MOIRÉ: Format-Preserving Encryption as Deception Middleware

Construction, adversarial analysis, and graduated deception architecture

David Kirsch - May 2026 - davidkirsch.me/research

This paper presents MOIRÉ, a construction that applies format-preserving encryption (NIST SP 800-38G) as a real-time deception layer between a production database and its application. The middleware produces a structurally faithful but specifically wrong version of reality: patterns, distances, and aggregate statistics survive the transformation while every individual value changes. We describe the core construction, analyze its security properties against three tiers of adversary knowledge, and identify fundamental vulnerabilities including a self-verification attack and known-plaintext weakness in the coordinate transformation. We then extend the analysis to MOIRÉ-aware adversaries - those who know deception systems exist and actively probe for them - and propose a graduated deception architecture that adapts distortion depth to threat level. We show that against sufficiently sophisticated adversaries, all full-environment deception systems converge toward selective distortion with tracers. Finally, we formalize the turned-asset doctrine: using the distortion layer to run compromised insiders as unwitting intelligence sources rather than revoking their access. The contribution is not any single primitive (all exist independently) but their composition into a coherent operational doctrine with explicit security bounds and failure modes.

1. Introduction

Deception technology has been part of the security toolkit for decades. Honeypots detect intrusions [3][4]. Honeytokens trip alarms when touched [14]. MITRE's Engage framework formalizes adversary engagement operations [5]. Commercial platforms from SentinelOne, CounterCraft, Acalvio, and Proofpoint deploy decoy servers, fake credentials, and synthetic data environments across enterprise networks [15-18].

All of them share a limitation: they're built to fool outsiders. An attacker who has never seen the real system can't tell the decoy from production. But an insider - someone who has legitimate access and knows what the real data looks like - will recognize a static decoy immediately. The record count is wrong. The activity patterns don't 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 as a real-time middleware layer. Instead of building 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 volumes, real statistical distributions. Every specific value is wrong. The construction is named after the moiré pattern - the interference effect that emerges when two regular patterns overlap at slightly different angles, producing a third pattern that looks structured but doesn't exist in either original.

We make three contributions. First, the MOIRÉ construction itself: a specification for format-preserving distortion middleware with data-type-specific transformations and session-keyed attribution. Second, an

adversarial analysis that goes beyond the standard threat model to consider MOIRÉ-aware adversaries who actively probe for deception, identifying seven detection techniques and corresponding countermeasures. Third, a graduated deception architecture that adapts distortion depth to threat level, acknowledging that full-environment distortion is the wrong response against sophisticated adversaries and converging on selective distortion with tracers as the stable equilibrium.

2. Prior Art

Format-preserving encryption.

Bellare, Rogaway, and Spies formalized FPE in 2009 [1]. NIST standardized two modes - FF1 and FF3-1 - in SP 800-38G [2]. 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. No prior work applies FPE as a deception mechanism.

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 [3][4]. MITRE's Engage framework (2022) structures adversary engagement into Expose, Affect, and Elicit operations [5]. Heckman, Stech et al. adapted Whaley's classical denial-and-deception model to cybersecurity [6]. NIST SP 800-53 Rev. 5 codifies deception in controls SC-26 and SC-30 [7]. Kahlhofer and Rass (2024) surveyed application-layer deception specifically and identified the subfield as underdeveloped [9]. Anagnostakis et al. introduced shadow honeypots (2005) [10]. Araujo and Hamlen developed honey-patches (2014) [11].

Cross-domain tracer data.

Peter Wright documented the "barium meal" technique in Spycatcher (1987) - distributing uniquely varied intelligence to suspected moles [12]. Tom Clancy coined "canary trap" for the same concept [13]. Thinkst operationalized a digital version with CanaryTokens [14]. The gap: no existing system applies format-preserving encrypted data as tracers designed to be attributable when the adversary acts on them in the physical world.

