Who Pays for Payment Fraud? Detection and Liability Rules under Strategic Fraudster Adaptation

Vimal Balasubramaniam Thomas Mosk Antoine Uettwiller*

June 16, 2026

Abstract

Payment fraud per capita has doubled across major economies since 2017, yet the intuitive regulatory response, shifting liability onto payment service providers (PSPs), ignores a central feature of fraud: fraudsters adapt. Using 1.19 million underground forum posts from CrimeBB, we provide the first direct micro-level evidence of fraudster adaptation: ChatGPT's release triggered a +97% increase in within-creator posts discussing AI–fraud, and stacked event studies across 27 anti-fraud technologies show countermeasure discussion growing after deployment while vulnerable-method discussion is unaffected. We embed these findings in a model of strategic adaptation and show that without liability PSPs underinvest, but under full liability competing PSPs overinvest: the optimal rule assigns partial liability. Reinvestment to counter decay generates a ratchet: each round of investment accelerates adaptation, making the next round necessary. Extending the framework to digital platforms, optimal liability rules require full liability for all detecting agents, plus a regulatory fee to prevent free riding.

Keywords: Payment fraud, fraud detection, operational risk, financial regulation, consumer protection

JEL Classification: G21, G23, G28, L51, O33, D62

*We thank Michele Fabi and Patrick Blonien (Discussants), Dirk Niepelt, Radoslawa Nikolowa, Maarten van Oordt, Tarun Ramadorai, Jason Sturgess, Konstantinos Zachariadis, and seminar participants at Queen Mary University of London, the 2nd Financial Fraud, Misconduct and Market Manipulation Conference, the CEPR Fintech and Digital Currencies RPN Annual Meeting 2025, the 2026 Fintech and Financial Institutions Research Conference at the Federal Reserve Bank of Philadelphia, and the 8th Future of Financial Information Conference. Balasubramaniam: Queen Mary University of London, CEPR. Email: v.balasubramaniam@qmul.ac.uk. Mosk: Queen Mary University of London. Email: t.mosk@qmul.ac.uk. Uettwiller: Queen Mary University of London. Email: a.uettwiller@qmul.ac.uk.

I Introduction

Payment fraud has emerged as one of the fastest-growing economic crimes of the digital age. Across the US, EU, UK, and China, fraud incidents per capita have roughly doubled since 2017, and fraud's share of total reported crime has risen in lockstep (Figure 1). The financial losses are large: in 2024 alone, US consumers lost $12.5 billion to payment fraud, UK consumers £341 million, and EEA consumers €4.3 billion.

Payment scams are difficult and costly for individual consumers to detect, whereas payment service providers (PSPs) can exploit scale economies and invest in advanced anti-fraud technologies. To align incentives, regulators have shifted liability onto PSPs, but coverage remains incomplete. In the US, Regulation Z requires credit card issuers to reimburse unauthorized fraud, and in practice all major card networks waive even the statutory $50 cap.[1] While most unauthorized fraud is reimbursed, authorized fraud, where the consumer is deceived into initiating the payment, is not covered by any federal rule, leaving a regulatory gap precisely for the fastest-growing fraud category.[2] The UK PSR's 2024 rules were the first to close this gap, requiring PSPs to reimburse victims of authorized push-payment fraud. Yet even where liability exists, this reasoning overlooks a central feature of fraud: fraudsters adapt. As PSPs intensify detection of prevalent scams, fraudsters substitute toward harder-to-detect techniques such as AI-generated impersonations (Stop Scams UK, 2023) or image-based malware delivery via messaging apps.[3] This adaptation creates a negative detection externality: when one PSP improves its models, it pushes fraud toward variants that are harder for all PSPs to detect or displaces it to other markets entirely.

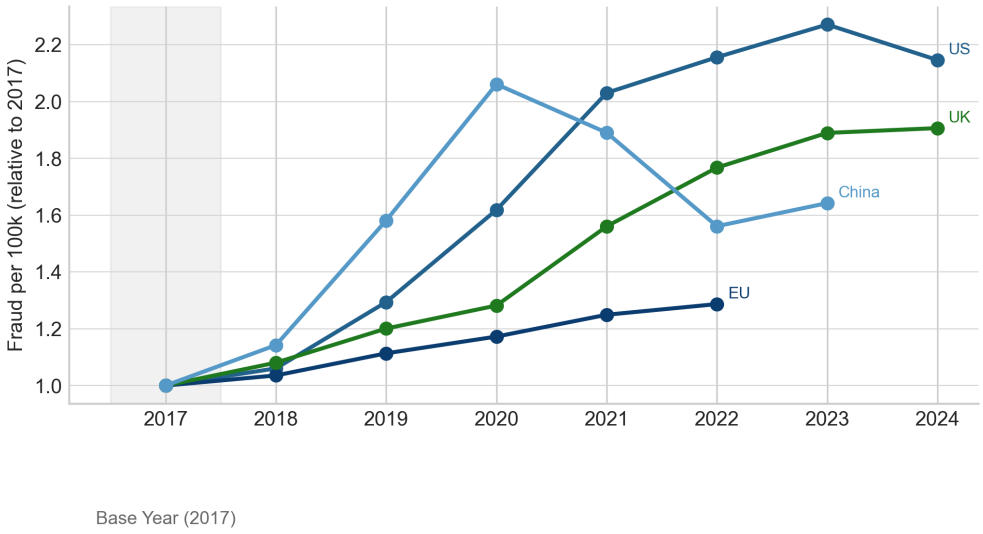
This paper studies how fraudster adaptation shapes the optimal intensity and allocation of detection investment. We provide the first direct micro-level evidence that criminal communities respond to detection technologies by discussing and disseminating countermeasures, consistent with substitution toward harder-to-detect methods. We then embed this evidence in a model that yields three results: partial liability corrects overinvestment driven by model decay in competitive markets, full liability for all corrects free-riding across PSPs and platforms, and partial liability again corrects overinvestment driven by cross-market displacement when agents detect entirely different fraud types.

[1] Peer-to-peer systems (Zelle, Venmo, PayPal) are covered by Reg E for unauthorized transactions only, with the same tiered exposure; and cryptocurrency transfers have no federal consumer-protection framework at all.

[2] In one case documented by TIME Magazine, a Wells Fargo customer authorized a $1,000 transfer to a scammer posing as the bank; Wells Fargo initially denied reimbursement, arguing that federal rules only cover unauthorized transactions. `https://time.com/6952817/financial-scams-banks`.

[3] See https://indianexpress.com/article/technology/he-downloaded-a-whatsapp-image-minutes-later-rs-2-lakh-was-gone-9951435/.

