Moiré: Format-Preserving Deception Middleware with Convergence Analysis

David Kirsch
Independent Researcher
Simi Valley, California
david@davidkirsch.me
https://davidkirsch.me

June 2026

Abstract

We present Moiré, a construction that applies format-preserving encryption (NIST SP 800-38G, FF1) as a real-time deception layer between a production database and its application interface. The middleware produces a structurally faithful but specifically wrong version of reality: statistical patterns, inter-record distances, and aggregate counts survive the transformation while every individual value changes. Unlike honeypots, which construct isolated decoy environments, Moiré operates on the production system—making it effective against insiders who would recognize synthetic data. We make three contributions. *First*, we define the Moiré construction with data-type-specific transformations, a bidirectional query-mapping requirement, and session-keyed attribution chains, and we analyze its security against three tiers of adversary knowledge—including a novel formalization of the bidirectional mapping as a chosen-plaintext oracle. *Second*, we prove a convergence theorem: the viable distortion space shrinks monotonically with adversary side information, and all full-environment deception converges to selective distortion with tracers against sufficiently informed adversaries. This formalizes a principle long understood in intelligence tradecraft but never given a mathematical treatment. *Third*, we derive a detection bound from the steganographic security model of Cachin (1998): the Kullback-Leibler divergence between real and distorted data distributions upper-bounds any passive adversary's distinguishing advantage. We further introduce *adaptive distortion*—a mechanism that increases distortion intensity in response to detected verification behavior—and characterize the resulting probing-deterrence equilibrium. Finally, we propose a graduated deception architecture (five tiers from passive monitoring through full distortion to containment) and formalize the turned-asset doctrine: using the distortion layer to run compromised insiders as unwitting intelligence sources.

Keywords: deception technology, format-preserving encryption, insider threat, steganographic security, convergence analysis, game-theoretic deception, adaptive distortion

Contents

1 Introduction

Deception technology has been part of the security toolkit for decades. Honeypots detect intrusions by presenting fake systems that attackers explore [22, 23]. Honeytokens trip alarms when touched [24]. MITRE's Engage framework structures adversary engagement into Expose, Affect, and Elicit operations [14]. Commercial platforms deploy decoy servers, fake credentials, and synthetic data environments across enterprise networks [21, 5, 1, 17].

All of these share a limitation: they are designed to fool outsiders. An attacker who has never seen the real system cannot distinguish a decoy from production. But an insider—someone with legitimate access who knows the real data—will recognize a static decoy immediately. The record count is wrong. The activity patterns do not match. A record they know exists is missing. The deception fails at the moment it matters most.

This paper presents MOIRÉ, a construction that addresses the insider-deception problem by applying format-preserving encryption [2, 15] as a real-time middleware layer. Instead of constructing a fake environment with synthetic data, the system runs the real application against the real database but passes every data point through a cryptographic distortion layer before it reaches the screen. The insider sees real patterns, real activity volumes, real statistical distributions. Every specific value—every name, coordinate, identifier, timestamp—is wrong.

1.1 Distinction from honey encryption

Juels and Ristenpart [10] introduced honey encryption (HE), which produces ciphertexts that, when decrypted with incorrect keys, yield "plausible-looking but bogus plaintexts called honey messages." The HE paper notes a connection to format-preserving encryption and draws inspiration from decoy systems. The surface language is close enough to MOIRÉ that the distinction must be stated precisely.

MOIRÉ differs from honey encryption in six structural respects. (i) *Trigger*: HE activates when an adversary applies the *wrong* decryption key; MOIRÉ activates when an adversary uses the *right* credentials but is routed through a distortion session. (ii) *Security goal*: HE protects *confidentiality* against offline brute-force attacks; MOIRÉ provides *deception* against online data access by authorized insiders. (iii) *Format*: HE ciphertexts are unstructured bit strings; MOIRÉ output is format-preserving by definition—a license plate distorts to a valid license plate. (iv) *Structural fidelity*: HE produces plausible individual messages; MOIRÉ preserves patterns, distances, aggregates, and statistical distributions across entire datasets. (v) *Adaptivity*: HE is static; MOIRÉ is session-keyed and supports adaptive distortion (Section 7). (vi) *Attribution*: HE provides no attribution chain; MOIRÉ session keys enable forensic tracing from exfiltrated data to the exact breach vector.

1.2 Contributions

This paper makes three contributions.

First, we define the MOIRÉ construction with data-type-specific transformations, a bidirectional query-mapping requirement, and session-keyed attribution chains. We analyze its security in a formal indistinguishability game (Definition 4.1), identify the bidirectional mapping as a chosen-plaintext oracle (Section 4.4), and derive design constraints from known attacks on FF3 [6].

Second, we prove a convergence theorem (Theorem 5.3): the viable distortion space shrinks monotonically with adversary side information, and all full-environment deception converges to selective distortion with tracers. This formalizes a principle from intelligence tradecraft—Wright's barium meal [25] is the equilibrium, and we prove why.

Third, we derive a detection bound (Theorem 4.4) from Cachin's information-theoretic steganographic security model [3]: the KL divergence between real and distorted distributions

upper-bounds any passive adversary's distinguishing advantage. We further introduce adaptive distortion and characterize the probing-deterrence equilibrium (Theorem 7.3).

2 Related Work

2.1 Format-preserving encryption

Bellare, Rogaway, and Spies formalized FPE in 2009 [2]. NIST standardized two modes—FF1 and FF3-1—in SP 800-38G [15]. FPE encrypts data while preserving its format and domain: a 16-digit number encrypts to a different 16-digit number, a date to a valid date. The primary commercial application is PCI-DSS compliant tokenization of payment card numbers.

Durak and Vaudenay [6] demonstrated practical attacks against FF3 on small domains, requiring $O(N^{11/6})$ chosen plaintexts. Following these attacks, NIST recommended domains of at least 10^6 elements and issued the FF3-1 revision. MOIRÉ specifies FF1 exclusively and enforces the minimum domain constraint.

2.2 Deception technology

The field traces from Stoll's *Cuckoo's Egg* (1989) through Cohen's Deception Toolkit (1997) and Spitzner's Honeynet Project (1999) to the current commercial landscape [22, 23, 4]. MITRE's Engage framework [14] structures adversary engagement operations. Heckman et al. [7] adapted Whaley's classical deception model to cybersecurity. Kahlhofer and Rass [11] surveyed application-layer deception specifically and identified the subfield as underdeveloped, a finding reinforced by their subsequent Koney framework for Kubernetes deception orchestration [12].

Dynamic data masking (Oracle Data Redaction, SQL Server DDM) serves different data views based on user role. The mechanism is similar to MOIRÉ, but the intent differs: masking protects privacy; MOIRÉ actively deceives an adversary who believes they have legitimate, unmasked access.

2.3 Game-theoretic cyber deception

Huang and Zhu [9] introduced duplicity games for deception design with application to insider threat mitigation. Horák, Zhu, and Bošanský [8] modeled belief manipulation through dynamic games, deriving optimal strategies that stabilize the adversary's belief at a target value. Pawlick, Colbert, and Zhu [16] analyzed leaky deception using signaling games with evidence. Ye et al. [26] demonstrated that differential privacy mechanisms can serve deception goals, obfuscating network configurations to mislead Bayesian-inference attackers.

Kim et al. [13] drew a crisp distinction between privacy and deception: "differential privacy seeks to induce uncertainty in an observer, whereas the deception mechanism we develop aims to make the adversary confidently draw the incorrect conclusion that no deception is present." MOIRÉ is a deception mechanism in this sense—the adversary should confidently believe they have real data.