Distinguishing MOIRÉ from existing approaches.

A fair objection is that MOIRÉ is "just encryption at the database layer" - a known technique rebranded. The distinction is purpose and architecture. Database encryption protects confidentiality from unauthorized access. MOIRÉ provides *authorized access to transformed data* - the user believes they have real access and acts accordingly. Honeypots are isolated decoy systems; MOIRÉ operates on the production system with real data. The structural fidelity requirement (preserving patterns, distances, and aggregates while changing specifics) creates engineering constraints that don't exist in either database encryption or honeypot construction. The contribution is the composition, not any individual primitive.

3. The MOIRÉ Construction

A middleware layer sits between the production database and the application's rendering layer. The application code is unmodified - same queries, same components, same business logic. Query results pass

through a distortion function parameterized by a session-specific cryptographic key.

The distortion function has five properties. It is **deterministic**: same input and same key always produce the same output. 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 street. It is **session-keyed**: different sessions produce different distortions. And it is **irreversible without the key**.

Critical requirement: bidirectional distortion. The middleware must intercept both query results (real → distorted) AND query inputs (distorted → real). Without bidirectional distortion, a user who searches for a distorted value they saw on screen gets empty results from the real database - an immediate tell. The middleware reverse-maps search terms through the distortion function before querying, then forward-maps results before rendering.

Data-type-specific 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 from key	Sequence, gaps, temporal clustering	Absolute dates
Categorical fields	Keyed substitution within equivalence class	Category structure	Specific values
Names, aliases	Keyed dictionary substitution	Format, counts	Every name
Aggregates, counts	Pass-through	Everything	Nothing

Table 1. Data-type-specific distortion transformations.

The coordinate rotation.

Geographic data requires special treatment. Random noise on individual coordinates destroys clustering and is statistically detectable. Instead, a single rigid-body transformation (rotation by angle θ plus translation [dx, dy]) is applied to all coordinates simultaneously. Two points 200 meters apart remain 200 meters apart. A corridor pattern stays a corridor - in a different part of the city. Parameters are derived from the session key via HKDF-SHA256.

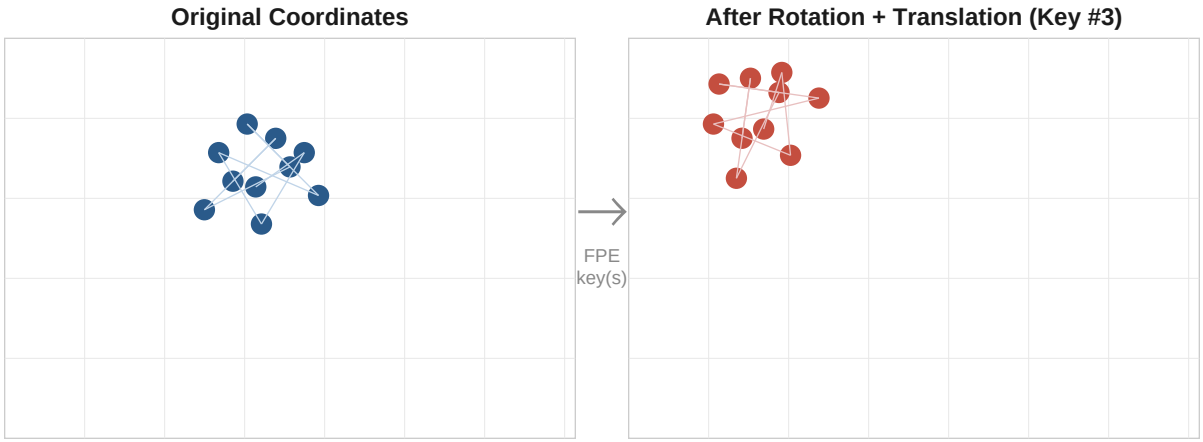

Figure 1. Rigid-body rotation preserves inter-point distances and clustering. Left: original coordinates. Right: same points after keyed rotation and translation.

