Journal of Economic Criminology 12 (2026) 100221

Contents lists available at ScienceDirect

Journal of Economic Criminology

journal homepage: www.journals.elsevier.com/journal-of-economic-criminology

Recurrence, scale, and risk-based monitoring in money laundering: Temporal displacement under cumulative enforcement

Endre J. Reite

NTNU Business School, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

JEL classification:
K42
G28
D82

Keywords:
Money laundering
enforcement
persistence
monitoring
displacement

ABSTRACT

Existing models of money laundering typically analyze how illicit actors adapt to enforcement by changing transaction size, fragmentation, or laundering method while treating interaction with the financial system as episodic. This paper instead models laundering as repeated interaction under risk-based monitoring. A launderer jointly chooses transaction scale and interaction frequency when monitoring intensity and compliance frictions rise with cumulative activity. Stronger recurrence-based monitoring lowers optimal frequency. When detection risk or monitoring frictions increase in both scale and persistence, the reduction in frequency is accompanied by larger transactions and greater organizational investment, implying consolidation rather than fragmentation. The paper therefore identifies a form of temporal displacement: enforcement reshapes the timing and concentration of laundering activity, so fewer transactions do not necessarily indicate deterrence.

1. Introduction

Economic models of money laundering typically analyze how illicit actors respond to enforcement by adjusting transaction characteristics or laundering methods. Launderers choose how to move illicit funds subject to detection risk, transaction costs, and penalties, consistent with the economic approach to crime (Becker, 1968). A central result of this literature is that enforcement rarely eliminates illicit activity. Instead, it reallocates behavior toward less intensively policed margins, generating displacement rather than deterrence (Stigler, 1970; Harrington, 1988; Polinsky and Shavell, 2007).

In the context of money laundering, this logic has been formalized in models that emphasize substitution across methods, channels, and transaction characteristics. Enforcement that targets conspicuous transactions induces shifts toward less visible methods, fragmentation, or alternative organizational forms (Takáts, 2011; Contreras and Villa Pérez, 2025). These models generate clear predictions about how laundering adapts when monitoring focuses on specific observable features. However, they share a strong and largely implicit assumption: interaction with the financial system is treated as episodic. Transaction attributes are optimized, but the frequency of interaction itself is fixed or left implicit. Persistence is therefore ruled out as a strategic margin.

This assumption sits uneasily with contemporary anti-money laundering practice. Modern AML systems do not evaluate transactions in isolation. Financial institutions aggregate transaction histories, update client risk profiles, and broaden scrutiny when concerns persist. Repeated interaction with the financial system therefore changes future monitoring conditions through mechanisms such as enhanced due diligence, expanded transaction review, repeated reporting, and, in some cases, termination of the customer relationship. Persistence is not merely a background feature of illicit activity; it is itself a source of cost and risk.

From the perspective of the monitored actor, this institutional design changes the nature of the enforcement problem. Each additional interaction with the financial system increases cumulative exposure to scrutiny, even when individual transactions are small or routine. The relevant choice is therefore not only how to structure a transaction, but how often to engage with monitored institutions at all. Existing models of money laundering abstract from this margin and, in doing so, miss a key channel through which enforcement reshapes behavior under risk-based AML regimes.

This paper develops a simple model in which monitoring frictions and detection risk increase with cumulative activity. A launderer chooses both transaction size and interaction frequency. Persistence therefore becomes costly in its own right, and enforcement directed at repeated activity induces adjustment along a temporal margin. As recurrence-based monitoring intensifies, optimal transaction frequency falls. When monitoring or detection links transaction scale to persistence, the remaining transactions become larger, generating consolidation rather than fragmentation.

E-mail address: endre.j.reite@ntnu.no.

https://doi.org/10.1016/j.jeconc.2026.100221
Received 25 December 2025; Received in revised form 31 March 2026; Accepted 5 April 2026
Available online 7 April 2026

The model also clarifies the role of organizational structure. Fixed costs associated with incorporation, intermediaries, or nominee arrangements enter as entry costs for persistent laundering. Structured laundering therefore reflects expected recurrence rather than transaction size alone. This mechanism helps explain why professional laundering arrangements emphasize continuity and organizational investment, and why intensified enforcement often induces reorganization rather than exit.

Although the analysis is motivated by money laundering, the underlying mechanism is more general. Any enforcement environment in which scrutiny and compliance costs escalate mechanically with repeated interaction creates incentives to manage cumulative exposure over time. Anti-money laundering provides a particularly clear application because history-dependent monitoring is institutionalized and explicit, but the logic extends to other settings characterized by cumulative enforcement.

The analysis examines how anti-money laundering enforcement changes laundering behavior when monitoring intensifies with repeated interaction rather than with transaction size alone. More specifically, when financial institutions aggregate behavior over time and escalate scrutiny as activity persists, do illicit actors fragment activity into many small transactions, or do they instead reduce interaction frequency and consolidate value into fewer, larger transactions? By modeling laundering frequency and transaction scale as joint strategic choices, the analysis studies whether risk-based AML generates temporal reorganization rather than conventional fragmentation or method substitution.

The paper proceeds as follows. Section 2 describes cumulative monitoring in contemporary AML enforcement. Section 3 situates the argument in the related literature. Section 4 presents the model. Section 5 characterizes the optimal choice of scale and persistence. Sections 6 through 11 discuss interpretation, extensions, empirical implications, and policy relevance. Section 12 concludes.