Panel A: Fraud per 100,000 inhabitants (Normalized to 2017)



Panel B: Fraud Share of Total Crime (Normalized to 2017)

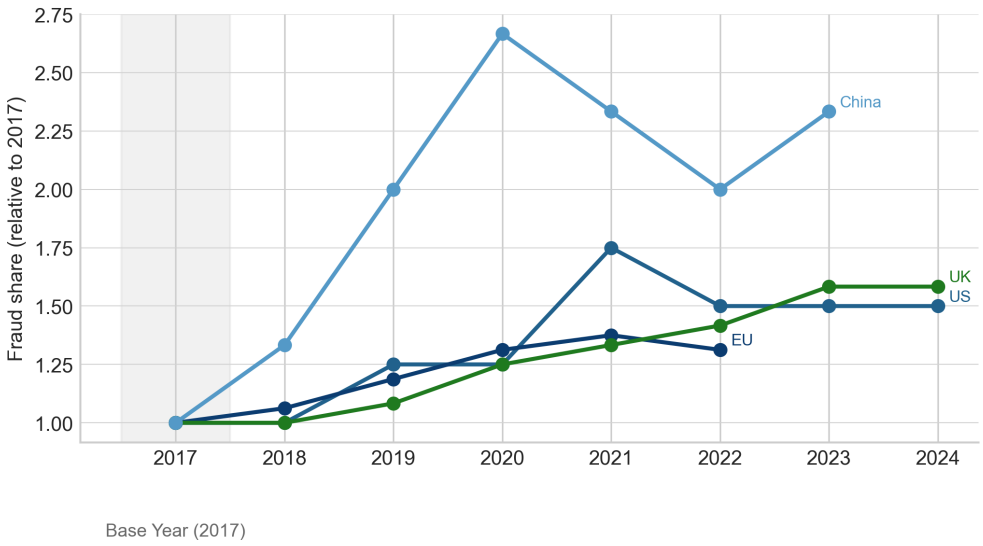


Figure 1
The Growth of Payment Fraud

Notes: Panel A shows fraud cases per 100,000 inhabitants in the EU, USA, UK, and China from 2017 to 2024, normalized relative to 2017. Panel B shows fraud as a share of total crime, also normalized relative to 2017. Sources: US: FBI UCR; EU: Eurostat; UK: ONS Crime in England and Wales; China: National Bureau of Statistics.

A critical assumption underlying both dynamics is that fraudsters adapt, developing and disseminating new circumvention techniques in response to detection technologies. The economics of crime literature shows that criminals respond to deterrence through spatial reallocation (Maheshri and Mastrobuoni, 2021), but not through method innovation. One reason is that official fraud statistics capture crime types but not the specific methods criminals use, making innovation difficult to observe in administrative data. We fill this gap using data from CrimeBB (Pastrana et al., 2018), one of the largest structured corpora of underground forum posts. From a panel of 1.19 million posts by 225,000 creators across five forums, we document two complementary patterns. First, the public release of ChatGPT in November 2022 serves as a natural experiment for the enabling-shock channel: within-creator AI–fraud co-discussion nearly doubled (+97%), with technically sophisticated creators driving early adoption and the effect subsequently broadening to less skilled creators, consistent with a general-purpose shock that lowers skill barriers for fraud. Second, stacked event studies around the deployment dates of 27 anti-fraud technologies reveal that countermeasure discussion grows monotonically after deployment, reaching significance at 24–35 months, while discussion of vulnerable methods is unaffected, consistent with substitution toward harder-to-detect methods.

We embed these empirical regularities in a model of payment fraud detection with strategic adaptation. The model yields four sets of results. First, when PSPs do not bear fraud losses, detection is underprovided; but under full liability in competitive markets, PSPs overinvest because each ignores how its detection intensity accelerates fraudster adaptation market-wide. The socially optimal rule assigns partial liability, with the optimal share strictly below one and increasing in the number of PSPs.[4] Second, when PSPs reinvest in detection to offset model decay, competition induces a ratchet effect: reinvestment restores model accuracy but simultaneously accelerates adaptation, raising the return to yet further investment. PSPs reinvest more frequently than a social planner would choose.

Third, fraud prevention depends on the joint effort of multiple agents, a PSP and digital platforms, whose detection efforts are complementary. A fraudulent transaction requires both a platform to fail to screen the solicitation and a PSP to fail to flag the payment; placing too much responsibility on one weakens incentives for the other, creating a free-rider problem. A Reuters investigation based on internal Meta documents revealed that the company internally projected earning approximately $16 billion annually from fraudulent advertisements while its platforms were estimated to be involved in one-third of all successful US scams, yet platforms face no statutory reimbursement obligation (Horwitz, 2025). The welfare-maximizing rule sets full liability for all agents whose efforts are complementary,

[4]We analyze consumer moral hazard in Internet Appendix H.

supplemented by a regulatory fee that internalizes the social cost of fraud. The UK PSR's 50/50 reimbursement split between sending and receiving PSPs falls short of this prescription, and the absence of platform liability represents the largest remaining regulatory gap.

Finally, we examine a third detection externality that arises when PSPs and platforms detect entirely different fraud types, for example, skimmed credit card fraud and eBay non-delivery fraud. Each domain's detection investment may displace fraud toward the other domain rather than eliminating it: a cross-market Whac-a-Mole. Because each domain ignores the migration costs it imposes on the other, full liability induces overinvestment and the optimal liability share is strictly below one for both agents.

Our paper makes three principal contributions. First, while the literature on fraud in financial markets has studied perpetrators (Dimmock et al., 2018; Egan et al., 2019; Piskorski et al., 2015), we shift focus to the institutions that prevent and compensate fraud. We develop a theory of payment fraud detection in which detection is endogenous: fraudsters adapt strategically to PSP investment, eroding its effectiveness over time. This contrasts with existing theories of information acquisition in financial markets (Hauswald and Marquez, 2006; Vives and Ye, 2025), which treat detection technology as fixed, and with FinTech competition models (Parlour et al., 2022), which abstract from adversarial adaptation. We derive optimal liability rules and show that the right rule depends on whether the dominant externality runs through adaptation, free-riding, or cross-market displacement.

Second, we provide the first direct micro-level evidence of criminal method innovation in response to detection technologies. Prior work documents the role of cryptocurrency in financing illegal activity (Foley et al., 2019; Griffin and Mei, 2024) and shows that profit pressure weakens bank incentives to report financial crime (Gao et al., 2020). We complement this evidence by showing that fraudsters respond to detection investment by shifting toward harder-to-detect methods, and that this response shapes optimal liability design.

Third, a growing literature studies the role of technology in financial decision-making (Fuster et al., 2022; Erel et al., 2021). We show that when both attackers and defenders innovate, technology investment generates externalities: detection improvements by one institution erode the effectiveness of others by shifting the distribution of fraud methods.

The remainder of the paper proceeds as follows. Section II presents empirical evidence of strategic adaptation and model decay using underground forum, Reddit, and Twitter data. Section III develops the theoretical model of fraud detection with strategic adaptation. Sections IV–VIII analyze optimal liability rules under various market structures and extensions. Section IX concludes.

II Empirical Evidence of Strategic Adaptation

Adversarial agents, whether fraudsters, spammers, or bot operators, adapt strategically to detection technologies, causing model performance to decay over time. We first develop a simple framework to study the effect of technological shocks on fraudster behavior and detection model performance. A critical assumption of the framework is that fraudsters adapt, developing and disseminating new circumvention techniques in response to technological changes. While the economics of crime literature establishes that criminals respond to deterrence and price incentives through spatial reallocation (Maheshri and Mastrobuoni, 2021) and cross-type substitution (Draca et al., 2019), there is no direct evidence that criminals adapt. One empirical challenge is that official fraud statistics report crime types but do not contain information about the specific methods used by criminals, making adaptation difficult to observe in administrative data. We use CrimeBB (Pastrana et al., 2018), one of the largest publicly available underground forum datasets, to study the effect of technological change on how criminals discuss and share new fraud methods. We present two analyses, exploiting the release of ChatGPT in November 2022 as a natural experiment for a general-purpose enabling shock, and a stacked event-study design covering 27 anti-fraud technology deployments.