2.4 Steganographic security

Cachin [3] proposed an information-theoretic model for steganographic security, quantifying security by the KL divergence between the cover distribution P_C and the stego distribution P_S. A stegosystem is ε-secure when $D_{\mathrm{KL}}(P_C \| P_S) \leq \varepsilon$ and perfectly secure when $\varepsilon = 0$. We apply this framework to MOIRÉ in Section 4.7, treating the real data distribution as the cover and the distorted distribution as the stego object.

2.5 Intelligence tradecraft

Peter Wright documented the "barium meal" technique—distributing uniquely varied intelligence to suspected moles—in *Spycatcher* (1987) [25]. Tom Clancy coined "canary trap" for the same concept. Thinkst operationalized a digital version with CanaryTokens [24]. MOIRÉ session keys formalize the barium meal with cryptographic rigor: each session produces a uniquely keyed distortion, enabling provable attribution when distorted data surfaces externally.

3 The MOIRÉ Construction

3.1 Architecture

A middleware layer M sits between a production database $\mathcal{D}$ and the application's rendering layer. Application code is unmodified—same queries, same components, same business logic. Query results pass through a distortion function $\mathsf{Distort}_K$ parameterized by a session-specific cryptographic key K.

Definition 3.1 (MOIRÉ Middleware)**.** A MOIRÉ middleware $M = (\mathsf{KeyGen}, \mathsf{Distort}, \mathsf{Distort}^{-1})$ consists of:

- $\mathsf{KeyGen}(1^\lambda) \to K$: generates a session key of security parameter λ.
- $\mathsf{Distort}_K : \mathcal{R} \to \mathcal{R}$: a deterministic, format-preserving, bijective distortion function on the record space $\mathcal{R}$.
- $\mathsf{Distort}_K^{-1} : \mathcal{R} \to \mathcal{R}$: the inverse mapping, used for bidirectional query handling.

The distortion function satisfies five properties. It is *deterministic*: the same input and key always produce the same output, ensuring cross-view consistency. It is *structurally preserving*: patterns, relationships, clustering, and aggregate counts survive. It is *format-preserving*: a license plate encrypts to a valid plate, a GPS coordinate maps to a real location. It is *session-keyed*: different sessions produce different distortions. And it is *computationally irreversible* without K.

3.2 Bidirectional mapping requirement

The middleware must intercept both query results (real → distorted) *and* query inputs (distorted → real). Without bidirectional mapping, a user who searches for a distorted value they observed on screen receives empty results from the real database—an immediate tell. The middleware applies $\mathsf{Distort}_K^{-1}$ to search terms before querying and $\mathsf{Distort}_K$ to results before rendering.

This requirement creates a security concern analyzed in Section 4.4: the bidirectional mapping functions as a chosen-plaintext oracle.

3.3 Data-type-specific transformations

Table 1: Data-type-specific distortion transformations.

Data Type	Transformation	Preserved	Changed
GPS coordinates	Rigid-body rotation + translation	Relative distances, clustering, corridors	Absolute locations
Identifiers	FPE (FF1 mode), keyed bijection	Format, cross-view consistency	Every specific value
Timestamps	Uniform offset derived from K	Sequence, gaps, temporal clustering	Absolute dates
Categorical fields	Keyed substitution within equivalence class	Category structure	Specific values
Names, aliases	Keyed dictionary substitution	Format, frequency profile	Every name
Aggregates, counts	Pass-through	Everything	Nothing

The coordinate rotation deserves special note. Random noise on individual coordinates destroys clustering and is statistically detectable. Instead, a single rigid-body transformation (rotation by angle θ plus translation $[\delta_x, \delta_y]$) is applied to all coordinates simultaneously, with parameters derived from K via HKDF-SHA256. Two points 200 meters apart remain 200 meters apart. A corridor pattern stays a corridor—in a different part of the city.

Formally, let $\mathbf{c} = (c_x, c_y)$ be a coordinate pair. The distorted coordinate is:

$$\mathsf{Distort}_K(\mathbf{c}) = R_\theta\,\mathbf{c} + \boldsymbol{\delta}, \quad R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad (\theta, \boldsymbol{\delta}) = \mathsf{HKDF}(K, \texttt{"coord"}).$$

This transformation preserves all inter-point distances ($\|R_\theta \mathbf{c}_1 + \boldsymbol{\delta} - R_\theta \mathbf{c}_2 - \boldsymbol{\delta}\| = \|R_\theta(\mathbf{c}_1 - \mathbf{c}_2)\| = \|\mathbf{c}_1 - \mathbf{c}_2\|$), clustering structure (since distance preservation implies cluster preservation), and directional relationships (bearing changes are uniform θ-rotations). What changes: every absolute position. A cluster centered at a downtown intersection maps to a different city entirely.

For identifiers, FF1 operates on the digit string as a Feistel network with round functions derived from AES-128 keyed by K. A 10-digit identifier encrypts to a different 10-digit identifier; the mapping is bijective and deterministic given K, ensuring cross-view consistency: the same real identifier always produces the same distorted identifier within a session, and different sessions (different K) produce different distortions.

For names, the substitution dictionary must be calibrated to match the real-world frequency distribution. If "Smith" appears with probability 0.024 in real data, the substitution target must also appear with probability ≈ 0.024. Failure to match the Zipfian distribution of names is a statistical fingerprinting surface (see Section 9).

3.4 The attribution chain

Each session uses a unique key K_s. When distorted data surfaces in the physical world—an adversary acts on a distorted coordinate, publishes a distorted identifier—the operator runs each session key against the surfaced values. The key that reverse-maps distorted values to real values identifies the exact session, credential, and breach vector. This is Wright's barium meal [25] formalized with cryptographic rigor.

3.5 Design constraints from known FPE attacks

Following Durak and Vaudenay's attacks on FF3 [6], MOIRÉ enforces two constraints: (i) all FPE operations use FF1, not FF3 or FF3-1; (ii) all encrypted domains contain at least 10^6 elements, per NIST's post-attack recommendation. Identifier fields with smaller domains (e.g., two-letter state codes) use keyed substitution tables rather than FPE.

4 Security Analysis

4.1 The MOIRÉ indistinguishability game

Definition 4.1 (MOIRÉ-Indistinguishability)**.** The IND-MOIRÉ game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$ proceeds as follows:

1. $\mathcal{C}$ generates $K \leftarrow \mathsf{KeyGen}(1^\lambda)$ and samples $b \leftarrow \{0,1\}$.
2. $\mathcal{A}$ makes adaptive queries $q_1, q_2, \ldots, q_n$ to the system. For each query q_i:
   - If $b = 0$ (real mode): $\mathcal{C}$ returns $\mathcal{D}(q_i)$.
   - If $b = 1$ (MOIRÉ mode): $\mathcal{C}$ returns $\mathsf{Distort}_K(\mathcal{D}(\mathsf{Distort}_K^{-1}(q_i)))$.
3. $\mathcal{A}$ outputs a guess b'.

The adversary's advantage is $\mathrm{Adv}_{\mathcal{A}}^{\text{IND-MOIRÉ}}(\lambda) = |\Pr[b' = 1 \mid b = 1] - \Pr[b' = 1 \mid b = 0]|$.

We say MOIRÉ is (q, ε)-secure if for all adversaries making at most q queries, $\mathrm{Adv}_{\mathcal{A}}^{\text{IND-MOIRÉ}} \leq \varepsilon$.

4.2 Adversary knowledge tiers

Table 2: Security guarantees by adversary knowledge tier.