2. Institutional background: cumulative monitoring in AML enforcement

Contemporary AML monitoring also has broader consequences, including de-risking and financial exclusion when institutions respond to persistent compliance concerns by restricting or terminating relationships (Pavlidis, 2023). That broader context matters here because the model abstracts from institutional detail but centers a specific feature of modern AML practice: scrutiny accumulates over time.

In anti-money laundering (AML) systems, transactions are not evaluated in isolation. Financial institutions aggregate client activity over time and assess it against the customer's risk profile, stated business purpose, and prior behavior. Monitoring intensity therefore depends not only on transaction-specific attributes, but also on recurrence, pattern, and persistence.

Transaction-monitoring systems evaluate activity within a broader behavioral context, including transaction frequency, timing, counterparties, and consistency with a customer's expected profile. A single transaction may not trigger serious concern on its own, whereas repeated or patterned activity often prompts escalation. Regulatory guidance explicitly endorses this aggregated, behavior-based approach as part of risk-based AML supervision (Financial Action Task Force, 2014; Basel Committee on Banking Supervision, 2017).

The probability that activity leads to deeper review also rises as concerns persist. When initial alerts cannot be resolved, institutions typically widen the review to longer histories, related accounts, and ownership structures. Escalation may involve enhanced due diligence, source-of-funds requests, extended transaction review, and repeated internal or external reporting. These steps impose real compliance frictions on the monitored actor and increase the likelihood that illicit activity is uncovered (Financial Action Task Force, 2014).

At the limit, institutions may terminate the relationship when perceived risk exceeds expected returns. For actors that rely on repeated access to formal finance, exclusion is itself a substantial penalty. The risk of exclusion may therefore rise even in the absence of a criminal sanction or final enforcement action.

The key implication is that AML enforcement is cumulative. Monitoring intensity, investigative scrutiny, and the threat of exclusion evolve with an actor's history of interaction rather than resetting after each transaction. Persistence is therefore not just a background feature of laundering; it is part of the cost structure. The model developed below abstracts from institutional detail, but it is designed to capture precisely this feature of contemporary AML regimes.

3. Related literature

This paper sits at the intersection of the economic literature on enforcement and the literature on money laundering under regulatory monitoring. A central result in the economics of enforcement is that when penalties or monitoring capacity are limited, enforcement often reallocates behavior toward less intensively policed margins rather than eliminating it outright (Becker, 1968; Stigler, 1970; Harrington, 1988; Polinsky and Shavell, 2007). For money laundering, the implication is that observed behavioral change following tighter enforcement should not automatically be read as a reduction in underlying illicit activity.

In money-laundering models, this logic has usually been developed through substitution across methods, channels, and transaction characteristics. Takáts (2011) shows how enforcement directed at conspicuous transactions can induce shifts toward less visible methods, while Contreras and Villa Pérez (2025) analyze strategic trade-offs between smurfing, layering, and over-diversification. These models generate clear predictions about fragmentation and channel choice when monitoring focuses on observable transaction features.

Recent work also clarifies adjacent margins of adaptation in money laundering. Tiwari et al. (2023) examine the choice among laundering techniques, Makmur (2024) highlights displacement toward informal remittance channels when enforcement is concentrated in formal finance, and Atkins (2024) shows that the deterrent effect of enforcement depends not only on detection but also on the credibility of sanctions. Against that background, the present paper focuses on a different margin of adjustment: how cumulative monitoring affects the frequency with which actors engage with monitored institutions.

The paper therefore brings together the enforcement logic of Becker (1968), Harrington (1988), and Polinsky and Shavell (2007) with laundering-specific accounts of adaptation under monitoring (Takáts, 2011; Contreras and Villa Pérez, 2025), while extending that literature by endogenizing persistence.

A common limitation of this literature is that laundering decisions are treated as episodic. The strategic problem concerns how to structure a transaction or which channel to use, while the frequency of interaction with the financial system is fixed or left implicit. That simplification is analytically convenient, but it rules out regimes in which monitoring responds mechanically to accumulated behavior rather than to isolated actions.

The question in this paper is different. Rather than asking how illicit actors adapt within a given level of interaction with the enforcement system, it asks how enforcement changes the level of interaction itself. That distinction matters because contemporary AML systems aggregate activity over time and escalate scrutiny as concerns persist, altering future compliance costs and detection risk even when transaction characteristics remain unchanged.

This focus on persistence also connects to work on the organization of criminal activity. Reuter (1983), Gambetta (1993), and Gambetta and Reuter (1995) emphasize that sustained illicit activity requires continuity, governance, and organizational investment rather than one-off optimization. Those insights motivate the treatment of fixed organizational costs here as entry costs for persistent laundering rather than as investments driven by transaction size alone.

The paper also relates to the broader literature on dynamic

enforcement and compliance history. Harrington (1988) shows how enforcement can depend on prior conduct, Kleven et al. (2011) document durable behavioral responses to audit threats, and Polinsky and Shavell (2000) provide a general treatment of repeat offenders. In many of those settings, however, history dependence operates through inference or discretionary targeting.

What is distinctive in AML is that history dependence is institutionalized and often mechanical. Risk classification, enhanced due diligence, suspicious activity reporting, and account termination all make scrutiny cumulative even without any formal model of latent criminal type. The framework below therefore abstracts from belief updating and strategic enforcement and instead treats cumulative monitoring as a feature of institutional design.

The paper complements existing theories of fragmentation, method substitution, and channel displacement by showing that cumulative monitoring turns persistence into a strategic margin. Enforcement then reshapes behavior not only by changing how laundering is conducted, but also by changing how often actors are willing to interact with monitored institutions.