II.A Technology Shocks, Fraud Composition, and Detection Decay

A fraudster allocates activity across K fraud types, indexed $k = 1, \ldots, K$, choosing shares $\pi_k \geq 0$ with $\sum_k \pi_k = 1$. A successful attack of type k yields payoff $V > 0$, and type k is detected with probability $\theta_k \in [0, 1]$. Scaling up any single type is increasingly costly, captured by a convex cost $\frac{1}{2}c\pi_k^2$. The fraudster's problem is:

$$\max_{\{\pi_k\}} \quad \sum_{k=1}^{K}(1-\theta_k)V\pi_k - \frac{c}{2}\sum_{k=1}^{K}\pi_k^2 \qquad \text{s.t.} \quad \sum_{k=1}^{K}\pi_k = 1, \quad \pi_k \geq 0. \tag{1}$$

Solving yields:

$$\pi_k^* = \frac{1}{K} + \frac{V}{c}\left(\bar{\theta} - \theta_k\right), \tag{2}$$

where $\bar{\theta}$ is the mean detectability across types. Fraudsters tilt activity toward types with below-average detectability, with the strength of substitution governed by V/c. Aggregate detection effectiveness is $\mathcal{A} = \sum_k \pi_k\theta_k$. A technology shock shifts detectabilities from $\boldsymbol{\theta}^0$ to

$\boldsymbol{\theta}^1$, induces a new allocation $\boldsymbol{\pi}^1$. The change in detection performance decomposes as:

$$\Delta\mathcal{A} = \underbrace{\sum_k \pi_k^0 \left(\theta_k^1 - \theta_k^0\right)}_{\text{technology effect}} + \underbrace{\sum_k (\pi_k^1 - \pi_k^0)\, \theta_k^1}_{\text{composition effect}}. \tag{3}$$

Because fraudsters substitute toward harder-to-detect types, the composition effect is always negative. As a result, aggregate detection performance can deteriorate even when detection technology improves: this is the Whac-A-Mole mechanism.

Equation (3) yields two implications. First, if a general-purpose enabling shock, such as the release of large language models, lowers net detectability across fraud types ($\delta_k = \theta_k^0 - \theta_k^1 > 0$ for all k), both the technology effect and the composition effect are negative, so $\Delta\mathcal{A} \leq 0$. Second, a targeted anti-fraud shock, such as EMV chip cards or 3-D Secure, raises detectability for a single fraud type j while leaving all others unchanged. This generates a positive direct effect, but also induces substitution toward harder-to-detect alternatives.[5] The Whac-A-Mole mechanism arises when the composition effect dominates, $\Delta\mathcal{A} < 0$.

Both results rest on the assumption that fraudsters shift their methods in response to technological change, that is, that $\theta_k^1 \neq \theta_k^0$. The empirical analysis below provide direct evidence for this assumption.

II.B Data and Measurement

II.B.1 CrimeBB Underground Forums

Our primary micro-level data come from CrimeBB, a structured corpus of underground forum posts maintained by the Cambridge Cybercrime Centre (Pastrana et al., 2018).[6] CrimeBB contains full-text posts, timestamps, pseudonymous user identifiers, and board-level metadata from several of the largest English-language hacking and fraud forums. We extract posts from five forums: Hackforums, Nulled, Breached, BreachForums, and Cracked, spanning a range of specializations from general hacking discussion and tool sharing to dedicated marketplaces for stolen credentials, carding, and account compromise.

The empirical analysis exploits two complementary sources of variation, each with distinct advantages. The first is the public release of ChatGPT on November 30, 2022, which serves as a natural experiment for a general-purpose enabling shock. Its key advantage is

[5] The net change in detection effectiveness is: $\Delta\mathcal{A} = \pi_j^0(\theta_j^1 - \theta_j^0) - \frac{V}{c}(\theta_j^1 - \theta_j^0)^2\left(1 - \frac{1}{K}\right)$.

[6] CrimeBB is available to vetted academic researchers through the Cambridge Cybercrime Centre (https://www.cambridgecybercrime.uk). A growing body of computational work has applied natural language processing to underground forum data to measure cybercriminal markets (Portnoff et al., 2017; Bhalerao et al., 2019).

precision: unlike most technological innovations, which diffuse gradually and lack a clearly defined adoption date, ChatGPT represents a single, sharp event that triggered an immediate and widely documented shift in public awareness and tool availability. The second source of variation consists of 27 targeted anti-fraud technology deployments spanning 2013–2023 (see Internet Appendix C). These events complement the ChatGPT analysis in two respects. First, they affect specific fraud types rather than the broad landscape of online activity, allowing us to test whether fraudster adaptation is targeted rather than general. Second, because they are distributed across different dates, fraud categories, and deploying institutions, they are not confounded by the large simultaneous shift in public discourse that accompanied ChatGPT's release.

II.B.2 Sample Selection

ChatGPT analysis. We define the sample window around the release date of November 30, 2022. The pre-period covers November 2021 through November 2022 (12 months), and the post-period extends from December 2022 through September 2024 (22 months). The asymmetry reflects data availability: Hackforums, Nulled, and BreachForums snapshots extend to September 2024, while Cracked and Breached end in June 2023.[7] We restrict the sample to English-language, non-gibberish posts of at least five words.[8] The final sample contains approximately 1.19 million posts by 225,534 unique creators. Hackforums is the largest forum by post volume (359,962 posts), followed by Breached (278,222) and Nulled (263,255). Each forum is organized into topical sub-boards such as Carding, Hacking Tutorials, and Marketplace, which we use as fixed effects. The median creator contributes 1 post over 1.5 months on average; 3.2% of creators post on more than one forum.

Stacked event study. For each of the 27 anti-fraud technology deployments, we retain all creator-months in the window $[T_{0j} - 36,\, T_{0j} + 60]$ months around the deployment date T_{0j}. The full CrimeBB corpus of approximately 8 million posts is scored against all 443 dictionary terms simultaneously using the Aho–Corasick algorithm (Aho and Corasick, 1975), producing for each technology j a creator-month panel with binary indicators (vuln_{ijt}, ctr_{ijt}) capturing whether a creator posted about vulnerable methods or countermeasures in a given month.

[7] All regressions include forum-board and creator fixed effects, which absorb level differences arising from differential coverage.

[8] Some gibberish-classified posts may represent coded or obfuscated communication among experienced forum participants. Because our identification relies on dictionary-based keyword matching, such posts would not match any dependent variable terms regardless of inclusion, so their exclusion tightens estimation without removing informative observations. To the extent that sophisticated adaptation shifts toward coded communication channels, our estimates are a lower bound on total adaptation.

Each technology yields approximately 1.7 million creator-months across roughly 1 million unique creators.