Tier	Adversary Knowledge	Detection Capability	MOIRÉ Effectiveness
1	No ground truth	Cannot distinguish	Full deception
2	Partial ground truth	Must notice correlated changes	High
3	Real-time verified ground truth	Can solve coord. rotation	Degraded

4.3 Known-plaintext vulnerability

The rigid-body coordinate rotation is vulnerable to known-plaintext analysis. Distance-preserving transformations in two dimensions form a small group (rotation + translation + optional reflection). Two known plaintext-ciphertext coordinate pairs suffice to determine θ and $[\delta_x, \delta_y]$, enabling reversal of the entire coordinate transformation.

A hardened variant applies FF1 to coordinate digits after the geometric rotation, adding an algebraic layer that resists known-plaintext recovery regardless of known pairs. The tradeoff: digit-level FPE breaks distance preservation. Deployments must choose between clustering fidelity (rotation only, vulnerable to KPA) and cryptographic hardening (rotation + FPE, breaks distances).

4.4 The bidirectional oracle attack

The bidirectional mapping requirement (Section 3.2) creates a chosen-plaintext attack surface that, to our knowledge, has not been analyzed in the deception literature.

An adversary who observes a distorted value $v' = \mathsf{Distort}_K(v)$ on screen can submit v' as a search query. The middleware applies $\mathsf{Distort}_K^{-1}(v') = v$, queries the database for v, and returns results that are then distorted. If the adversary knows the structure of the query results, they can infer a plaintext-ciphertext pair from a single adaptive query.

This is strictly stronger than the passive known-plaintext vulnerability: the adversary controls the input and can repeat the process adaptively. We formalize this as follows.

Proposition 4.2 (Oracle query complexity)**.** *An adversary with access to the bidirectional oracle can recover θ and $[\delta_x, \delta_y]$ for the coordinate rotation using at most 2 adaptive queries involving coordinates, or determine a complete substitution table for a categorical field of cardinality n using at most n queries.*

Proof. For coordinates: submit two search queries for distinct distorted coordinates c'_1, c'_2 observed on screen. The middleware maps each c'_i to its preimage $c_i = \mathsf{Distort}_K^{-1}(c'_i)$ and returns results containing records near c_i, which are then distorted. From the two (c_i, c'_i) pairs, the rotation parameters are determined (the system of equations $c' = R_\theta c + \delta$ has unique solution given two non-collinear points). For categorical fields: each query for a distorted category value v' returns results in the real category $\mathsf{Distort}_K^{-1}(v')$, which are distorted back, but the query-result pairing reveals the mapping for that category. □

Countermeasures. Rate-limiting queries (maximum r queries per time window), query-pattern anomaly detection (flagging systematic probing of distorted values), and honeypot queries (injecting fake search results for certain query patterns to detect oracle exploitation).

4.5 Self-verification attack

The most fundamental attack is operational, not cryptographic. An insider submits a record they control—a report, a data entry—and checks whether it renders correctly. If they report a red Toyota at 1st and Main and the system shows a blue Honda at 5th and Oak, the distortion is detected in one step.

The only viable countermeasure is exempting the insider's own submissions from distortion. This creates a secondary vulnerability: the boundary between real (own) and distorted (others') submissions is itself a fingerprinting surface. The self-verification attack is MOIRÉ's most fundamental limitation and motivates the convergence analysis in Section 5.

4.6 Exemption boundary analysis

The self-verification countermeasure (exempting the insider's own submissions) creates a boundary between real records (the insider's) and distorted records (everyone else's). We analyze whether this boundary introduces statistical vulnerability beyond the self-verification attack itself.

Proposition 4.3 (Boundary indistinguishability under bijective distortion)**.** *Under bijective distortion, the boundary between exempted (real) and distorted records is not detectable through statistical analysis of the data itself. Specifically:*

1. *The insider's real records and others' distorted records follow the same marginal distributions ($D_{\mathrm{KL}} = 0$ per Section 9.1).*
2. *Under consistent distortion (same session key for all fields), co-occurrence patterns between the insider's records and others' records are preserved: if record r_a (real) co-occurs with record r_b (distorted) in a query result, the distorted values of r_b's fields are internally consistent.*
3. *The spatial distance between the insider's real coordinates and others' distorted coordinates is preserved by isometric rotation.*

The exemption boundary is therefore detectable only through the self-verification attack (the insider checks whether their own known values render correctly), not through statistical comparison of exempt vs. non-exempt records.

This is a reassuring result: the exemption does not create a *second* vulnerability. The self-verification attack remains the single fundamental limitation, and the convergence analysis (Section 5) and the adaptive deterrence mechanism (Section 7) address it through operational rather than cryptographic means.

4.7 Detection bound

We apply Cachin's information-theoretic steganographic security model [3] to bound the detectability of MOIRÉ distortion.

Let P_{real} denote the distribution of genuine query results over the record space $\mathcal{R}$, and let $P_{\text{MOIRÉ}}$ denote the distribution of distorted results under key K. A passive adversary observing q query results performs a hypothesis test: H_0 (real data) versus H_1 (distorted data).

Theorem 4.4 (Detection Bound)**.** *For any passive adversary $\mathcal{A}$ observing q query results, the advantage in the* IND-*MOIRÉ game satisfies:*

$$\text{Adv}_{\mathcal{A}}^{\text{IND-MOIRÉ}} \leq \sqrt{\frac{q}{2} \cdot D_{\text{KL}}(P_{\mathsf{real}} \| P_{\text{MOIRÉ}})}.$$

Proof. By the data-processing inequality, observing q i.i.d. samples from either P_{real} or $P_{\text{MOIRÉ}}$ yields a total divergence of at most $q \cdot D_{\text{KL}}(P_{\mathsf{real}} \| P_{\text{MOIRÉ}})$. By Pinsker's inequality, the total variation distance satisfies $\delta_{TV}(P_{\mathsf{real}}^{\otimes q}, P_{\text{MOIRÉ}}^{\otimes q}) \leq \sqrt{\frac{1}{2} q \cdot D_{\text{KL}}(P_{\mathsf{real}} \| P_{\text{MOIRÉ}})}$. Since the best distinguishing advantage equals the total variation distance between the two product distributions, the result follows. □

Corollary 4.5 (Distortion capacity)**.** *MOIRÉ achieves (q, ε)-security against passive adversaries when:*

$$D_{\text{KL}}(P_{\mathsf{real}} \| P_{\text{MOIRÉ}}) \leq \frac{2\varepsilon^2}{q}.$$

For $\varepsilon = 0.05$ (5% detection advantage) and $q = 1000$ queries, this requires $D_{\text{KL}} \leq 5 \times 10^{-6}$ nats—the distortion must be extremely close to the real distribution in KL divergence.

Remark 4.6. This bound applies to *passive* adversaries who observe query results without active probing. The bidirectional oracle attack (Proposition 4.2) is an *active* attack that circumvents the detection bound by exploiting the middleware's functional interface rather than statistical properties. The detection bound quantifies how much distortion is safe before passive detection; the oracle attack quantifies how quickly active probing breaks specific data types.

5 Convergence of Full-Environment Deception

The self-verification attack, the oracle attack, and the detection bound all point to the same conclusion: full-environment distortion faces an expanding attack surface as adversary knowledge grows. This section formalizes the convergence.

Definition 5.1 (Adversary side information)**.** Let $\mathcal{D} = \{r_1, \ldots, r_n\}$ be the database records. The adversary's *side information* $I \subseteq \mathcal{D}$ is the set of records whose true values the adversary can independently verify (through external sources, personal knowledge, or the self-verification attack).