4. Model and methodological positioning

The model follows the economic literature on crime and public enforcement by treating laundering as an optimization problem under detection risk, monitoring frictions, and sanctions. It extends money-laundering models that explain adaptation through transaction size, fragmentation, or substitution across methods and channels, but differs from them in one specific respect: it makes the frequency of interaction with the monitored financial system explicit. The objective is not to reproduce the full institutional complexity of AML systems, but to isolate a mechanism that existing models largely leave implicit: under history-dependent scrutiny, recurrence itself becomes part of the optimization problem.

Consider a launderer who chooses transaction size $x \geq 0$ and laundering frequency $R \geq 0$, where R is interpreted as the expected number of interactions with the formal financial system within a given aggregation window.

Expected total payoff is given by:

$$\Pi(x,R) = R\pi(x,R) - \mathbb{1}\{R>0\}F$$

where $F \geq 0$ is a fixed organizational cost capturing incorporation, nominee arrangements, or other investments required to operate repeatedly within the financial system.

Per-episode expected payoff is

$$\pi(x,R) = p(x,R)[(1-\tau)x - c] - (1-p(x,R))\kappa - \varphi(x,R),$$

where $p(x,R) \in (0,1)$ denotes the probability that a transaction is not detected, $\tau \in (0, 1)$ is an effective tax or wedge applied to laundered funds, $c > 0$ is a per-transaction cost, $\kappa > 0$ is the penalty incurred upon detection, and $\varphi(x, R)$ captures monitoring frictions that arise even when transactions are not formally detected.

Detection risk is assumed to be separable in transaction size and recurrence:

$$p(x,R) = p_x(x)p_R(R), \quad p_x'(x) < 0, \quad p_R'(R) < 0.$$

Larger transactions are more likely to attract scrutiny, and repeated interaction with the financial system reduces the probability that any given transaction escapes detection. This separability does not preclude interaction in marginal effects, as transaction size and recurrence may still jointly influence detection risk.

Monitoring frictions are assumed to contain both a scale-dependent component and a recurrence-dependent component:

$$\varphi(x,R;\eta) = \varphi_0(x) + \eta h(R), \eta > 0, h'(R) > 0, h''(R) > 0.$$

Here $\varphi_0(x)$ captures scale-dependent frictions, while h(R) captures frictions that rise with repeated interaction. The parameter η indexes the intensity of recurrence-based monitoring. This specification makes clear that total monitoring frictions are the sum of a scale term and an η-weighted recurrence term.

The convexity of h(R) captures escalating scrutiny and operational frictions that intensify with persistence. These frictions include repeated documentation requests, enhanced due diligence, account restrictions, and the growing risk of relationship termination.

The baseline specification therefore treats cumulative monitoring as operating primarily through recurrence. Cross-effects between scale and recurrence are introduced later as extensions rather than maintained assumptions, which allows the core result on laundering frequency to be derived under minimal structure.

Although the model is static for tractability, it should be read as a reduced-form representation of behavior within a fixed aggregation window during which financial institutions accumulate past activity and adjust scrutiny accordingly. In that sense, current recurrence affects both current costs and the monitoring conditions faced in subsequent interactions.

The model abstracts from institutional detail while capturing a central feature of modern AML regimes: monitoring frictions that escalate with cumulative activity. This structure allows laundering frequency to emerge as an endogenous choice variable alongside transaction size.

5. Optimal scale and persistence under recurrence-based monitoring

This section characterizes the launderer's joint choice of transaction size and laundering frequency under cumulative monitoring. The analysis highlights a frequency–scale trade-off that arises when monitoring frictions escalate with repeated interaction.

To isolate the role of recurrence, first abstract from the fixed organizational cost $F$and consider the launderer's problem of choosing $x$and $R$to maximize expected profits

$$\Pi(x,R) = R\,\pi(x,R).$$

Interior solutions satisfy the first-order conditions

$$\frac{\partial \Pi}{\partial x} = R\,\pi_x(x,R) = 0, \frac{\partial \Pi}{\partial R} = \pi(x,R) + R\,\pi_R(x,R) = 0,$$

which jointly determine the optimal transaction size $x^*$and laundering frequency R^*.

Let recurrence-based monitoring intensity be indexed by $\eta > 0$, and suppose monitoring frictions take the form

$$\phi(x,R;\eta) = \widetilde{\phi}(x,R) + \eta h(R),$$

so that an increase in ηraises the marginal cost of persistence.

Proposition 1. Frequency–scale trade-off

An increase in recurrence-based monitoring intensity strictly reduces optimal laundering frequency. Conditional on monitoring frictions or detection technologies that generate scale–recurrence interactions, optimal transaction size increases. In the absence of such interactions, transaction size need not respond.

Formally,

$$\frac{\partial R^*}{\partial \eta} < 0 (\text{unconditionally}),$$

$$\frac{\partial x^*}{\partial \eta} > 0 \text{if scale and recurrence interact.}$$

5.1. Proof

Expected profits are given by

$$\Pi(x,R;\eta) = R\,\pi(x,R;\eta).$$

Define

$$F_1(x,R;\eta) \equiv \frac{\partial \Pi}{\partial x} = R\,\pi_x(x,R;\eta),$$

$$F_2(x,R;\eta) \equiv \frac{\partial \Pi}{\partial R} = \pi(x,R;\eta) + R\,\pi_R(x,R;\eta).$$

Interior optimality implies $F_1 = 0$and $F_2 = 0$. Since $R > 0$, the first condition is equivalent to $\pi_x(x,R;\eta) = 0$.