Table 1
CrimeBB Sample Construction and Descriptive Statistics

Forum	Raw Posts	Clean Posts	% Kept	Unique Creators	Boards
Breached	737,922	278,222	37.7	47,896	71
Breachforums	332,891	166,796	50.1	28,147	72
Cracked	898,326	122,619	13.6	52,528	104
Hackforums	392,578	359,962	91.7	43,977	142
Nulled	1,385,758	263,255	19.0	60,357	82
Total	3,747,475	1,190,854	31.8	225,534	471

Notes: Posts from CrimeBB PostgreSQL snapshots (Cambridge Cybercrime Centre). Sample window: 12 months pre to 22 months post ChatGPT launch (Nov 2021–Sep 2024). Clean posts: English-language, non-gibberish, $\geq$5 words. Hackforums and Nulled snapshots extend to Sep 2024; Cracked and Breached snapshots end Jun 2023.

II.B.3 Dependent Variables

We construct five binary indicators at the post level that capture, with increasing specificity, the intersection of generative AI and fraud discussion. The measures are designed to distinguish genuine AI–fraud integration from a general increase in AI awareness following ChatGPT's release.

The broadest measure, $\mathbf{1}$(GenAI), equals one if a post mentions any generative AI term, including ChatGPT, GPT, LLM, deepfake, or Stable Diffusion, and captures general AI awareness without requiring fraud-related content. $\mathbf{1}$(Fraud) equals one if a post mentions any fraud-related term such as phishing, carding, ransomware, or credential stuffing, and serves as a placebo: if ChatGPT merely increased AI discussion without affecting fraud behavior, this measure should be unaffected.

The three remaining measures capture the intersection of AI and fraud discussion with increasing precision. $\mathbf{1}$(GenAI $\times$ Fraud) equals one if a post contains both a GenAI and a fraud term. $\mathbf{1}$(Fraud Method) equals one if a post mentions a specific fraud method or tool, such as a phishing kit, fullz, CVV, or combo list, rather than fraud in general. $\mathbf{1}$(GenAI $\times$ Method) equals one if a post contains both a GenAI term and a specific fraud method term, and is the most restrictive measure. A finding in which only $\mathbf{1}$(GenAI) increases after ChatGPT's release would indicate that the shock raised AI awareness without penetrating into fraud operations; a finding in which the intersection measures also increase would indicate integration of AI into concrete fraud activity. The complete keyword lists are reported in Internet

Appendix D.

II.B.4 Technology Events and Dictionaries

We identify 27 anti-fraud technology deployments spanning 2013–2023 across nine fraud categories: carding, account takeover, banking, identity, payment systems, SIM and telecom, AI techniques, circumvention, and cryptocurrency fraud. Each deployment event T_0 is anchored to a precisely dated landmark, a product launch, regulatory milestone, or market-defining acquisition, to ensure the event date is not endogenous to forum activity. The full list with deployment rationales is in Internet Appendix C.

For each technology, we construct two curated keyword dictionaries. *Vulnerable-method terms* (178 unique terms) describe the attack vector targeted by the technology; *countermeasure terms* (270 unique terms) describe techniques for circumventing the deployed defense. The asymmetry reflects the greater diversity of circumvention strategies relative to the attack vectors they target. All dictionaries are in Internet Appendix D.

II.B.5 Technical Sophistication

To test whether the ChatGPT shock differentially affected creators of varying skill levels, we construct a pre-period technical sophistication score for each creator. The score counts pre-period posts containing at least one term from a curated dictionary of 212 technical terms spanning 13 categories: exploitation and vulnerability research, malware development, reverse engineering, network attacks, OPSEC, and others.[9] Because the score is computed entirely from pre-period posts, it is a fixed creator attribute immune to contamination by the ChatGPT shock.

The distribution is heavily right-skewed: 97.1% of pre-period creators have zero sophistication posts, reflecting the concentration of technical expertise among a small core of underground forum participants. Among the 4,332 creators with at least one sophistication post, the median is 1 and the 90th percentile is 3. We use both the continuous measure $\log(1 + \text{sophistication posts})$ and a binary indicator $\text{TechSoph}^{\text{hi}}$ in the interaction specifications.

II.B.6 Descriptive Statistics

Table 2 reports the prevalence of each dependent variable before and after ChatGPT's release. In the pre-period, only 0.11% of posts mention a GenAI term; post-ChatGPT, this

[9]The full dictionary is in Internet Appendix E. Categories covering deepfakes and AI-generated content are excluded to avoid overlap with the dependent variables.

rises to 0.79%, a seven-fold increase. More importantly, the intersection measures also surge: $\mathbf{1}(\text{GenAI} \times \text{Fraud})$ rises from 0.023% to 0.133%, a 5.8-fold increase, indicating that AI adoption on these forums penetrates into fraud-specific discussion. Meanwhile, $\mathbf{1}(\text{Fraud})$ remains roughly stable at approximately 6.3%, suggesting that the overall volume of fraud discussion did not change but its technological composition did. At the creator level, 1.1% of creators ever mention a GenAI term, while 9.9% mention a fraud term; 582 creators (0.3%) account for all posts in which AI and fraud co-occur.

Table 2
Fraud Discussion Before and After ChatGPT

	Pre-ChatGPT		Post-ChatGPT	
Variable	Count	%	Count	%
$\mathbf{1}$(GenAI)	766	0.11	4,050	0.79
$\mathbf{1}$(Fraud)	42,943	6.34	32,080	6.25
$\mathbf{1}$(GenAI × Fraud)	157	0.02	685	0.13
$\mathbf{1}$(Tech Sophistication)	9,167	1.35	9,270	1.81
Posts	677,451		513,403	
Unique creators	148,122		95,453	

Notes: Pre-ChatGPT: Nov 2021–Nov 2022. Post-ChatGPT: Dec 2022 onward. Each indicator equals 1 if the post contains at least one matching term from the respective dictionary. GenAI × Fraud requires co-occurrence of both AI and fraud terms in the same post.

II.C The ChatGPT Natural Experiment

II.C.1 Empirical Strategy

The release of ChatGPT on November 30, 2022 provides a single, precisely dated shock to the availability of generative AI tools. ChatGPT reached over 100 million users within weeks of its release, making sophisticated language generation, code writing, and content creation accessible to anyone with an internet connection, including participants in underground fraud forums. From the perspective of Section II.A, this is a general-purpose enabling shock that lowers the skill barrier for multiple fraud types simultaneously.

We estimate the effect of this shock on the five dependent variables using within-creator variation. The main specification is a linear probability model with creator fixed effects:

$$y_{it} = \alpha_i + \beta \cdot \text{After}_t + \gamma \cdot X_{it} + \varepsilon_{it}, \tag{4}$$

where $y_{it} \in \{0, 1\}$ is one of the five dependent variables for creator i in post t, α_i are creator fixed effects, $\text{After}_t = \mathbf{1}[t \geq \text{Dec 2022}]$, and X_{it} includes an indicator for thread-initiating

posts and the log of the creator's pre-period posting count. The dependent variable is standardized to zero mean and unit standard deviation, so $\hat{\beta}$ is interpretable in standard deviation units; multiplying by SD(y) recovers the raw probability change. Standard errors are clustered at the creator level. Creator fixed effects ensure identification comes from within-creator changes in posting behavior, not from compositional shifts in who is active on the forums.[10]

The main threat to identification is confounding events coinciding with the ChatGPT release. We address this in three ways: event-study plots documenting flat pre-trends, creator fixed effects absorbing all time-invariant heterogeneity, and interaction specifications testing whether effects vary with pre-existing creator characteristics as predicted by the skill-barrier-lowering mechanism of Section II.A.