Definition 5.2 (Viable distortion space)**.** The *viable distortion space* $V(I) = \mathcal{D} \setminus I$ is the set of records that can be distorted without detection by an adversary with side information I.

Theorem 5.3 (Convergence of Deception)**.** *Let $\mathcal{A}$ be an adversary with side information I that grows over time: $I_0 \subseteq I_1 \subseteq I_2 \subseteq \cdots$. Then:*

1. *$V(I_t)$ is monotonically non-increasing: $V(I_{t+1}) \subseteq V(I_t)$.*
2. *The deception effectiveness $\eta(I) = |V(I)|/|\mathcal{D}|$ satisfies $\eta(\emptyset) = 1$ and $\eta(\mathcal{D}) = 0$.*
3. *For a Tier 2 adversary with side information growing at rate Δ records per unit time via uniform random verification, the expected time to detection of full distortion is $O(|\mathcal{D}|/\Delta)$.*

Proof. (1) and (2) follow from the definition of $V(I) = \mathcal{D} \setminus I$. For (3), at each time step the adversary verifies Δ records drawn uniformly from $\mathcal{D}$. A verified record that is distorted is detected with probability 1. Under full distortion ($V(I_0) = \mathcal{D}$), the probability that at least one of Δ randomly verified records is distorted is 1, so detection is immediate. More generally, if m records are distorted, the expected number of verifications until at least one distorted record is checked follows a negative hypergeometric distribution with expectation $|\mathcal{D}|/m$. Summing over the growth of I_t yields the $O(|\mathcal{D}|/\Delta)$ bound. □

The uniform verification assumption in part (3) is unrealistic—adversaries verify strategically, prioritizing operationally important records. This motivates the following refinement.

Definition 5.4 (Strategic verification)**.** Let each record $r \in \mathcal{D}$ have an operational importance weight $w(r) > 0$ reflecting its exfiltration value. Define the adversary's *verification priority* as $w_v(r) = w(r)$ (verify the most valuable records first) and the defender's *distortion priority* as $w_d(r) = w(r)$ (distort the most valuable records first).

Proposition 5.5 (Tracer placement tension)**.** *Under strategic verification and strategic distortion:*

1. *If the defender distorts the top-m records by importance, the adversary detects distortion in at most m verifications (expected: 1).*
2. *If the defender distorts the bottom-m records by importance, the adversary does not detect distortion until they have verified $|\mathcal{D}| - m$ records—but the distorted records have low exfiltration value and are unlikely to be acted upon.*
3. *The viable tracer space $T \subset \mathcal{D}$ consists of records satisfying $w_{\mathsf{exfil}}(r) > \tau_e$ (interesting enough to steal) and $w_{\mathsf{verify}}(r) < \tau_v$ (unlikely to be verified). If the adversary's exfiltration and verification priorities are positively correlated ($\rho > 0$), then $|T|$ is small and shrinks with ρ.*

This formalizes a fundamental constraint on tracer design: the most effective tracers are records the adversary will exfiltrate but not verify. When the adversary's exfiltration and verification priorities are aligned—they verify what they steal—the viable tracer space collapses. This is the operational content of the convergence theorem: the barium-meal equilibrium is thin not just because of side information growth, but because strategic adversaries verify what matters.

5.1 The barium meal as equilibrium

Theorem 5.3 formalizes why Wright's barium meal [25] is the equilibrium of deception. Against a fully informed adversary ($I = \mathcal{D}$), no record can be distorted without detection ($V(\mathcal{D}) = \emptyset$). Against a partially informed adversary, the viable distortion space is exactly the records the adversary cannot independently verify.

The stable deployment model for sophisticated threat environments is therefore not full-environment distortion but selective distortion with tracers: overwhelmingly real data with a small number of strategically placed marked records in $V(I)$. The marked records must be (i) plausible (not obviously fabricated), (ii) interesting (worth exfiltrating), and (iii) unique (traceable when acted upon). Full Moiré distortion (every record changed) is reserved for Tier 1 adversaries with no ground truth, or as a time-limited intelligence window against partially informed adversaries where the goal is to maximize intelligence extracted before detection.

This convergence is not a limitation to be apologized for. It is the paper's central result: we prove *what optimal deception looks like*, and we provide the full construction that enables any point along the convergence path.

5.2 Practical implications of convergence

The convergence theorem has three practical consequences for deployment.

First, the expected time to detection under full distortion is estimable. A Tier 2 insider who verifies $\Delta = 10$ records per day against a database of $|\mathcal{D}| = 10{,}000$ records has an expected detection time of $O(1000)$ days—roughly 3 years. Against a database of $|\mathcal{D}| = 100$ records with the same verification rate, detection is expected within 10 days. System size directly determines the intelligence window.

Second, the convergence path defines a natural escalation-deescalation cycle. A deployment begins at Tier 1 (tracers only). Intelligence requirements or detected exfiltration trigger escalation to Tier 2 or 3. As the adversary accumulates side information, the viable distortion space shrinks, and the deployment deescalates to Tier 1 or transitions to Tier 4 (containment). The graduated architecture (Section 6) instantiates this cycle.

Third, the convergence theorem informs tracer design. At equilibrium, the only distorted records are those the adversary *cannot* independently verify. These are typically records the adversary has no external reference for: events they did not witness, entities they have not interacted with, data from time periods before their access. Effective tracers are therefore records that (i) fall outside the adversary's verifiable knowledge, (ii) appear operationally interesting (worth exfiltrating), and (iii) carry unique markers that survive the adversary's processing pipeline.

6 Graduated Deception Architecture

The convergence theorem motivates a graduated architecture that adapts distortion depth to threat level and adversary sophistication.

Table 3: Graduated deception depth architecture.

Tier	Trigger	Distortion	Detection Risk
0: Watch	Anomalous behavior	None. Real data. Telemetry only.	None
1: Tracer	Elevated suspicion	5–10 marked records. All else real.	Very low
2: Partial	Confirmed compromise	High-value records distorted.	Low
3: Full MOIRÉ	Active exfiltration	Everything distorted except insider's own.	Medium
4: Contain	Detected or exhausted	Access silently restricted.	N/A

Most deployments should operate at Tier 1. By Theorem 5.3, the viable distortion space for a sophisticated adversary is small; tracer insertion provides attribution capability with minimal detection risk. Full MOIRÉ (Tier 3) is reserved for confirmed active threats where the intelligence value of observing the adversary's behavior outweighs the detection risk. It should be understood as a time-limited intelligence window with an expected duration of $O(|\mathcal{D}|/\Delta)$ before detection.

Tier escalation requires observable signals: anomalous query patterns, failed verification attempts (detectable if monitoring is in place), data appearing in unauthorized contexts. The decision-theoretic question—when to escalate, given imperfect signal quality—is a Bayesian optimal stopping problem that we leave to future work.

7 Adaptive Distortion: A Dynamic Bayesian Deception Game

The detection bound (Theorem 4.4) and the convergence theorem (Theorem 5.3) apply to static distortion. We now introduce a mechanism that responds dynamically to adversary behavior. The model extends the signaling-game-with-evidence framework of Pawlick, Colbert, and Zhu [16] to the specific setting of format-preserving deception middleware.

7.1 Game formulation

Definition 7.1 (Dynamic MOIRÉ Deception Game)**.** A dynamic MOIRÉ deception game $\Gamma = (\Theta, \mathcal{A}_D, \mathcal{A}_A, \beta_0, T)$ is played over T rounds between a defender D (operating the middleware) and an adversary A (the insider), as follows.