Under the maintained specification, monitoring intensity enters payoffs only through $\phi(x,R;\eta)$. In the benchmark case where monitoring intensity does not affect the marginal cost of scale, $\phi_{x\eta} = 0$, implying

$$\frac{\partial F_1}{\partial \eta} = 0.$$

By contrast,

$$\frac{\partial F_2}{\partial \eta} = -(\phi_\eta + R\phi_{R\eta}) = -(h(R) + Rh'(R)) < 0,$$

since $h'(R) > 0$.

Let Jdenote the Jacobian of the system $(F_1,F_2)$with respect to (x,R). Standard second-order conditions for a strict local maximum imply $R\pi_{xx} < 0$and $\det(J) > 0$. By the implicit function theorem,

$$\begin{pmatrix} \partial x^*/\partial \eta \\ \partial R^*/\partial \eta \end{pmatrix} = -J^{-1}\begin{pmatrix} F_{1\eta} \\ F_{2\eta} \end{pmatrix}.$$

Since $F_{1\eta} = 0$, it follows that

$$\frac{\partial R^*}{\partial \eta} = \frac{-(R\pi_{xx})F_{2\eta}}{\det(J)} < 0,$$

because $R\pi_{xx} < 0$, $F_{2\eta} < 0$, and $\det(J) > 0$.

Similarly,

$$\frac{\partial x^*}{\partial \eta} = \frac{(R\pi_{xR})F_{2\eta}}{\det(J)}.$$

Since $F_{2\eta} < 0$, the sign of $\partial x^*/\partial \eta$is opposite that of π_{xR}. Hence $\partial x^*/\partial \eta > 0$whenever $\pi_{xR} < 0$, that is, whenever higher recurrence raises the marginal cost of scale. A sufficient condition is a positive scale–recurrence interaction in monitoring frictions, for example $\widetilde{\phi}(x,R) = \phi_0(x) + sxR$with $s > 0$, or detection technologies in which transaction size becomes more informative when activity is persistent. ∎

5.2. Interpretation and examples

The proposition isolates a simple mechanism. When monitoring costs escalate with repeated interaction, launderers optimally reduce the number of transactions to limit cumulative scrutiny. Persistence becomes costly. If enforcement technologies also link scale to persistence, the reduction in frequency lowers the marginal cost of concentrating volume and the remaining transactions become larger.

This logic contrasts with size-based enforcement. When monitoring focuses on transaction size but abstracts from history, launderers fragment transactions to remain below detection thresholds. Under recurrence-based monitoring, frequent low-value transactions are precisely what generate scrutiny. The dominant response can therefore be temporal consolidation rather than fragmentation.

The mechanism is consistent with escalation practices in AML compliance. Repeated alerts commonly trigger broader review of accounts, counterparties, and historical transactions, making subsequent activity more costly regardless of its size. In that environment, stronger recurrence-based monitoring induces fewer monitored interactions and may also induce greater organizational investment when persistent access to the formal financial system remains valuable.

Symbol	Definition	Behavioral role	Observability
x	Transaction size	Scale of laundering per episode	Observable (via SARs, FIU data)
R	Laundering frequency (recurrence)	Strategic choice of interaction rate	Observable (transaction count per entity)
$\pi(x,R)$	Per-episode expected payoff	Drives optimal laundering strategy	Modeled
$p(x,R)$	Probability of non-detection	Determines expected gain vs. penalty	Latent (proxied via SAR outcomes)
$\phi(x,R)$	Monitoring frictions (non-detection costs)	Captures compliance burden and scrutiny	Partially observable (e.g., EDD flags, account restrictions)
F	Fixed organizational cost	Entry cost for structured laundering	Latent (inferred from complexity of laundering setup)
η	Intensity of recurrence-based monitoring	Enforcement parameter affecting marginal cost of R	Partially observable (via policy thresholds, escalation rules)
τ	Effective tax/wedge on laundered funds	Reduces net return from laundering	Modeled or estimated
κ	Penalty upon detection	Shapes deterrence effect	Policy parameter (observable)

6. Persistence as an enforcement state variable

Persistence can be treated as an enforcement state variable rather than as a background feature of illicit activity. In standard economic models of crime, expected enforcement costs depend mainly on contemporaneous actions, transaction size, method, or other observable features, while past behavior either does not enter the payoff function or matters only through belief updating by authorities (Becker, 1968; Polinsky and Shavell, 2007). As a result, the frequency of illicit interaction is usually fixed or implicit.

Modern AML enforcement departs from that abstraction. Regulatory guidance and supervisory practice emphasize cumulative, actor-based monitoring in which scrutiny escalates with repeated interaction rather than solely through inference about latent criminal type (Financial Action Task Force, 2014; Basel Committee on Banking Supervision, 2017). Empirical evidence that behavioral history improves the identification of future suspicious activity is consistent with this institutional logic (Reite et al., 2025).

In this setting, laundering frequency influences the enforcement environment itself. Each additional interaction can intensify later monitoring through enhanced due diligence, broader transaction review, repeated suspicious activity reporting, and the threat of relationship termination (Levi and Reuter, 2006; Lord, 2018). From the launderer's perspective, persistence matters not only because exposure is repeated, but because the conditions of scrutiny worsen as activity continues.

This creates a sharp contrast with size-based enforcement. When monitoring is driven mainly by transaction size, the marginal cost of scale is high but largely independent of history, which encourages fragmentation strategies such as smurfing (Takáts, 2011; Contreras and Villa Pérez, 2025). Under recurrence-based monitoring, by contrast, the marginal cost of an additional interaction rises with cumulative activity, discouraging frequent engagement with the monitored system.

