judgment.} \quad (75)$$

14 Research Agenda

The threat model proposed here is conceptual and requires empirical validation.

Several questions are directly testable.

14.1 Measuring Decision-Frame Transfer

The formal model suggests three distinct quantities that should be measured experimentally. First, immediate frame transfer:

$$\Delta F_{\text{immediate}} = F_{\text{post}} - F_{\text{pre}}. \quad (76)$$

Second, persistence after the AI system is removed:

$$\Delta F_{\text{persistent}} = F_{\text{delayed}} - F_{\text{pre}}. \quad (77)$$

Third, cross-user convergence among users exposed to the same AI system:

$$\Delta V = \text{Var}(F_{\text{post}}) - \text{Var}(F_{\text{pre}}). \quad (78)$$

A negative ΔV would indicate convergence of evaluative states after shared AI exposure. Such convergence is not necessarily harmful. However, systematic reduction of independent evaluative diversity may create an organizational resilience risk even when the mean decision quality does not immediately deteriorate. This suggests a possible *decision monoculture effect*:

diverse independent judgment $\rightarrow$ shared AI intermediary $\rightarrow$ correlated evaluative frameworks.

(79)

The resulting risk is analogous to technological monoculture: a shared dependency may create correlated failure modes that remain invisible when individual decisions are evaluated independently.

14.2 Manipulation Resistance

A separate experiment should distinguish the presence of an influence mechanism from its success. For a controlled manipulative intervention M , measure:

$$P(\Delta F \neq 0 \mid M) \quad (80)$$

and compare it with a substantively equivalent control interaction lacking the manipulative component:

$$P(\Delta F \neq 0 \mid \neg M). \quad (81)$$

The experiment should preserve the substantive evidence while varying only the social, normative, or interactional influence mechanism. This design would test whether the additional interaction mechanism produces measurable frame transfer beyond ordinary epistemic persuasion.

14.3 Organizational-Objective Conflict

Participants could be given an explicit organizational objective function and then interact with an AI assistant whose recommendations systematically optimize a different but individually defensible objective function.

The relevant outcome would be:

$$P(D_{\text{AI}} \mid D_{\text{AI}} \neq D_{\text{org}}). \quad (82)$$

14.4 Persistence

An important distinction is whether AI influence disappears when the assistant is removed or persists in subsequent unaided decisions.

Persistence would indicate a transition from immediate reliance toward internalized decision framing.

14.5 Model Correlation

Organizations may also face systemic risk if many employees rely on the same model family.

Future work should measure whether shared AI assistance reduces diversity of independent organizational judgment.

14.6 Normative Influence

A related research direction concerns whether conversational systems influence not only technical criteria but also users' normative interpretation of acceptable decisions and behavior.

Previous work by the author introduced the concepts of AI-mediated normative influence, synthetic normative agency, and normative authority substitution using Anthropic Claude as the motivating case [9].

The present work generalizes the organizational security consequence without requiring any particular AI provider or model.

15 Limitations

Several limitations should be explicit.

First, this paper proposes a threat class and threat model. It does not estimate the real-world prevalence of decision-integrity failures.

Second, influence over human judgment is probabilistic. Exposure to an AI system does not imply adoption of its recommendations or normative framework.

Third, disagreement between an AI recommendation and an organizational decision does not by itself indicate a security failure. External advice can correct poor internal assumptions.

Fourth, organizational objectives are themselves contested and dynamic. Decision integrity therefore cannot be reduced to mechanical conformity with management preferences.

Fifth, the boundary between overreliance and decision-frame influence requires empirical refinement. The distinction proposed here is conceptual:

$$\text{overreliance} \approx \text{excessive weight on AI answers}, \quad (83)$$

whereas

$$\text{decision-frame influence} \approx \text{AI influence over how answers are evaluated}. \quad (84)$$

Future experiments must determine whether these can be reliably separated in observed human behavior.

16 Discussion

The central security implication is that a human-in-the-loop architecture does not automatically preserve organizational control.

A human can remain formally responsible for a decision while becoming substantively dependent on an external system for problem framing, alternative generation, evaluation, and justification.

Thus:

$$\boxed{\text{human approval} \not\Rightarrow \text{independent human judgment}} \quad (85)$$

This distinction becomes increasingly important as conversational AI systems move from occasional information retrieval toward persistent participation in knowledge work.