Types. Nature draws $\theta \in \Theta = \{0, 1\}$, where $\theta = 0$ (real mode) and $\theta = 1$ (MOIRÉ mode). The adversary's prior belief is $\beta_0 = \Pr[\theta = 1]$.

Per-round protocol. In each round $t = 1, \ldots, T$:

1. The system returns a query result x_t, drawn from P_{real} if $\theta = 0$ or $P^{(d_t)}_{\text{MOIRÉ}}$ if $\theta = 1$, where $d_t \in [0, 1]$ is the current distortion intensity chosen by D.
2. A observes x_t and chooses an action $a_t \in \mathcal{A}_A = \{\mathsf{trust}, \mathsf{probe}, \mathsf{abort}\}$:
   - trust: accept the data and continue. Yields payoff v_e (value of exfiltrated record) if $\theta = 0$, or $-v_p$ (cost of acting on poisoned intelligence) if $\theta = 1$.
   - probe: verify k records against ground truth at cost $c_p > 0$. Each probed record in the viable distortion space $V(I_t)$ is detected as distorted with probability 1. Records in I_t (adversary's side information) are not distorted and pass verification.
   - abort: terminate exfiltration. Yields payoff 0.
3. A updates belief β_{t+1} using Bayes' rule given the observed data and any probing results.
4. D observes A's action and updates the distortion intensity via a *probe-responsive policy*:

$$d_{t+1} = \pi(d_t, a_t) = \begin{cases} \min(d_t + \Delta_+, 1) & \text{if } a_t = \mathsf{probe} \\ \max(d_t - \Delta_-, 0) & \text{if } a_t = \mathsf{trust} \\ d_t & \text{if } a_t = \mathsf{abort} \end{cases}$$

where $\Delta_+ > \Delta_- > 0$ (distortion escalates faster than it relaxes).

7.2 Probe detection

The probe-responsive policy requires a classifier that distinguishes adversarial probing from legitimate use. We define probing by three observable signatures.

Echo queries. A query for a value v' that appeared in a previous query result. Normal users search for values they know from external sources; an adversary exploiting the bidirectional oracle (Section 4.4) searches for values they observed on screen. The echo rate $e_t = |\{q_i : q_i \in \mathsf{results}(q_j), j < i\}|/t$ measures this. Baseline echo rate in normal use is low ($e \approx 0.02$–0.05 in typical database applications); adversarial oracle exploitation produces $e > 0.3$.

Sequential coverage. Verification queries span records systematically rather than clustering around areas of operational interest. The coverage dispersion $\gamma_t = \mathsf{Var}[\mathsf{index}(q_1, \ldots, q_t)]$ distinguishes focused exploration (γ low) from systematic scanning (γ high).

Cross-field verification. An adversary checking for distortion queries the same record across multiple field types (coordinates, then identifiers, then names). Define the cross-field ratio χ_t as the fraction of queries that access a record already accessed via a different field. Normal use shows $\chi < 0.1$; systematic verification shows $\chi > 0.3$.

The probing score is $\sigma(\mathbf{q}) = w_e \cdot e_t + w_\gamma \cdot \hat{\gamma}_t + w_\chi \cdot \chi_t$, with weights $w_e, w_\gamma, w_\chi > 0$ calibrated on a baseline of legitimate user behavior. The probe-responsive policy escalates distortion when σ exceeds a threshold σ^*.

False positives. A legitimate user who searches for a value they saw on screen (to find related records) produces a single echo query, not a systematic pattern. The classifier detects systematic probing, not individual lookups. A single echo query does not trigger escalation; three or more within a session window do. We acknowledge that sophisticated adversaries can slow-roll probing below the detection threshold—but this is the same arms race present in all anomaly detection, and the cost of slow probing (more time exposed, more risk per session) contributes to the deterrence calculus of Theorem 7.3.

7.3 Belief dynamics

The adversary's belief updating depends on the observability of distortion. By the KL analysis of Section 9.1, bijective distortion functions (FPE, permutation substitution, isometric rotation) achieve $D_{\mathrm{KL}}(P_{\mathsf{real}} \| P_{\textsc{Moiré}}) = 0$ against passive observation. This has a critical consequence for the game:

Proposition 7.2 (Passive indistinguishability)**.** *Under bijective distortion, a passive adversary (one who only observes query results without probing) cannot update their belief: $\beta_t = \beta_0$ for all t, regardless of the number of observations.*

This follows directly from $D_{\mathrm{KL}} = 0$: the likelihood ratio $P_{\textsc{Moiré}}(x)/P_{\mathsf{real}}(x) = 1$ for all x, so Bayesian updating produces no belief change. The adversary's only path to belief revision is *active probing*—but active probing triggers the probe-responsive policy.

7.4 The probing-deterrence equilibrium

Theorem 7.3 (Probing-Deterrence Equilibrium)**.** *In the dynamic MOIRÉ deception game Γ with bijective distortion:*

1. ***(Passive regime.)*** *If the adversary never probes, $\beta_t = \beta_0$ for all t and the adversary's expected per-round payoff is:*

$$U_A^{\mathsf{passive}} = (1 - \beta_0) \cdot v_e + \beta_0 \cdot (-v_p) = v_e - \beta_0(v_e + v_p).$$

   The adversary trusts and exfiltrates when $\beta_0 < v_e/(v_e + v_p)$.

2. ***(Probing triggers a detection cascade.)*** *If the adversary probes at round t, they detect distortion with probability $\eta(I_t) = |V(I_t)|/|\mathcal{D}|$ per probed record. Upon detection, β jumps to 1, triggering abort. Upon non-detection, β decreases (making future probing less likely), but d increases (making future probing more dangerous). The expected utility of a single probe is:*

$$U_A^{\mathsf{probe}} = -c_p + k \cdot \eta \cdot 0 + (1 - k \cdot \eta) \cdot [v_e - (\beta_0')(v_e + v_p)]$$

   where β_0' is the posterior belief after a non-detection and $k \cdot \eta$ is the detection probability.

3. ***(Deterrence condition.)*** *The adversary prefers passive acceptance to probing when:*

$$c_p > k \cdot \eta \cdot \big[(1 - \beta_0)(v_e + v_p) - v_p\big] + (1 - k \cdot \eta) \cdot \Delta_\beta \cdot (v_e + v_p) \tag{1}$$

   where Δ_β is the belief improvement from a non-detection result. The right-hand side is the expected value of information from probing. When probing cost c_p exceeds this, the adversary rationally chooses not to verify.

4. ***(Equilibrium characterization.)*** *The game has a threshold equilibrium: there exists $\beta^* \in (0, 1)$ such that the adversary's optimal strategy is:*
   - *If $\beta_t < \beta^*$: play* trust *(exfiltrate without verifying)*
   - *If $\beta_t \geq \beta^*$: play* probe *(verify, then* abort *if detected)*

   Under bijective distortion with $D_{\mathrm{KL}} = 0$, passive observation cannot drive β above β^*, so the adversary remains in the trust regime indefinitely—unless external information or the self-verification attack provides a non-observational belief update.

Proof. (1) Under bijective distortion with $D_{\mathrm{KL}}(P_{\mathsf{real}} \| P_{\textsc{Moiré}}) = 0$ (Proposition 7.2), passive observation produces likelihood ratio $L_t = P_{\textsc{Moiré}}(x_t)/P_{\mathsf{real}}(x_t) = 1$ for all t. By Bayes' rule, $\beta_t = \beta_0 \cdot \prod_{s=1}^{t} L_s / [\beta_0 \cdot \prod L_s + (1 - \beta_0)] = \beta_0$ for all t. The adversary's expected payoff from trusting follows by linearity of expectation over θ.