II.C.2 Main Results

Table 3 reports the main results. The coefficient on **1**(GenAI) is large and highly significant, while the coefficients on **1**(Fraud) and **1**(Method) are small and statistically indistinguishable from zero. Crucially, both intersection measures are positive and significant. Column (1) shows a broad increase in AI-related discussion around the ChatGPT release, consistent with the shift in public discourse that accompanied the launch. However, ChatGPT did not increase general fraud discussion, but it did increase discussion that combines AI with fraud contexts and with specific fraud methods. The effect sizes are economically meaningful: AI–fraud co-discussion nearly doubles relative to its pre-period baseline, while overall fraud discussion is unchanged. Wald tests (Internet Appendix Table B2) confirm that the AI–fraud intersection grew significantly faster than both general AI discussion and overall fraud discussion, indicating that AI tools are being integrated into concrete fraud activity rather than merely generating idle curiosity.

[10]The identifying assumption is that, conditional on creator fixed effects, ChatGPT is the dominant shock to AI-related fraud discussion over this period. Earlier models such as GPT-3 (June 2020) and DALL-E 2 (April 2022) attracted attention but achieved nothing comparable in terms of mass adoption. Google Bard and Meta's LLaMA both post-date our treatment cutoff.

Table 3
The Effect of ChatGPT on Underground Forum Discussion

	1(AI)	**1**(Fraud)	**1**(AI×Fr)	**1**(Method)	**1**(AI×M)
after_chatgpt	0.0960 * **	−0.0096	0.0254 * **	−0.0047	0.0138 * **
	(0.0088)	(0.0122)	(0.0049)	(0.0065)	(0.0044)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Creator	Creator	Creator	Creator	Creator
N	1,190,854	1,190,854	1,190,854	1,190,854	1,190,854
R^2	0.2043	0.5019	0.1897	0.3288	0.1896
Mean(y)	0.0040	0.0631	0.0007	0.0196	0.0003
SD(y)	0.0635	0.2431	0.0266	0.1387	0.0173

Notes: Sample: All boards. Y standardised to zero mean and unit standard deviation; X in natural units. $\hat{\beta}$ = effect of 0→1 change in SD(y) units. Creator FE + controls (initial post indicator, log reputation). Clustered standard errors (creator) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

II.C.3 Heterogeneity: Experience and Technical Sophistication

To test the skill-barrier-lowering mechanism, we interact After_t with two pre-determined creator characteristics. Table 4 reports the experience interaction. More experienced creators adopted AI terminology faster, but less experienced creators drove the increase in fraud method discussion. This pattern is consistent with a diffusion dynamic in which technically literate creators were early adopters, while ChatGPT provided the greatest marginal benefit to creators who previously lacked the expertise to generate convincing content or code.

Table 4
Early Adopters: The Role of Forum Experience

	1(AI)	**1**(Fraud)	**1**(AI×Fr)	**1**(Method)	**1**(AI×M)
after_chatgpt	0.0634 * **	0.0207	0.0186 * **	0.0279 * *	0.0202 * **
	(0.0105)	(0.0137)	(0.0071)	(0.0129)	(0.0074)
after × log_rep	0.0067 * **	−0.0063	0.0014	−0.0068 * **	−0.0013
	(0.0024)	(0.0043)	(0.0013)	(0.0024)	(0.0015)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Creator	Creator	Creator	Creator	Creator
N	1,190,854	1,190,854	1,190,854	1,190,854	1,190,854
R^2	0.2043	0.5019	0.1897	0.3288	0.1896
Mean(y)	0.0040	0.0631	0.0007	0.0196	0.0003
SD(y)	0.0635	0.2431	0.0266	0.1387	0.0173

Notes: Sample: All boards. after × log_rep = after_chatgpt × log(1+pre-period posts). Y standardised to zero mean and unit standard deviation; X in natural units. Creator FE + controls (initial post indicator, log reputation). Clustered standard errors (creator) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5 reports the interaction with pre-period technical sophistication. Technically

sophisticated creators drove both AI adoption and AI–fraud co-discussion, while less sophisticated creators drove the increase in pure fraud and method discussion. Together, the two interaction tests characterize the distributional incidence of the shock: sophisticated creators were first movers in integrating AI into fraud applications, while the broadening of fraud method discussion was concentrated among less skilled participants for whom ChatGPT lowered entry barriers.

Table 5
Early Adopters: The Role of Technical Sophistication

	$\mathbf{1}$(AI)	$\mathbf{1}$(Fraud)	$\mathbf{1}$(AI×Fr)	$\mathbf{1}$(Method)	$\mathbf{1}$(AI×M)
after_chatgpt	0.0767 * **	−0.0002	0.0163 * **	0.0099	0.0112 * **
	(0.0074)	(0.0112)	(0.0047)	(0.0067)	(0.0041)
after × log_tech_soph	0.0164 * **	−0.0080*	0.0077 * *	−0.0124*	0.0022
	(0.0060)	(0.0045)	(0.0038)	(0.0065)	(0.0046)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Creator	Creator	Creator	Creator	Creator
N	1,190,854	1,190,854	1,190,854	1,190,854	1,190,854
R^2	0.2044	0.5019	0.1897	0.3289	0.1896
Mean(y)	0.0040	0.0631	0.0007	0.0196	0.0003
SD(y)	0.0635	0.2431	0.0266	0.1387	0.0173

Notes: Sample: All boards. after × log_tech_soph = after_chatgpt × log(1 + pre-period tech-sophistication posts). Y standardised to zero mean and unit standard deviation; X in natural units. Creator FE + controls. Clustered standard errors (creator) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

II.C.4 Event Study

Figure 2 plots event-study coefficients from a variant of equation (4) that replaces After_t with month-by-month dummies relative to December 2022. Pre-period coefficients are close to zero and statistically indistinguishable from the reference period, indicating no pre-existing trend in fraud- and AI-related discussion prior to the ChatGPT release. Post-release coefficients show a sharp break at the treatment date, rising to a peak around months +4 to +7 before partially moderating, yet remaining significantly above zero through the end of the sample. The results are consistent with rapid initial adoption followed by sustained diffusion through the underground community.