The attribution chain.

Each session uses a unique key. When distorted data surfaces in the physical world, 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 [12] formalized with cryptographic rigor - a provable attribution chain from physical-world adversary behavior through digital deception to the point of compromise.

4. Security Analysis

Adversary knowledge tiers.

Tier	Adversary Knowledge	Detection Capability	MOIRÉ Effectiveness
1	No ground truth. Never seen real system.	Cannot distinguish from production.	Full deception. Indistinguishable.
2	Partial ground truth. Knows some real values.	Difficult. Must notice multiple correlated changes.	High. Consistent distortion resists casual verification.
3	Verified real-time ground truth on specific data points.	Can solve coordinate rotation with 2 known point correspondences.	Degraded. Coordinate transformation reversible.

Table 2. Security guarantees by adversary knowledge tier.

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 are sufficient to determine the rotation angle and translation vector, enabling reversal of the entire coordinate transformation.

A hardened variant applies FPE to coordinate digits after the geometric rotation. This adds an algebraic layer that can't be solved without the key, regardless of known pairs. The tradeoff: digit-level FPE breaks distance preservation. A hybrid approach uses rotation-only for visual map rendering (preserving clustering) and rotation-plus-digit-FPE for exported numeric data (hardened against cryptanalysis).

Self-verification attack.

The most dangerous and simplest attack against MOIRÉ is not cryptographic but operational. An insider submits a record they control - a sighting, a report - 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, they've detected the distortion in one step. This requires no technical sophistication.

The only viable countermeasure is exempting the insider's own submissions from distortion. Their records render correctly; all other records are distorted. This creates a secondary vulnerability: the insider can map the boundary between their real submissions and the distorted background. This boundary is itself a detection signal. The self-verification attack is MOIRÉ's most fundamental limitation and cannot be fully mitigated within the full-distortion model.

5. The Adversarial Meta-Game

The preceding analysis assumes an adversary who doesn't know MOIRÉ exists. A qualitatively different threat emerges when the adversary expects deception systems and has developed detection techniques. This section maps that landscape.

Detection techniques.

Test	Method	Sophistication	Countermeasure
Self-verification	Submit controlled record, check rendering	Low	Exempt insider's own submissions from distortion
External cross-reference	Verify data points against external source	Low-Medium	No full defense. Accept as inherent limitation.
Social consistency	Ask other users to verify specific records	Medium	Place all suspected compromised users in same session key
Statistical fingerprinting	Test coordinate clustering vs road grid alignment	High	Road-grid-aware coordinate transformation
Temporal probing	Submit record, time how long until visible	Medium	Match ghost-mode update frequency to production
Behavioral probe	Trigger actions that should cause downstream effects	Medium-High	Ghost-mode actions trigger real-looking downstream effects within ghost session
Canary detection	Examine credential for signs of planting	High	Canary credentials indistinguishable from real

Table 3. MOIRÉ detection techniques and countermeasures.

Attack surfaces created by deception.

MOIRÉ introduces vulnerabilities that don't exist without deception. The distortion middleware itself becomes the highest-value target in the architecture - compromise the key schedule and the entire deception becomes transparent. Key management creates a new secret category to protect. The boundary between real and distorted data is a fingerprinting surface. False positive deployment suppresses real intelligence while the operator works with phantom data. Most critically, 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 or reports.

6. Graduated Deception Architecture

The self-verification attack and the adversarial meta-game demonstrate that full-environment distortion is the wrong response against sophisticated adversaries. The detection surface is too large and the countermeasures introduce their own vulnerabilities. A more resilient architecture adapts distortion depth to threat level.