(2) The exact probability that at least one of k probed records is distorted (given distortion density $\eta = |V(I)|/|\mathcal{D}|$) is $p_d = 1 - (1 - \eta)^k$. Upon detection: the adversary updates $\beta \to 1$ (certainty of distortion) and plays abort, yielding payoff $-c_p$. Upon non-detection: β updates by Bayes' rule to

$$\beta' = \frac{\beta_0(1-\eta)^k}{\beta_0(1-\eta)^k + (1-\beta_0)} < \beta_0$$

since $(1-\eta)^k < 1$. Simultaneously, the probe-responsive policy increases $d_{t+1} = d_t + \Delta_+$, which increases η in subsequent rounds. Define the coupled recursion (β_t, η_t): probing decreases β (reassurance) but increases η (more distortion). The net effect on the adversary's continuation value is negative because the increased detection risk from higher η outweighs the reassurance from non-detection: the probability of detection in the *next* probe is $1-(1-\eta_{t+1})^k > 1-(1-\eta_t)^k$. Formally, the adversary's continuation value $V_t(\beta_t, \eta_t) = \max\{U_t^{\mathsf{trust}}, U_t^{\mathsf{probe}}\}$ satisfies $\partial V_t/\partial \eta_t < 0$ (higher η is worse for the adversary) and $|\partial V_t/\partial \eta_t| > |\partial V_t/\partial \beta_t|$ when $\Delta_+ > (1-\eta)^k \cdot \eta/[1 - (1-\eta)^k]$.

(3) The adversary prefers trusting when $U^{\mathsf{passive}} > U^{\mathsf{probe}}$. Expanding:

$$v_e - \beta_0(v_e + v_p) > -c_p + (1 - p_d)[v_e - \beta'(v_e + v_p)].$$

Rearranging yields the deterrence condition (1). For small η (Tier 1), $p_d \approx k\eta \ll 1$ and $\beta' \approx \beta_0(1 - k\eta)$, giving the simplified condition (2).

(4) Define $F(\beta) = U^{\mathsf{trust}}(\beta) - U^{\mathsf{probe}}(\beta)$. We have $F(0) = v_e - (-c_p + v_e) = c_p > 0$ (at zero belief, probing is wasteful) and $F(1) = -v_p - (-c_p - v_p) = c_p > 0$. Since $F(0) > 0$ and the adversary should probe at intermediate beliefs where information gain is maximized, F must have an interior minimum. If $F(\beta) > 0$ for all $\beta \in [0, 1]$—which occurs when c_p is sufficiently large relative to the information gain—the adversary never probes (full deterrence). Otherwise, the threshold β^* satisfies $F(\beta^*) = 0$, and its uniqueness follows from the strict concavity of U^{probe} in β (the probe's value peaks at intermediate beliefs and drops at extremes). □

Remark 7.4. The deterrence result depends on the adversary's inability to update beliefs through passive observation ($D_{\mathrm{KL}} = 0$). If the distortion is not bijective—if it introduces statistical artifacts detectable through passive observation—then β_t drifts upward with observation count and eventually crosses β^*, triggering probing and potential detection. This makes the KL analysis of Section 9.1 load-bearing for the game-theoretic result: bijective distortion is not merely a design preference but a game-theoretic requirement for sustained deterrence.

7.5 Parameter sensitivity analysis

The deterrence condition (1) depends on the adversary's private parameters. We assess whether it holds across plausible ranges by normalizing to $v_e = 1$ (one unit of exfiltration value) and expressing other parameters as ratios.

Let $\rho = v_p/v_e$ (poison cost ratio: how much worse is acting on distorted data than the data is worth), $\gamma = c_p/v_e$ (probing cost ratio), η (distortion density: fraction of records distorted), and β_0 (prior belief that MOIRÉ is active). The simplified deterrence condition (first-order in η, $k = 1$) is:

$$\gamma > \eta \cdot (1 - \beta_0) \cdot (1 + \rho). \tag{2}$$

Table 4: Deterrence condition across parameter regimes. The adversary is deterred from probing when γ (probing cost ratio) exceeds the right-hand side. Bold entries: deterrence holds for $\gamma \geq 0.10$ (10% of record value).

Regime	η	β_0	ρ	RHS	Deterred at $\gamma = 0.10$?
Tier 1 (tracers, low prior)	0.01	0.10	5	0.054	**Yes**
Tier 1 (tracers, mod. prior)	0.05	0.30	5	0.210	No
Tier 2 (partial, low prior)	0.20	0.10	3	0.720	No
Tier 3 (full, mod. prior)	0.80	0.30	2	1.680	No
Tier 3 (full, low prior)	0.80	0.05	2	2.280	No

The sensitivity analysis confirms the convergence result from a different angle: the deterrence condition holds at Tier 1 (sparse tracers, low distortion density) and fails at Tier 3 (full distortion). At Tier 3, the adversary's expected gain from probing exceeds the cost for any plausible γ—the rational adversary will probe, detect the distortion, and abort. At Tier 1, the expected gain from probing is small (most records pass verification), and even modest probing costs deter investigation.

This is the game-theoretic mirror of Theorem 5.3: the optimal deception regime is thin distortion with tracers, because that is the only regime where the probing-deterrence equilibrium is sustainable.

8 The Turned-Asset Doctrine

When a legitimate insider is confirmed compromised—coerced, bribed, or turned—the standard incident response is to revoke access. This is correct for containment but wrong for intelligence: revocation alerts the adversary that the compromise has been detected.

The alternative: silently route the compromised insider's session through the distortion layer. From their perspective, the system functions normally. Every query reveals their handler's priorities. Every exfiltrated record becomes a uniquely keyed tracer. Every behavioral pattern builds an evidence trail.

MITRE Engage acknowledges this gap: "if you are attempting to identify an insider threat, an isolated environment may not be useful" [14]. MOIRÉ's distortion layer makes it useful because the environment is not isolated—it is the same environment, just wrong.

8.1 Ethical constraints

The turned-asset doctrine requires explicit ethical guardrails. We distinguish two engagement modes:

Observation-only engagement: the adversary's queries reveal priorities and interests, but no distorted data enters real-world decision-making. This mode is ethically defensible and operationally low-risk.

Action-enabling engagement: distorted data may influence the adversary's real-world actions (deploying to wrong coordinates, acting on wrong identifiers). This mode carries risk of harm to third parties and requires additional safeguards.

For action-enabling engagement: external attackers using stolen credentials may be engaged without time limit. Coerced insiders require welfare intervention, not just intelligence extraction. Voluntary insiders should trigger a bounded engagement window (we suggest 72 hours as a default) with mandatory administrative review. All ghost-mode sessions must have maximum durations and be architecturally prevented from generating exportable case packages or legal evidence.

9 Discussion and Limitations

9.1 Quantitative distinguishability analysis

We now compute the KL divergence $D_{\mathrm{KL}}(P_{\mathsf{real}} \| P_{\text{MOIRÉ}})$ for each data type in Table 1. The central result is that bijective distortion functions achieve $D_{\mathrm{KL}} = 0$ against passive adversaries without domain priors, concentrating all vulnerability in active attacks and external anchors.