The frequency-scale trade-off derived in Section 5 can therefore be read as a standard substitution effect along a previously unmodeled margin. As recurrence-based monitoring intensifies, the shadow price of persistence rises. Actors substitute away from repeated interaction and, when scale and recurrence interact, toward fewer and larger transactions. This response does not require learning, belief updating, or strategically sophisticated enforcement. It follows mechanically from institutional rules that condition scrutiny on accumulated activity.

Under cumulative monitoring, the relevant question is not only how to launder, but how often to engage with the monitored system. Persistence becomes an object of optimization in its own right.

7. Discussion: episodic versus persistent laundering

The discussion turns on a distinction between episodic and persistent laundering defined by temporal organization rather than by transaction form alone. That distinction emerges endogenously from the interaction between the timing of illicit revenue, cumulative monitoring, and fixed organizational costs. In that respect, the argument complements earlier work on fragmentation, substitution, and organizational adaptation by shifting attention from how funds are disguised to how repeated exposure to monitored institutions is managed over time.

Episodic laundering is more attractive when illicit proceeds arise from one-off or infrequent events, or when the fixed costs of establishing a persistent laundering channel are large relative to expected recurrence. In those settings, actors have little reason to invest in organizational infrastructure or to manage long-run monitoring exposure. They may instead tolerate relatively high per-transaction detection risk in order to minimize repeated interaction with formal institutions. Large corruption cases are a useful illustration: proceeds arrive in discrete amounts, and the main problem is integration rather than flow management.

Persistent laundering characterizes activities that generate a continuing stream of illicit revenue. Here the relevant constraint is not the visibility of any single transaction but the cumulative exposure created by repeated interaction. When monitoring frictions escalate with recurrence, the framework suggests temporal consolidation: fewer laundering episodes, each of larger scale, combined with stronger incentives to invest in organizational arrangements that help sustain access. That interpretation is consistent with empirical accounts of professional laundering that emphasize continuity, intermediaries, and repeated access to the financial system (Levi and Reuter, 2006; Lord, 2018).

What the model adds to that literature is a more specific enforcement mechanism. Repeated interaction becomes costly not only because exposure is repeated, but because monitoring itself escalates as activity persists. That distinction helps explain why organizational investment can be rational even when transaction size alone does not justify it.

The same logic helps explain heterogeneity across crime types without resorting to ad hoc differences in sophistication or preferences. Drug trafficking, large-scale fraud, and smuggling often generate steady cash flow and therefore make persistent laundering and organizational investment more attractive. Episodic strategies are more plausible when revenue is irregular or when recurrence itself is prohibitively costly because cumulative monitoring is intense.

Cumulative monitoring also changes the form of displacement. Under size-based or method-based enforcement, displacement operates across channels or transaction forms. Under recurrence-based monitoring, it also operates across time. Fewer observed transactions may therefore reflect a reorganization of exposure under cumulative scrutiny rather than lower laundering demand, which is consistent with the broader literature showing that tighter AML pressure can reallocate illicit activity rather than eliminate it (Contreras and Villa Pérez, 2025).

8. Heterogeneity, channel displacement, and institutional context

The strength of recurrence-based monitoring and the cost of organizational investment are unlikely to be uniform across actors, channels, or institutional environments. This heterogeneity generates systematic variation in laundering strategies, even when underlying criminal activities are similar.

Differences across actors arise naturally from variation in organizational capacity. Professional laundering organizations face lower effective fixed costs of operation and are better equipped to manage compliance frictions associated with cumulative monitoring. For these actors, persistence is a feasible and often optimal strategy. Opportunistic or small-scale offenders, by contrast, face higher organizational costs and weaker capacity to navigate monitoring systems. For them, episodic laundering remains relatively more attractive, even when it entails higher per-transaction exposure. The model thus predicts sorting across laundering strategies based on organizational capacity rather than differences in risk preferences or sophistication.

Temporal displacement also interacts with channel choice. Recurrence-based monitoring is most effective where identity persistence and transaction linkage are strong, as in regulated banking systems. When monitoring frictions escalate steeply with repeated interaction in those environments, actors have stronger incentives to shift activity toward channels in which continuity is harder to observe or aggregate. The implication is that recurrence-based monitoring may induce displacement not only across time, but also toward settings where transaction histories are more weakly linked across persons, accounts, or platforms. This complements recent work showing that AML scrutiny concentrated in formal finance may leave informal remittance channels comparatively exposed (Makmur, 2024).

In crypto-asset markets, persistence may be harder to monitor because activity can be distributed across wallets, platforms, and jurisdictions. The FATF travel rule (Recommendation 16) is intended to strengthen continuity by requiring originator and beneficiary information to accompany transfers, making repeated activity easier to link beyond the banking sector (Financial Action Task Force, 2021, 2025). For the mechanism developed here, that distinction matters. When linkage is weak, spacing and identity resets are cheaper. When linkage is credible, recurrence becomes more expensive and temporal consolidation becomes more plausible.

Institutional context further conditions these responses. Regulatory harmonization efforts, such as attempts to standardize AML supervision and monitoring practices across jurisdictions, can be interpreted as attempts to align recurrence-based monitoring intensity. The framework suggests that harmonization may have heterogeneous effects across institutional environments. In institutional environments with high organizational costs and strong enforcement capacity, intensified monitoring may suppress persistent laundering altogether or induce relocation. In environments where organizational costs are lower and enforcement capacity is weaker, the same regulatory tightening may instead increase incentives for structured, persistent laundering by raising the returns to organizational investment.