Tier	Trigger	Distortion	Detection Risk
0: Watch	Anomalous behavior detected	None. Real data. Behavioral telemetry only.	None
1: Tracer insertion	Elevated suspicion. Repeated anomalies.	5-10 marked records planted. All other data real.	Very low. 90-95% of data passes verification.
2: Partial distortion	Confirmed compromise of specific data categories.	High-value records distorted. Aggregates and most data real.	Low. Casual verification passes.
3: Full MOIRÉ	Confirmed active threat. Active exfiltration detected.	Everything distorted except insider's own submissions.	Medium. Systematic probing will detect.
4: Contain	Deception detected or intelligence window exhausted.	Access silently restricted. No deception.	N/A. Transition to containment.

Table 4. Graduated deception depth architecture.

Most deployments should operate at Tier 1. Tracer insertion provides attribution capability with minimal detection risk. The marked records must be plausible (not obviously fabricated), interesting (worth exfiltrating), and unique (traceable when acted upon). The insider verifies 20 records against external ground truth and finds them all correct. The 3 marked records they don't verify are the ones that trace back to them.

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, not a permanent deception. The design goal is to maximize intelligence extracted before detection, not to prevent detection indefinitely.

7. Convergence Toward Selective Distortion

Against sufficiently sophisticated adversaries, all full-environment deception systems face the same fundamental constraint: the more the adversary knows independently of the system, the narrower the viable distortion space becomes. A brand-new insider can be fully deceived. A veteran with two years of system experience and extensive external knowledge is nearly impossible to fully deceive. The deception space shrinks monotonically with adversary knowledge.

This convergence is not a novel insight - it's the foundational assumption of adversarial security. What it implies for MOIRÉ is specific: the stable deployment model for sophisticated threat environments is not full-environment distortion but selective distortion with tracers. This is where intelligence tradecraft has always settled. Wright's barium meal worked because most of the intelligence was real [12]. The canary trap works because most of the document is genuine. The deception is thin but invisible.

The MOIRÉ construction remains valuable at Tier 3 (full distortion) for unsophisticated adversaries and for the initial phase of an engagement before the adversary begins systematic probing. The graduated architecture acknowledges this by treating full MOIRÉ as one tier in a spectrum, not the default operating mode.

8. The Turned-Asset Doctrine

The distortion middleware enables a broader operational 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. It is 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" [5]. The distortion layer makes it useful because the environment isn't isolated - it's the same environment, just wrong. The primitives required for this construction exist in published work and in some cases are patented (Rapid7 US 11,303,675 [19]; SentinelOne US 11,038,658 [20]). What doesn't exist is the strategic doctrine that composes them with format-preserving distortion into a sustained insider engagement protocol.

Ethical constraints apply. External attackers using stolen credentials can be engaged indefinitely. Coerced insiders require welfare intervention, not just intelligence extraction. Voluntary insiders should trigger a 72-hour maximum engagement window before mandatory review. Ghost-mode sessions must have maximum durations and administrative oversight.

9. Limitations and Open Problems

Media-rich data. The construction handles structured data (text, numbers, coordinates). Images and documents resist format-preserving transformation. A photo of a red Honda doesn't become a photo of a blue Toyota through FPE. Media-rich applications require supplementary handling - reference image substitution or media suppression - that weakens the deception.

Open source exposure. If the system's source code is public, the distortion middleware's existence is known. Per Kerckhoffs's principle, security should depend on the key, not the algorithm's secrecy. But knowledge of the middleware's existence changes the adversary's approach from naive to MOIRÉ-aware, collapsing the threat model to Section 5.

Legal uncertainty. Silently routing a session through modified data may implicate wiretap statutes. The construction is architecturally identical to A/B testing or feature flags (the system serves different content based on session state), but legal counsel should evaluate jurisdiction-specific implications before deployment.

Formal security proof. The composed construction (FPE + rigid-body transformation + keyed substitution + bidirectional query mapping) has not been formally proven secure under stated assumptions. The security analysis in Section 4 is descriptive, not provable.