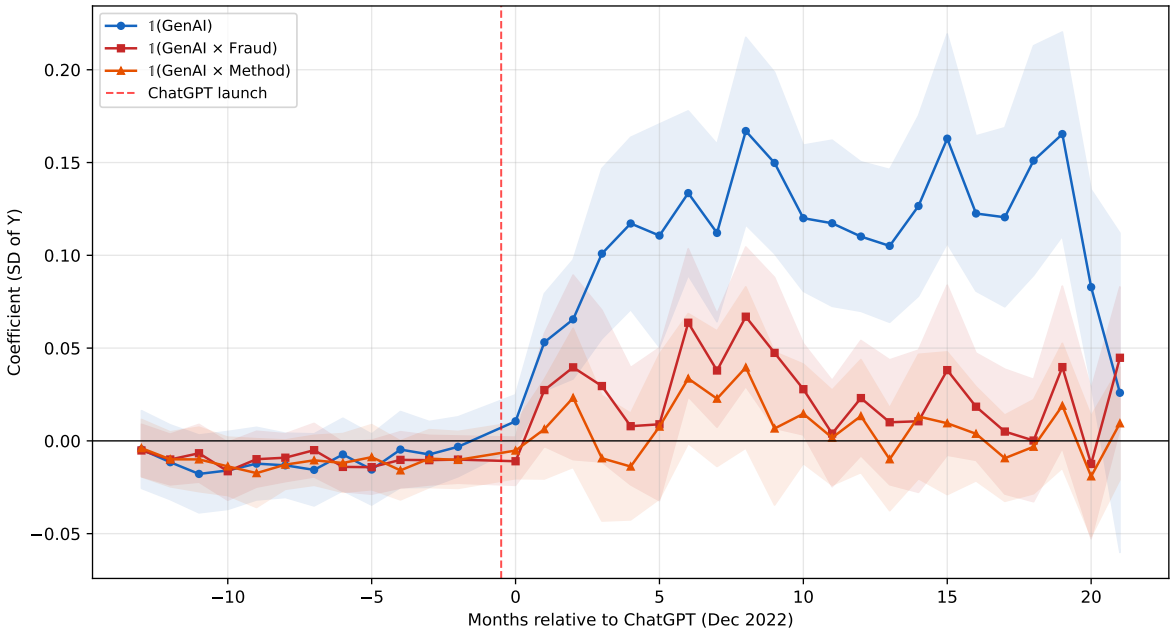


Figure 2
AI–Fraud Discussion Around the ChatGPT Release

Notes: Event-study coefficients from equation (4) with month-by-month dummies relative to December 2022 ($m = -1$ as reference). The dependent variable is standardized to zero mean and unit standard deviation. Shaded bands indicate 95% confidence intervals based on creator-clustered standard errors.

II.D Adversarial Adaptation to Anti-Fraud Technologies

II.D.1 Empirical Strategy

We complement the ChatGPT analysis with a stacked event-study design that examines fraudster adaptation to 27 targeted anti-fraud technology deployments spanning 2013–2023. While the ChatGPT analysis identifies the effect of a general-purpose enabling shock, the stacked design tests the targeted-shock mechanism of Section II.A: whether anti-fraud technology deployments trigger the accumulation of countermeasure knowledge without suppressing discussion of the targeted attack methods.

For each technology j with deployment date T_{0j}, we estimate:

$$y_{it} = \alpha_i + \gamma_t + \sum_{k \neq \text{ref}} \beta_k \cdot \mathbf{1}[\text{bin}(t - T_{0j}) = k] + \varepsilon_{it}, \tag{5}$$

where $y_{it} \in \{\text{vuln}_{it}, \text{ctr}_{it}\}$, α_i are creator fixed effects, γ_t are calendar-month fixed effects, and the β_k are event-time coefficients relative to the reference window $[-3, -1]$ months before

T_0. Event-time bins are: < -24, $[-24, -13]$, $[-12, -7]$, $[-6, -4]$, $[-3, -1]$ (reference), $[0, 2]$, $[3, 5]$, $[6, 11]$, $[12, 23]$, $[24, 35]$, $[36, 47]$, $48+$. Identification relies on within-creator, within-calendar-month variation; pre-deployment coefficients serve as a falsification test. Standard errors are heteroskedasticity-robust (HC1). Per-technology estimates are pooled via inverse-variance weighted meta-analysis, which mitigates the concern that any single deployment coincides with a confounding event, since the 27 technologies span different dates, fraud categories, and deploying institutions.

II.D.2 Results: The Asymmetry of Adaptation

Figure 3 presents the pooled event-study estimates. Pre-deployment coefficients are flat for both outcomes, suggesting relatively little confounding before each technology deployment. The right panel shows a gradual but significant increase in countermeasure discussion following technology deployment, rising monotonically after T_0 and reaching statistical significance at 24–35 months post-deployment. The left panel tells a starkly different story: technology deployment does not suppress discussion of vulnerable methods. Forum participants continue discussing targeted attack methods at the same rate after defenses are deployed.

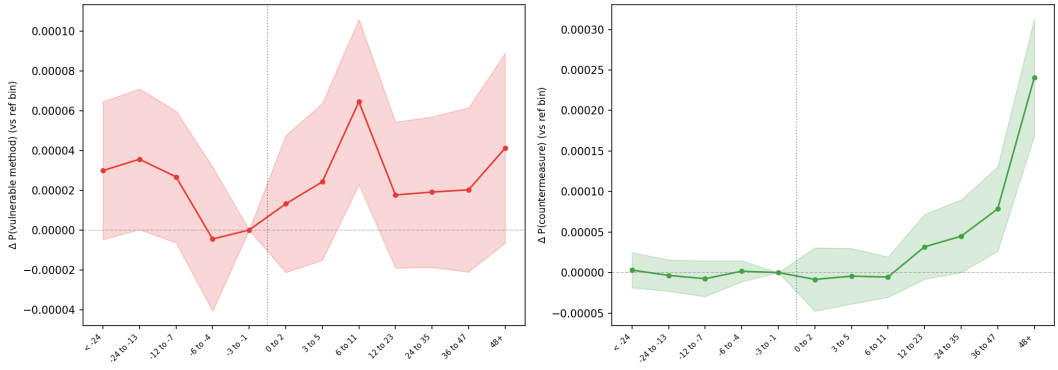


Figure 3
Asymmetric Adaptation After Anti-Fraud Technology Deployment

Notes: Inverse-variance weighted meta-analytic event-study coefficients pooled across 27 anti-fraud technology deployments. The left panel shows the probability of vulnerable-method discussion; the right panel shows the probability of countermeasure discussion. Event-time bins are relative to each technology's deployment date T_0, with $[-3, -1]$ months as the reference. Creator and calendar-month fixed effects included. Shaded bands indicate 95% confidence intervals based on HC1 standard errors.

This asymmetry is the empirical signature of the Whac-A-Mole mechanism. Technology deployment adds countermeasure knowledge to the underground community without removing existing attack method knowledge: fraudsters learn to circumvent new defenses without abandoning prior techniques.

II.D.3 Heterogeneity Across Technologies

The pooled estimates mask substantial variation across the 27 technologies. We characterize this variation using two per-technology outcome measures: the long-run event-study coefficient at 48+ months and the convergence ratio, the 12–23 month coefficient divided by the 48+ month coefficient. Figure 4 presents two views of this heterogeneity. The left panel plots convergence speed against long-run magnitude: cryptocurrency-related technologies dominate the extremes of the magnitude axis. The right panel reveals a clearer pattern: older technologies show higher convergence ratios, indicating that circumvention knowledge for earlier deployments is closer to saturation. The Spearman rank correlation between convergence speed and technology age is $\rho = +0.59$, suggesting that model decay unfolds over years rather than months, with adaptation accelerating as the community accumulates experience with a given technology.