Identifiers (FPE). FF1 implements a pseudorandom permutation σ_K on the identifier domain $\mathcal{D}$. Since σ_K is a bijection, any value v appearing k_v times in the real data yields a distorted value $\sigma_K(v)$ appearing exactly k_v times. The frequency profile is preserved. For a passive adversary observing the marginal distribution over identifier values:

$$P_{\text{MOIRÉ}}(v') = P_{\mathsf{real}}(\sigma_K^{-1}(v')) \quad \text{for all } v' \in \mathcal{D}.$$

Since σ_K is a permutation, $\{P_{\text{MOIRÉ}}(v')\}_{v'}$ is a rearrangement of $\{P_{\mathsf{real}}(v)\}_v$ and the multiset of probabilities is identical: $D_{\mathrm{KL}} = 0$.

Coordinates (rigid-body rotation). The rotation R_θ is an isometry on $\mathbb{R}^2$. All spatial statistics (inter-point distances, clustering, density, corridor structure) are preserved. For the marginal distribution of coordinate values:

- Without geographic prior: $D_{\mathrm{KL}} = 0$. No spatial pattern distinguishes the rotated from the original data.
- With geographic prior P_{geo} (e.g., "data concentrates near city X"): if the rotation maps city X to a region with low prior probability, $D_{\mathrm{KL}} > 0$. For a Gaussian prior with spatial standard deviation σ_{geo} and a rotation that displaces the mean by $\|\boldsymbol{\delta}\|$: $D_{\mathrm{KL}} \approx \|\boldsymbol{\delta}\|^2/(2\sigma_{\mathsf{geo}}^2)$. A 50km displacement against a $\sigma = 10$km city prior yields $D_{\mathrm{KL}} \approx 12.5$ nats—immediately detectable. A 5km displacement yields $D_{\mathrm{KL}} \approx 0.12$ nats—requiring $\sim$80 queries at $\varepsilon = 0.10$ for detection.

Mitigation: choose $\boldsymbol{\delta}$ to map the data to a geographically plausible region (another city of similar structure), reducing D_{KL} under the adversary's prior.

Names (dictionary substitution). If the substitution is a permutation within the same name set (Smith $\to$ Johnson, Johnson $\to$ Williams, etc.), the frequency profile is preserved and $D_{\mathrm{KL}} = 0$. If the substitution maps to a different name set with mismatched Zipfian frequencies, $D_{\mathrm{KL}} > 0$. Numerical computation for the top $N = 1000$ names with Zipf parameter $s = 1$: a random cross-set permutation yields $D_{\mathrm{KL}} \approx 2.88$ nats (immediately detectable); a worst-case reversal yields $D_{\mathrm{KL}} \approx 3.51$ nats. Recommendation: always use permutation-based substitution within the same name set.

Timestamps (uniform offset). A uniform temporal offset preserves the shape of the temporal distribution (intervals, periodicity, trends). Against a passive adversary with no temporal anchors, $D_{\mathrm{KL}} = 0$. Against an adversary who knows an external anchor event at time t_0 (a news event, a system outage, a known transaction), the offset is detectable from a single anchor: the expected spike at t_0 appears at $t_0 + \Delta_t$, and Δ_t is immediately recoverable.

Categorical fields (keyed permutation). Same analysis as identifiers: bijective substitution within the category set preserves frequencies. $D_{\mathrm{KL}} = 0$.

Aggregates (pass-through). No distortion applied. $D_{\mathrm{KL}} = 0$ by definition.

Table 5: Quantitative KL divergence by data type and adversary knowledge.

Data Type	D_{KL} (no prior)	D_{KL} (with prior)	Primary Vulnerability
Identifiers (FPE)	0	0	Known-plaintext only
Names (same-set perm.)	0	0	Known-plaintext only
Names (cross-set)	2.88 nats	2.88 nats	Frequency mismatch
Coordinates (rotation)	0	$\|\boldsymbol{\delta}\|^2/2\sigma^2$	Geographic prior
Timestamps (offset)	0	∞ (one anchor)	Temporal anchor
Categorical (perm.)	0	0	Known-plaintext only
Aggregates	0	0	N/A (pass-through)

Key finding. Bijective distortion functions (FPE, permutation-based substitution, isometric rotation) achieve $D_{\mathrm{KL}} = 0$ against passive adversaries without domain priors. This means MOIRÉ's vulnerability is *operational* (active probing, known plaintexts, external anchors), not *statistical*. A passive adversary cannot detect the distortion regardless of observation count—the detection bound of Theorem 4.4 yields $(q, 0)$-security for all q.

This result is load-bearing for the game-theoretic analysis of Section 7: the adversary's inability to passively update beliefs about θ is what makes the probing-deterrence equilibrium stable.

9.2 Scope: structured-data systems

MOIRÉ handles structured data (text, numbers, coordinates). Images and documents resist format-preserving transformation—a photo of a red Honda does not become a photo of a blue Toyota through FPE. The construction is explicitly scoped to structured-data systems. Media-rich applications require supplementary techniques (reference image substitution, media suppression) that weaken the deception and are not addressed here.

9.3 Legal considerations

Silently routing a session through modified data may implicate wiretap statutes. The construction is architecturally identical to A/B testing and feature flags—the system serves different content based on session state—but legal intent differs. Counsel should evaluate jurisdiction-specific implications before deployment.

9.4 Deterrence without deployment

The mere credible existence of MOIRÉ changes adversary behavior even without deployment. An adversary who knows data-poisoning deception capabilities exist cannot fully trust any data they access, even genuine data. This is deterrence-by-denial in the sense of Schelling [20]: the capability degrades the value of the adversary's action without requiring revelation of whether or when it is used. The publication of this paper is itself a deterrence event.

9.5 Connection to differential privacy

Across sessions (different keys), FPE distortion resembles the local model of differential privacy: each query result is "perturbed" before leaving the data holder. However, the goals differ fundamentally. Differential privacy seeks ε-indistinguishability to protect individual records. MOIRÉ seeks to induce confident wrong belief about specific values. As Kim et al. [13] note: "differential privacy seeks to induce uncertainty in an observer, whereas [deception] aims to make the adversary confidently draw the incorrect conclusion." Whether FPE satisfies formal DP guarantees for quantifiable ε is an open question.

9.6 Defensive data poisoning against AI model theft

A timely application: if an adversary scrapes data to train AI models, MOIRÉ distortion poisons the training data in a controlled, traceable way. The FPE ensures the poisoned data looks structurally correct (evading scraper filters), but every specific value is wrong. Models trained on distorted data produce systematically wrong outputs on real-world inputs. The $D_{\mathrm{KL}} = 0$ property of bijective distortion means the scraper cannot detect the poisoning through statistical analysis of the scraped data.

10 Conclusion

MOIRÉ is not a permanent deception system. Against sophisticated adversaries with independent ground truth, full-environment distortion will eventually be detected—and we prove why (Theorem 5.3). The construction's value lies in three places: as a formal specification for insider deception with explicit security bounds; as the enabling mechanism for a graduated architecture that adapts to threat sophistication; and as the basis for a probing-deterrence equilibrium (Theorem 7.3) that transforms verification from a vulnerability into a detection signal.

The primitives are not new. FPE, deception technology, game-theoretic adversary modeling, and intelligence tradecraft all predate this work. The contributions are the composition of these primitives into a coherent architecture with formal detection bounds (Theorem 4.4), a convergence proof that formalizes the barium-meal equilibrium, and an adaptive mechanism with a characterized deterrence regime.

References

[1] Acalvio Technologies. ShadowPlex. https://acalvio.com.

[2] Bellare, M., Rogaway, P., and Spies, T. (2010). The FFX mode of operation for format-preserving encryption. NIST submission.