Adaptive deception. A future direction: compute distortion lazily at query time, adapting based on detected probing behavior. Records being systematically verified become more accurate; unverified records remain distorted. This adaptive approach requires behavioral telemetry and a decision engine that classifies query patterns, with rate limiting to prevent the adversary from using the adaptation mechanism as an oracle.

10. Conclusion

MOIRÉ is not a silver bullet. Against sophisticated, MOIRÉ-aware adversaries with independent ground truth, full-environment distortion will eventually be detected. The construction's value lies in three places: as a Tier 3 response against unsophisticated threats where full distortion is effective; as the enabling mechanism for a graduated deception architecture that adapts to threat sophistication; and as the formal basis for a turned-asset doctrine that transforms insider compromise from a containment problem into an intelligence opportunity.

The primitives are not new. FPE, deception technology, and intelligence tradecraft all predate this work. The contribution is their composition into a coherent architecture with explicit security bounds, identified failure modes, and an honest assessment of where the construction works, where it degrades, and where it fails.

References

[1] Bellare, M., Rogaway, P., Spies, T. "The FFX Mode of Operation for Format-Preserving Encryption." NIST submission, 2010.

[2] NIST SP 800-38G. "Recommendation for Block Cipher Modes of Operation: Methods for Format-Preserving Encryption." 2016.

[3] Spitzner, L. "Honeypots: Tracking Hackers." Addison-Wesley, 2003.

[4] Stoll, C. "The Cuckoo's Egg." Doubleday, 1989.

[5] MITRE Engage Framework. engage.mitre.org. 2022.

[6] Heckman, K., Stech, F., Thomas, R., Schmoker, B., Tsow, A. "Cyber Denial, Deception and Counter Deception." Springer, 2015.

[7] NIST SP 800-53 Rev. 5. Controls SC-26, SC-30, SI-20.

[8] NIST SP 800-160 Vol. 2. "Developing Cyber-Resilient Systems." 2021.

[9] Kahlhofer, M., Rass, S. "Application Layer Cyber Deception without Developer Interaction." arXiv:2405.12852, 2024.

[10] Anagnostakis, K. et al. "Detecting Targeted Attacks Using Shadow Honeypots." USENIX Security, 2005.

[11] Araujo, F., Hamlen, K. "From Patches to Honey-Patches." ACM CCS, 2014.

[12] Wright, P. "Spycatcher." Viking, 1987.

[13] Clancy, T. "Patriot Games." Putnam, 1987.

[14] Thinkst Applied Research. "Canarytokens.org." canarytokens.org.

[15] SentinelOne / Attivo Networks. "ThreatDefend Platform." sentinelone.com.

[16] CounterCraft. "Deception-Powered Threat Intelligence." countercraftsec.com.

[17] Acalvio Technologies. "ShadowPlex." acalvio.com.

[18] Proofpoint / Illusive Networks. "Shadow." proofpoint.com.

[19] Rapid7. US Patent 11,303,675. "Containing compromised credentials using deception systems." 2022.

[20] SentinelOne / Attivo. US Patent 11,038,658. "Deceiving attackers in endpoint systems." 2021.

[21] Clark, J., Hengartner, U. "Panic Passwords: Authenticating under Duress." USENIX HotSec, 2008.

[22] Valeros, V., Rigaki, M., Garcia, S. "Attacker Profiling Through Analysis of Attack Patterns in Geographically Distributed Honeypots." arXiv:2305.01346, 2023.

[23] Huang, Y., Zhu, Q. "Duplicity Games for Deception Design with an Application to Insider Threat Mitigation." arXiv:2006.07942, 2020.

[24] Nissen, A., Kulyk, O. "The Password You Hope You Never Use." ACM CSCW Companion, 2025.

[25] Whaley, B. "Stratagem: Deception and Surprise in War." Artech House, 2007.

[26] Cohen, F. "The Deception Toolkit." 1997. all.net/dtk.