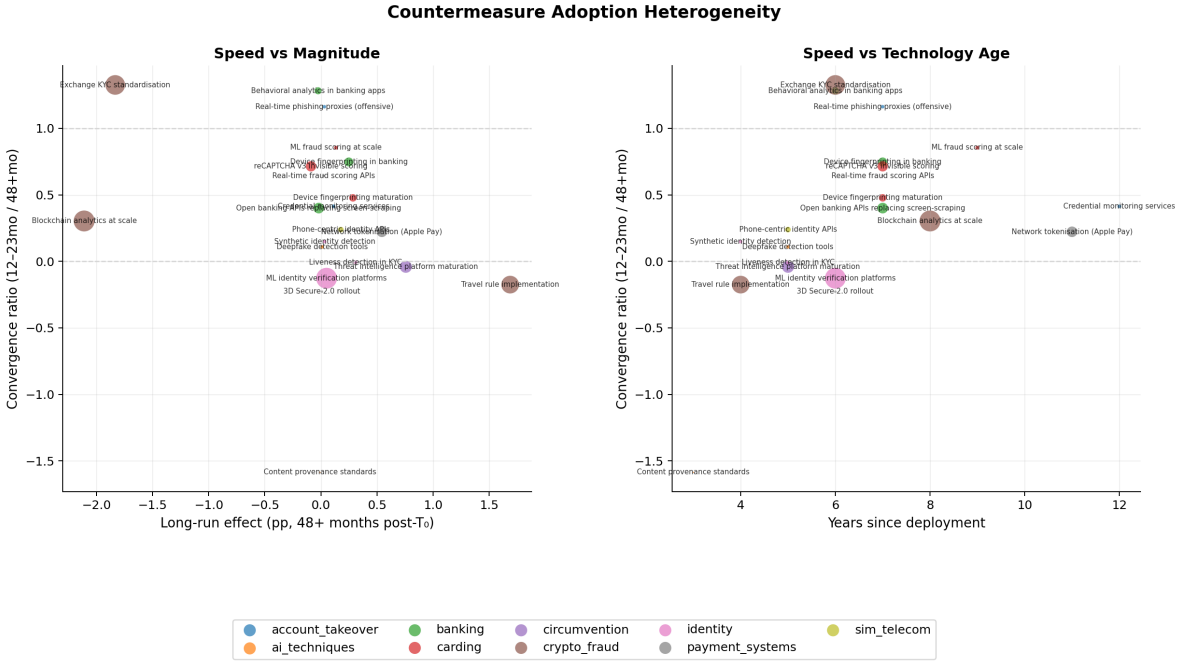


Figure 4
Adaptation Speed and Magnitude Across Technologies

Notes: The left panel plots convergence speed (12–23 month coefficient divided by the 48+ month coefficient) against long-run magnitude (48+ month event-study coefficient) for each of the 27 technologies. The right panel plots convergence speed against years since technology deployment. Spearman rank correlation: $\rho = +0.59$.

II.E Corroborating Evidence

The framework of Section II.A predicts two aggregate implications of fraudster adaptation: the composition of fraud types shifts over time as fraudsters substitute toward harder-to-detect methods, and detection model performance decays as the attack distribution diverges from the training data. FBI IC3 and FTC Sentinel complaint data show that the composition of fraud types shifts markedly over time: phishing and impersonation fraud grow steadily by complaint count, while investment fraud dominates financial losses and rises sharply after 2020. A prominent example of composition-driven substitution is EMV chip card adoption: Hayashi et al. (2015) documents that card-not-present fraud rose sharply in the United Kingdom, France, and Canada immediately following EMV deployment, as fraudsters substituted toward online channels where chip authentication offered no protection. Internet Appendix F.1 provides corroborating evidence using approximately 500,000 Reddit submissions from /r/Scams over 2012–2024, documenting substitution across scam types including the displacement of Bitcoin scams following Twitter's 2018 crackdown on cryptocurrency fraud accounts. Internet Appendix F.2 documents detection model decay using Twitter bot-detection data: classifiers trained on early-generation bots exhibit sharp declines in accuracy as more sophisticated social bots become prevalent, with performance declining monotonically as the composition shifts toward harder-to-detect types.

The empirical evidence documents a problem, adaptation erodes detection, but it cannot, on its own, prescribe a solution. If the data showed only that PSPs underinvest in detection, the policy response would be straightforward: shift liability to PSPs and let the market supply the missing effort. But the evidence also shows that adaptation is endogenous to detection itself, that it accumulates over time, and that its severity varies with market conditions. These facts raise a harder question: how should liability rules be designed when the very act of investing in detection accelerates the process that undermines it? Answering this question requires understanding how liability, competition, and adaptation interact in equilibrium, objects the data alone cannot identify. The model we develop below translates the empirical regularities into a framework that delivers these equilibrium objects, and in doing so yields results that are not obvious from the data. In particular, the model shows that full liability, the natural first response to underinvestment, overshoots: it induces PSPs to invest beyond the social optimum, because each provider ignores how its detection effort contributes to market-wide adaptation. The optimal rule is partial liability, and its level depends on market structure and the severity of decay. The next section develops this argument in detail.

III The Model

We embed the adaptation mechanism documented in Section II in a model of payment fraud detection that is a special case of the framework in Section II.A, restricted to two fraud types and with detection intensity chosen endogenously by payment service providers (PSPs). A regulator sets a liability parameter that determines the share of fraud losses each PSP must reimburse to consumers. This liability rule, together with competition among intermediaries and the option to retrain detection models, interacts with adversarial adaptation to determine welfare outcomes.

The base model has two types of agents: one fraudster and payment service providers (PSPs) in a two-period economy. All parties are risk neutral. We assume a zero discount rate throughout the analysis.[11]

Payment service providers are homogeneous financial intermediaries that may represent either traditional banks or FinTech companies. For simplicity, we do not distinguish between these two types of institutions in the analysis. PSPs provide payment services over two periods ($t = 0, 1$) and invest in detection of transactions that may be fraudulent. The PSP processes M transactions in each period, which we normalize to one for simplicity: $M = 1$. A share $\alpha \in [0, 1]$ of transactions is exposed to fraud attempts in each period. There are two types of fraud: Type A and Type B. PSPs can invest in detection technology to reduce the likelihood that fraud attempts succeed. The detection technology is more effective against Type A fraud. Specifically, the success probability of a Type A fraud attempt in the first period is given by $\phi_{0i}^{A} = 1 - I_{0i}$, where $I_{0i} \in [0, 1]$ denotes the detection intensity of PSP i at time $t = 0$. Type B fraud is more difficult to detect. Its success probability is given by $\phi_{0i}^{B} = 1 - \theta I_{0i}$, where $\theta \in [0, 1]$ captures the relative detectability of Type B fraud. When $\theta = 1$, both fraud types are equally detectable. When $\theta = 0$, Type B fraud is entirely undetectable by the detection technology. The investment I_{0i} imposes a cost on PSPs, which, following Hauswald and Marquez (2006) and Carletti et al. (2023), we model as quadratic: $\frac{1}{2} I_{0i}^2$.

A share α of transactions are targeted by fraudsters, and we assume targeted consumers authorize the payment, consistent with the APP fraud setting in which victims are deceived into initiating transfers voluntarily. Only PSP detection can prevent a targeted transaction from resulting in fraud.[12]

[11] Risk neutrality and zero discounting are without loss of generality for the qualitative results. Risk-averse PSPs would invest more aggressively in detection to reduce variance in fraud losses, strengthening the overinvestment result under full liability and reinforcing the case for partial liability. A positive discount rate attenuates but does not eliminate the model decay externality, since the externality operates through period-2 losses whose weight falls with discounting.