These interactions help explain why empirical assessments of AML effectiveness often yield heterogeneous results. Changes in enforcement intensity along one margin may induce reallocation across actors, channels, or jurisdictions rather than proportional reductions in illicit activity. By treating persistence as a strategic margin, analysis clarifies how recurrence-based monitoring reshapes the composition and organization of laundering activity in ways that depend critically on institutional context.

9. Comparing size-based and persistence-based AML regimes

The contrasts across enforcement architectures are summarized in Table 2.

Under size-based monitoring, enforcement intensity increases with transaction scale but is largely independent of transaction history. The marginal cost of scale is high, while the marginal cost of an additional transaction is comparatively low. The canonical response in this environment is fragmentation: illicit funds are divided into smaller transactions to remain below detection thresholds. This logic underpins standard smurfing models and predicts an increase in transaction counts when enforcement tightens (Takáts, 2011; Contreras and Villa Pérez, 2025).

Under method-based enforcement, monitoring targets specific laundering techniques or channels. Behavioral responses take the form of substitution across methods—shifting between channels with different visibility profiles—while leaving the temporal structure of activity largely unchanged. These regimes shape how laundering is

Table 2
Behavioral responses under alternative AML monitoring regimes.

Monitoring regime	Primary enforcement signal	Typical laundering adaptation	Empirical signature
Size-based monitoring	Transaction amount relative to thresholds	Fragmentation / 'smurfing' to remain below thresholds	More transactions, smaller average size; clustering around thresholds
Method-based monitoring	High-risk channels, products, or typologies	Channel substitution (shift to less-monitored methods)	Shifts in method mix; stable counts/size within surviving channels
Recurrence-based monitoring	Cumulative activity or recurrence over a look-back horizon	Temporal consolidation: fewer episodes, potentially larger transactions; organizational investment to sustain access	Fewer transactions, larger average size; spacing or bursts depending on window length and update speed

conducted, but not how often actors interact with the financial system.

Recurrence-based monitoring operates differently. When enforcement intensity escalates with cumulative interaction, the marginal cost of an additional transaction rises with recurrence. Frequent engagement with the monitored system becomes costly even when individual transactions are small or routine. The mechanism implies temporal consolidation: fewer transactions of larger scale, potentially accompanied by greater organizational investment to manage scrutiny. In this setting, laundering adapts by economizing on interaction frequency rather than by fragmenting transactions or switching methods.

When enforcement combines these approaches—as is increasingly the case in contemporary AML regimes—observed laundering strategies reflect the interaction of multiple enforcement margins. Size-based instruments encourage fragmentation, persistence-based instruments encourage consolidation, and method-based instruments induce channel substitution. This interaction helps explain why empirical patterns often appear heterogeneous and why one-dimensional enforcement assessments may misclassify displacement as deterrence. A decline in transaction counts, for example, may reflect intensified recurrence-based monitoring rather than a reduction in illicit activity.

By isolating persistence as a strategic margin, the paper provides a framework for interpreting these mixed enforcement environments. Differences in observed laundering behavior need not reflect different levels of sophistication or intent. They may instead reflect which enforcement margin—size, method, or persistence—is most salient in a given institutional setting.

10. Empirical implications and mapping to data

The model yields empirically testable implications that differ sharply from those generated by size-based or method-based enforcement models. These implications concern the temporal organization of laundering activity rather than its aggregate volume or choice of technique. Consistent with existing empirical findings, intensified enforcement is expected to reshape behavior rather than eliminate illicit flows.

The central prediction is that increases in recurrence-based monitoring reduce transaction frequency while leaving total illicit volume ambiguous. When monitoring intensity escalates with cumulative interaction, launderers economize on the number of monitored transactions. Conditional on enforcement technologies that link scale to persistence, this reduction in frequency is accompanied by an increase in average transaction size. Observed consolidation should therefore be interpreted as a behavioral response to cumulative monitoring rather than as evidence of deterrence.

These predictions map naturally to observable measures. Laundering frequency corresponds to transaction counts per entity or beneficial owner within a defined aggregation window. Persistence can be proxied by the duration and continuity of account activity or by repeated interaction with compliance functions. Monitoring pressure may be measured using indicators such as the initiation of enhanced due diligence, the intensity of suspicious activity reporting, or account-level risk scores, all of which explicitly incorporate behavioral history in contemporary AML systems (Financial Action Task Force, 2014; Basel Committee on Banking Supervision, 2017).

Exogenous changes in recurrence-based monitoring provide natural settings for empirical evaluation. Regulatory or technological shifts that shorten aggregation windows, lower transaction-count thresholds, or expand the use of automated behavioral scoring increase the marginal cost of persistence.

The model implies that stronger recurrence-based monitoring reduces transaction frequency. When scale interacts with recurrence, the adjustment may also take the form of larger average transaction size, conditional on a given volume of illicit funds. These comparative statics contrast with size-based enforcement models, which predict greater fragmentation when enforcement tightens (Takáts, 2011; Contreras and Villa Pérez, 2025).

Several data environments are well suited to evaluating these implications. Financial intelligence unit microdata and panels of suspicious activity reports can reveal transaction frequency, escalation, and reporting intensity over time. Supervisory datasets increasingly contain longitudinal risk assessments and escalation flags that reflect cumulative monitoring. Enforcement case files and leak-based datasets can provide complementary evidence on organizational structure and persistence across entities. Existing empirical work already shows that incorporating behavioral history improves detection performance, which is consistent with the paper's emphasis on persistence as a monitoring target (Reite et al., 2025).