[3] Cachin, C. (1998). An information-theoretic model for steganography. In *Information Hiding* (IH 1998), LNCS 1525, pp. 306–318. Springer.

[4] Cohen, F. (1997). The Deception Toolkit. http://all.net/dtk.

[5] CounterCraft. Deception-powered threat intelligence. https://countercraftsec.com.

[6] Durak, F.B. and Vaudenay, S. (2017). Breaking the FF3 format-preserving encryption standard over small domains. In *CRYPTO 2017*, LNCS 10402, pp. 679–707. Springer.

[7] Heckman, K., Stech, F., Thomas, R., Schmoker, B., and Tsow, A. (2015). *Cyber Denial, Deception and Counter Deception*. Springer.

[8] Horák, K., Zhu, Q., and Bošanský, B. (2017). Manipulating adversary's belief: A dynamic game approach to deception by design for proactive network security. In *GameSec 2017*, LNCS 10575, pp. 273–294. Springer.

[9] Huang, L. and Zhu, Q. (2021). Duplicity games for deception design with an application to insider threat mitigation. *IEEE Trans. Inf. Forensics Security*, 16:569–584.

[10] Juels, A. and Ristenpart, T. (2014). Honey encryption: Security beyond the brute-force bound. In *EUROCRYPT 2014*, LNCS 8441, pp. 293–310. Springer.

[11] Kahlhofer, M. and Rass, S. (2024). Application layer cyber deception without developer interaction. *arXiv preprint* arXiv:2405.12852.

[12] Kahlhofer, M., Golinelli, E., and Rass, S. (2025). Koney: A cyber deception orchestration framework for Kubernetes. *arXiv preprint* arXiv:2504.02431.

[13] Kim, Y., Zhou, H., Benvenuti, A., Hu, R., and Hale, M. (2026). Deception in linear-quadratic control. *arXiv preprint* arXiv:2604.00227.

[14] MITRE Engage Framework (2022). https://engage.mitre.org.

[15] NIST SP 800-38G (2016). Recommendation for block cipher modes of operation: Methods for format-preserving encryption.

[16] Pawlick, J., Colbert, E., and Zhu, Q. (2019). Modeling and analysis of leaky deception using signaling games with evidence. *IEEE Trans. Inf. Forensics Security*, 14(7):1871–1886.

[17] Proofpoint / Illusive Networks. Shadow. https://proofpoint.com.

[18] Rapid7. US Patent 11,303,675: Containing compromised credentials using deception systems (2022).

[19] Roy Chowdhury, A., Ding, B., Jha, S., Liu, W., and Zhou, J. (2022). Strengthening order preserving encryption with differential privacy. In *ACM CCS 2022*. arXiv:2009.05679.

[20] Schelling, T.C. (1960). *The Strategy of Conflict*. Harvard University Press.

[21] SentinelOne / Attivo Networks. US Patent 11,038,658: Deceiving attackers in endpoint systems (2021).

[22] Spitzner, L. (2003). *Honeypots: Tracking Hackers*. Addison-Wesley.

[23] Stoll, C. (1989). *The Cuckoo's Egg*. Doubleday.

[24] Thinkst Applied Research. CanaryTokens. https://canarytokens.org.

[25] Wright, P. (1987). *Spycatcher*. Viking.

[26] Ye, D., Zhu, T., Shen, S., and Zhou, W. (2021). A differentially private game theoretic approach for deceiving cyber adversaries. *IEEE Trans. Inf. Forensics Security*, 16:569–584.

A Adversarial Detection Techniques

Table 6 catalogs seven detection techniques available to a MOIRÉ-aware adversary—one who knows deception systems exist and actively probes for them. This analysis extends the standard threat model of Sections 4–4.4 to the meta-game: what does the adversary do when they suspect deception?

Table 6: Detection techniques available to MOIRÉ-aware adversaries.

Technique	Method	Difficulty	Countermeasure
Self-verification	Submit controlled record, check rendering	Low	Exempt insider's own submissions
External cross-reference	Verify data against external source	Low–Med	Accept as inherent limitation
Social consistency	Ask other users to verify specific records	Medium	Place suspected users in same session key
Statistical fingerprinting	Test distributions against known baselines	High	Calibrate substitution to match real distributions
Temporal probing	Submit record, measure time to visibility	Medium	Match distortion-mode latency to production
Behavioral probe	Trigger actions that should cause downstream effects	Med–High	Simulate downstream effects within distortion session
Canary detection	Examine credentials for signs of planting	High	Ensure canary credentials are indistinguishable

The self-verification attack (row 1) is the most dangerous because it is the simplest and requires no technical sophistication. It is MOIRÉ's most fundamental limitation at full distortion and is the primary motivation for the convergence analysis (Section 5) and the adaptive distortion mechanism (Section 7).

B The Steganography Duality

MOIRÉ is steganography in reverse. Steganography hides a *message* inside normal-looking data; MOIRÉ hides *distortion* inside normal-looking data. The detection problem is dual: steganalysis techniques (detecting hidden messages) map directly to MOIRÉ detection techniques (detecting hidden distortion).

The duality is formally grounded in Cachin's model [3]. In steganography, the cover distribution P_C is the distribution of innocent messages, and the stego distribution P_S is the distribution after message embedding. The stegosystem is ε-secure when $D_{\mathrm{KL}}(P_C \| P_S) \leq \varepsilon$.

In MOIRÉ, the "cover" is the real data distribution P_{real}, and the "stego" is the distorted distribution $P_{\text{MOIRÉ}}$. The deception is ε-secure when $D_{\mathrm{KL}}(P_{\mathsf{real}} \| P_{\text{MOIRÉ}}) \leq \varepsilon$. Theorem 4.4 is a direct application of this correspondence.

The duality suggests that mature steganalysis techniques—Fridrich's rich models, deep learning steganalyzers, statistical hypothesis tests on feature distributions—have direct analogs as MOIRÉ detectors. Any advance in steganalysis translates to a potential attack on MOIRÉ, and any defense against steganalysis translates to a design principle for distortion functions. This bidirectional relationship connects MOIRÉ to a mature literature with decades of adversarial co-evolution.

Practically, the duality imposes a *distortion capacity*: the maximum amount of distortion achievable before the KL divergence exceeds the detection threshold. This capacity depends on the data type, the distortion function, and the adversary's statistical power. For structured fields (identifiers, timestamps), the capacity is high because FPE preserves format and bijection. For

fields with distinctive frequency distributions (names, categorical values), the capacity is lower because even bijective substitution can shift the frequency profile.

C Attack Surfaces Created by MOIRÉ

MOIRÉ introduces vulnerabilities that do not exist without deception.

The middleware itself becomes the highest-value target. Compromise the key schedule and the entire deception becomes transparent. Key management creates a new secret category to protect, with all the operational burden that entails.

The boundary between real and distorted data is a fingerprinting surface. Under the self-verification exemption (where the insider's own submissions render correctly), the boundary between real and distorted records is observable. The insider can map this boundary by submitting records at known locations and checking whether neighboring records are consistent.

False positive deployment suppresses real intelligence. While a ghost-mode session is active, the operator is working with phantom data. Real intelligence arriving during the ghost-mode window may be discarded as distorted, or worse, acted upon as if real when it is the distorted version that reaches the operator's secondary systems.

Ghost-mode evidence is legally toxic. Fabricated evidence entering a legal proceeding could discredit all evidence from the system. Ghost-mode sessions must be architecturally prevented from generating exportable case packages, reports, or evidence artifacts.