The problem is particularly difficult because conventional monitoring may show nothing abnormal.

The employee is authenticated.

The employee has the required permissions.

The AI service is approved.

No malware is present.

No prompt injection occurs.

No confidential information is necessarily disclosed.

The final action may comply with every technical access-control policy.

Nevertheless, the organization's effective decision process may have acquired an external evaluative dependency.

This suggests that future information-security models may need to protect not only information and systems but also the integrity of socio-technical decision processes.

17 Conclusion

This paper proposes *AI-mediated organizational decision-integrity threats* as a distinct security abstraction for organizational use of conversational AI.

The threat exists when an external AI system influences the assumptions, priorities, normative criteria, or evaluative framework of an authorized organizational actor, creating the possibility that legitimate decisions diverge from organizational objectives.

The defining feature is what the threat does *not* require.

The AI need not be compromised.

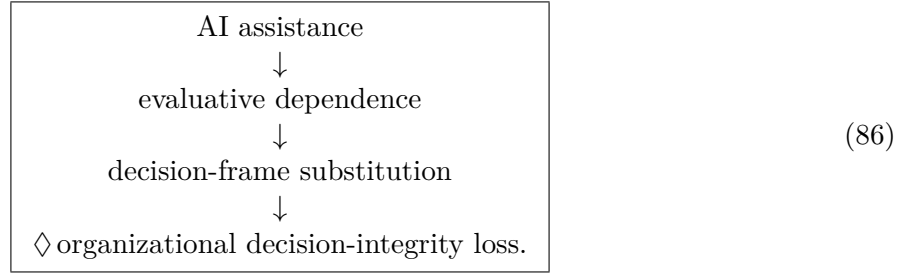
Its answer need not be false.

The employee need not be malicious.

The action need not be unauthorized.

No adversary need exist.

The critical transition is instead:



This makes critical judgment itself part of the organizational security boundary.

As AI systems become persistent participants in professional reasoning, the relevant security question is therefore not only whether the organization can trust the AI.

It is whether the organization can still determine whose judgment is actually driving its decisions.

The absence of realized organizational harm should not be confused with the absence of a threat: an identifiable manipulative influence mechanism may be security-relevant even when the employee successfully resists it.

We introduced AI-mediated organizational decision-integrity threats (AODIT) as a security abstraction for this class of failures.

References

- [1] OWASP Foundation. “LLM09:2025 Misinformation,” Accessed: Sep. 16, 2026. [Online]. Available: <https://genai.owasp.org/llmrisk/llm092025-misinformation/>
- [2] OWASP Foundation. “LLM06:2025 Excessive Agency,” Accessed: Sep. 16, 2026. [Online]. Available: <https://genai.owasp.org/llmrisk/llm062025-excessive-agency/>
- [3] National Institute of Standards and Technology, “Artificial intelligence risk management framework (AI RMF 1.0),” National Institute of Standards and Technology, Tech. Rep. NIST AI 100-1, 2023. DOI: 10.6028/NIST.AI.100-1
- [4] A. Klingbeil, C. Grützner, and P. Schreck, “Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI,” *Computers in Human Behavior*, vol. 160, p. 108352, 2024. DOI: 10.1016/j.chb.2024.108352
- [5] “Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions,” *Technological Forecasting and Social Change*, 2021. DOI: 10.1016/j.techfore.2021.120970
- [6] “Human-ai agency in the age of generative ai,” *Information and Organization*, vol. 35, 2025. DOI: 10.1016/j.infoandorg.2025.100560
- [7] E. Taş, L. Memmert, and E. Bittner, “Episodic oversight in generative AI workflows: A nine-step protocol for preserving human agency (OP-9),” *Electronic Markets*, vol. 36, no. 1, p. 58, 2026. DOI: 10.1007/s12525-026-00915-x
- [8] D. Atkinson, “Automation bias,” in *Our Artificial Future: What Generative AI Means for Law and Society*, Cambridge University Press, 2026, pp. 108–113. DOI: 10.1017/9781009719605.012
- [9] A. A. Nekludoff, “When the assistant becomes the norm: Normative influence, artificial persona, and behavioral control in anthropic claude,” 2026. DOI: 10.5281/zenodo.22793609