The framework also suggests implications for organizational form. Conditional on a given volume of illicit funds, laundering conducted through persistent interaction with the financial system may be associated with greater legal and organizational complexity, including multiple entities, intermediaries, and layered ownership structures. These features reflect the role of fixed organizational costs in managing cumulative scrutiny rather than attempts to conceal individual transactions. Episodic laundering, by contrast, should exhibit higher per-transaction exposure but lower organizational investment, consistent with typologies observed in enforcement cases (Levi and Reuter, 2006; Lord, 2018).

Finally, the framework clarifies why declines in transaction counts should not be interpreted mechanically as deterrence. Reductions in frequency accompanied by increases in transaction size are consistent with intensified recurrence-based monitoring even when total illicit flows remain unchanged. Distinguishing temporal displacement from genuine deterrence therefore requires joint analysis of frequency, scale, and organizational structure rather than reliance on transaction counts alone.

11. Policy implications

The main policy implication is interpretive. Under cumulative, actor-based monitoring, a decline in transaction counts is not sufficient evidence of deterrence. The same pattern may reflect temporal consolidation: actors reduce the number of monitored interactions while concentrating value into fewer, larger transactions. Evaluating AML effectiveness therefore requires information on frequency, transaction size, persistence, and organizational structure rather than counts alone.

This perspective also clarifies how aggregate enforcement outcomes should be read. System-level evidence on AML effectiveness does not by itself identify the behavioral margin through which adaptation occurs. The contribution here is narrower: it identifies temporal reorganization as one such margin and shows why fewer observed transactions need not imply less laundering.

A second implication concerns spillovers across channels. Recurrence-based monitoring in one part of the financial system can increase incentives to move activity toward settings in which continuity is harder to observe, such as cash-intensive businesses, informal remittance corridors, opaque corporate vehicles, or segments of virtual-asset markets. Makmur (2024) illustrates the broader point: when scrutiny is concentrated in formal finance, activity can shift toward channels in which continuity is harder to reconstruct. In adjacent domains, travel-rule requirements are intended to preserve identity linkage across crypto-asset transfers (Financial Action Task Force, 2021, 2025). Persistence-based strategies are therefore only as effective as the system's ability to keep identity and transaction history linkable across those domains.

A third implication is that recurrence-sensitive detection has limited bite unless escalation is paired with credible consequences. Atkins (2024) argues that criminal prosecution can transmit a different deterrence signal than repeated civil fines in AML enforcement. In the terms of the model, weak sanctions reduce the shadow cost of detection and make cooling-off, clustering, or channel switching more attractive than genuine deterrence. Recurrence should therefore be treated as a first-order risk signal rather than as a by-product of transaction screening.

These points reinforce the limits of one-dimensional enforcement. Size-based tools encourage fragmentation, method-based tools encourage substitution, and persistence-based tools encourage temporal consolidation or channel shifting when used in isolation. Effective AML supervision therefore requires coordinated attention to scale, frequency, and organizational form, ideally aggregated at the level of beneficial ownership rather than individual accounts. More generally, policy evaluation should distinguish deterrence from displacement by examining how multiple margins move together over time.

12. Conclusion

This paper develops a simple theoretical model in which laundering frequency—the rate at which actors interact with the formal financial system—is an endogenous choice under cumulative monitoring. That move departs from models that focus on transaction size, fragmentation, or method choice under essentially episodic enforcement. When monitoring intensity and compliance frictions rise with repeated activity, persistence itself becomes costly.

The central result is a frequency-scale trade-off. As recurrence-based monitoring intensifies, actors optimally reduce the number of monitored interactions. When detection or monitoring links transaction scale to persistence, the reduction in frequency is accompanied by consolidation into fewer, larger transactions. The paper therefore identifies a form of temporal displacement that complements existing accounts of fragmentation and channel substitution.

The model also clarifies the role of organizational structure. Fixed organizational costs operate as entry costs for persistent laundering rather than as simple responses to transaction size. Structured laundering is therefore tied to expected recurrence and to the need to manage cumulative scrutiny over time. That helps explain why professional laundering arrangements emphasize continuity and why enforcement pressure often induces reorganization rather than exit.

Although the analysis is motivated by money laundering, the underlying mechanism is more general. Wherever scrutiny escalates mechanically with repeated interaction, actors have incentives to manage cumulative exposure rather than isolated actions. AML is a particularly clear application because history-dependent monitoring is explicit and institutionalized, but similar dynamics may arise in other settings characterized by cumulative supervision.

The framework necessarily abstracts from several features that may matter in practice, including strategic interaction within laundering networks, learning by enforcement authorities, and explicit channel choice. A richer treatment of those elements would require a more fully dynamic model. Even so, treating persistence as a strategic margin offers a tractable way to understand how modern enforcement regimes reshape illicit behavior over time.

Funding

This work was supported by the Research Council of Norway under Grant No. 360861 and Grant No. 341289.

CRediT authorship contribution statement

Endre J. Reite: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

AI-assisted tools (GPT-5.2) were used solely for language editing and stylistic refinement. All analysis, arguments, and conclusions are the author's own.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author thanks participants in the FSA Regulatory Sandbox Project 2025 for helpful comments on earlier drafts of the paper.

Appendix A. Robustness

A.1 Non-separable detection risk

The baseline model assumes separability of detection risk in transaction size and recurrence. Allowing for interaction between these dimensions strengthens the frequency–scale trade-off.

Suppose the probability of non-detection satisfies

$$p(x,R), p_x < 0, p_R < 0, p_{xR} < 0.$$

The negative cross-partial implies that transaction size becomes more informative when activity is persistent. Per-episode expected payoff is

$$\pi(x,R) = p(x,R)[(1-\tau)x - c] - (1-p(x,R))\kappa - \phi(x,R).$$

Recurrence then affects the marginal profitability of scale both through monitoring frictions and through detection risk. An increase in recurrence-based monitoring raises the marginal cost of persistence directly via $\phi$and indirectly via p_{xR}. The tendency toward temporal consolidation is therefore

reinforced.

A.2 Discrete interpretation of laundering frequency

Laundering frequency is modeled as continuous for tractability. A discrete formulation leads to the same qualitative comparative statics.

Let the launderer choose an integer number of transactions $n \in \mathbb{N}$. Expected profit is

$$\Pi(x, n) = n\,\pi(x, n),$$

with monitoring frictions increasing and convex in n. If

$$\phi_n > 0, \phi_{nn} > 0,$$

the marginal cost of additional transactions rises with recurrence. An increase in recurrence-based monitoring reduces the optimal number of transactions. Conditional on total volume, this implies larger transaction size.

A.3 Organizational entry and fixed costs

Fixed organizational costs are treated as independent of transaction size in the main text. Allowing organizational costs to vary weakly with scale does not affect the role of persistence.

Suppose entry costs take the form

$$F(x) = F_0 + \psi x, \psi \geq 0.$$

Organizational investment remains unattractive for one-off activity and becomes optimal only when laundering is expected to persist. Expected recurrence, rather than scale alone, governs entry.

References

Atkins, M., 2024. Should banks face criminal prosecution for breaches of Money Laundering Regulations or are civil fines effective? Analysis of the significance of the first ever criminal conviction of a bank (NatWest) for breaches of the money laundering regulations. J. Econ. Criminol. 6, 100097. https://doi.org/10.1016/j.jeconc.2024.100097.

Basel Committee on Banking Supervision, 2017. *Sound management of risks related to money laundering and financing of terrorism*. Bank for International Settlements, Basel.

Becker, G.S., 1968. Crime and punishment: an economic approach. J. Political Econ. 76 (2), 169–217. https://doi.org/10.1086/259394.

Contreras, A., Villa Pérez, E., 2025. Strategic choices in money laundering: smurfing, layering, and financial over-diversification. J. Econ. Criminol. 11, 100199. https://doi.org/10.1016/j.jeconc.2025.100199.

Financial Action Task Force, 2014. *Risk-based approach guidance for the banking sector*. FATF, Paris.

Financial Action Task Force, 2021. *Updated guidance: a risk-based approach to virtual assets and virtual asset service providers*. FATF, Paris.

Financial Action Task Force, 2025. *FATF updates standards on recommendation 16 on Payment transparency*. FATF, Paris.

Gambetta, D., 1993. The Sicilian Mafia: The Business of Private Protection. Harvard University Press, Cambridge, MA.

Gambetta, D., Reuter, P., 1995. Conspiracy Among the Many: The Mafia in Legitimate Industries. In: Fielding, N.G., Clarke, A., Witt, R. (Eds.), The Economic Dimensions of Crime. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-62853-7_5.

Harrington, W., 1988. Enforcement leverage when penalties are restricted. J. Public Econ. 37 (1), 29–53. https://doi.org/10.1016/0047-2727(88)90003-5.

Kleven, H.J., Knudsen, M.B., Kreiner, C.T., Pedersen, S., Saez, E., 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. Econometrica 79 (3), 651–692. https://doi.org/10.3982/ECTA9113.

Levi, M., Reuter, P., 2006. Money laundering. Crime. Justice 34 (1), 289–375. https://doi.org/10.1086/501508.

Lord, N., 2018. The corporate context of financial crime. In: Simpson, S., Lord, N. (Eds.), The Oxford Handbook of White-Collar Crime. Oxford University Press, Oxford.

Makmur, K.L., 2024. Why only scrutinise formal finance? Money laundering and informal remittance regulations in Indonesia. J. Econ. Criminol. 6, 100111. https://doi.org/10.1016/j.jeconc.2024.100111.

Pavlidis, G., 2023. The dark side of anti-money laundering: Mitigating the unintended consequences of financial action task force standards. J. Econ. Criminol. 2, 100040. https://doi.org/10.1016/j.jeconc.2023.100040.

Polinsky, A.M., Shavell, S., 2000. The economic theory of public enforcement of law. J. Econ. Lit. 38 (1), 45–76. https://doi.org/10.1257/jel.38.1.45.

Polinsky, A.M., Shavell, S., 2007. The theory of public enforcement of law. In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics, 1. Elsevier, Amsterdam, pp. 403–454. https://doi.org/10.1016/S1574-0730(07)01006-7.

Reite, E.J., Karlsen, J., Westgaard, E.G., 2025. Improving client risk classification with machine learning to increase anti-money laundering detection efficiency. J. Money Laund. Control 28 (1), 93–107. https://doi.org/10.1108/JMLC-03-2024-0040.

Reuter, P., 1983. Disorganized crime: The economics of the visible hand. MIT Press.

Stigler, G.J., 1970. The optimum enforcement of laws. J. Political Econ. 78 (3), 526–536. https://doi.org/10.1086/259646.

Takáts, E., 2011. A theory of "crying wolf": the economics of money laundering enforcement. J. Law Econ. Organ. 27 (1), 32–78. https://doi.org/10.1093/jleo/ewp018.

Tiwari, M., Ferrill, J., Gepp, A., Kumar, K., 2023. Factors influencing the choice of technique to launder funds: the APPT framework. J. Econ. Criminol. 1, 100006. https://doi.org/10.1016/j.jeconc.2023.100006.