[12] We relax this assumption in Internet Appendix H, where consumers choose a vigilance level that reduces

We assume the share of Type A fraud in the first period is one ($p_0 = 1$). All fraud attempts that are flagged by detection are blocked by the PSP; attempts that evade detection result in actual fraud cases. PSPs use the same fraud detection model in both periods (we relax this assumption in Section V).

In period 1, the fraudster employs Type A fraud. At the end of the period, after observing the average detection intensity $\bar{I}_0 = \frac{1}{N}\sum_{i=1}^{N} I_{i0}$, the fraudster may reduce the share of Type A fraud in period 2 to $p_1 \in [0, 1]$ and shift the remaining share $(1 - p_1)$ to an undetectable Type B fraud. This shift entails a convex cost of $\frac{1}{2}(1 - p_1)^2$. In each successful fraudulent transaction, the fraudster earns a payoff of $V > 0$.

Figure 5 illustrates the sequence of events. At time $t = 0$, PSPs choose a fixed detection intensity I_0, which governs the sensitivity of their fraud detection systems. In each period ($t = 1, 2$), PSPs detect fraud, but any undetected fraudulent transactions result in realized fraud. After period 1, the fraudster observes the average detection intensity across PSPs and may adjust its strategy. Specifically, the fraudster can reduce the share of detectable (Type A) fraud and shift to an undetectable (Type B) variant, incurring a convex cost of adaptation. This strategic adjustment reflects model decay from the perspective of the PSPs. The interaction between PSPs and the fraudster can be interpreted as a Stackelberg Security Game (SSG) in which PSPs act as defenders (leaders) that commit to detection intensities before observing fraud behavior, and the fraudster acts as a follower that best responds by adapting its attack type. Our model extends the standard SSG framework (see Korzhyk et al. (2011); Tambe (2011)) by introducing multiple defenders, continuous detection intensities rather than discrete defense allocations, and an endogenous "model decay" externality through the fraudster's adaptive response.[13]

Key variables of the model, including those introduced here and in subsequent sections, are tabulated in Table 6. Proofs of all propositions are in Appendix J.

their exposure to fraud.

[13] An analogous mechanism operates in epidemiology, where pathogen populations evolve resistance under selective pressure from drugs or vaccines, degrading treatment efficacy over time (Laxminarayan and Brown, 2001). In both settings, adaption endogenously shifts the distribution of attacks toward variants that are harder to counter.

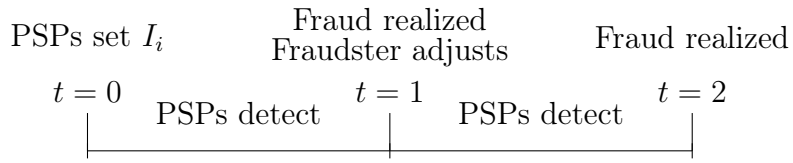


Figure 5
Timeline of Events

Notes: At $t = 0$, PSPs choose detection intensity I_i. In each period ($t = 1, 2$), PSPs detect fraud; undetected attempts result in realized fraud. After period 1, the fraudster observes average detection intensity and may shift from Type A to Type B fraud at a convex cost.

Table 6
Baseline Parameters for Numerical Analysis

Parameter	Description	Range	Baseline Value
α	Share of transactions exposed to fraud	$[0, 1]$	0.3
V	Loss from a successful fraud attempt	$\mathbb{R}_+$	3
N	Number of competing PSPs	$\mathbb{N}_+$	1, 2, 5
θ	Relative detectability of Type B fraud	$[0, 1]$	0.3, 0.5
γ	PSP liability share (regulatory choice)	$[0, 1]$	Endogenous
p_t	Share of Type A fraud in period t	$[0, 1]$	Endogenous
I_{0i}	Detection intensity of PSP i at $t = 0$	$[0, 1]$	Endogenous
ϕ_{it}	Fraud success rate for PSP i in period t	$[0, 1]$	Endogenous

Notes: Baseline parameter values used in the numerical analysis. "Endogenous" indicates that the variable is determined in equilibrium.

III.A Fraud Type Choice

In this section, we characterize the fraudster's strategic response to PSP detection intensity and derive the model decay function used in the remainder of the paper. The fraudster observes the average detection intensity across PSPs and adjusts the mix of fraud types accordingly. A higher intensity induces a shift from detectable (Type A) to harder-to-detect (Type B) fraud, at a convex cost. This endogenous adjustment introduces model decay and generates a feedback loop that weakens the effectiveness of fraud detection over time.

We begin by analyzing the fraudster's choice of fraud type at $t = 1$. The fraudster determines at $t = 1$ the optimal share of type A fraud p_1, given the average fraud success rate in period $t = 0$ to maximize its profits:

$$\Pi_{fraudster} = (p_1(1 - \bar{I}_0) + (1 - p_1)(1 - \theta\bar{I}_0))\alpha V - (1/2)(1 - p_1)^2, \tag{6}$$

where $\bar{I}_0 = (1/N)\sum_{i=1}^{N} I_{i0}$ is the average detection intensity of the N PSPs in period $t = 0$

and V represents the fraud benefits. Taking the first order conditions with respect to p_1 yields an optimal share of type A fraud in period $t = 1$ of:

$$p_1^* = 1 - (1 - \theta)\alpha V \bar{I}_0 \tag{7}$$

The fraud success rate in period 1 ϕ_{i1} is the weighted average of the fraud success rate of type A and B:

$$\phi_{i1} = p_1(1 - I_{i0}) + (1 - p_1)(1 - \theta I_{i0}) = 1 - I_{i0} + (1 - \theta)^2 \alpha V \bar{I}_0 I_{i0}, \tag{8}$$

The fraudster's optimal response in equation (7) reveals a key externality: higher aggregate detection intensity $\bar{I}_0$ reduces detectable Type A fraud but *endogenously* increases undetectable Type B fraud if $\theta < 1$. The fraudster's equilibrium choice captures precisely the asymmetric adaptation documented in Section II.D: higher aggregate detection intensity induces a shift toward Type B fraud without eliminating Type A, consistent with the monotonic growth of countermeasure discussion that does not displace vulnerable-method discussion.

The model decay described above is closely related to the phenomenon of concept drift studied in the computer science and machine learning literature. Concept drift refers to the evolution of the underlying data-generating process in ways that invalidate a model's predictive accuracy (Bayram et al., 2022; Bolton and Hand, 2002). Concept drift manifests through the growing prevalence of more sophisticated and less detectable fraud types, mirroring the increase in type B fraud in our model, which degrades model performance if left unaddressed. A related phenomenon is documented empirically by Rajan et al. (2015), who show that default models in mortgage markets systematically underpredict risk when lenders adjust loan characteristics in response to securitization incentives. In both settings, agent behavior endogenously shifts the relationship between observed inputs and outcomes, leading to model decay. Our empirical evidence from CrimeBB (Section II.D) provides direct micro-level evidence for this mechanism: the monotonic growth of countermeasure discussion after technology deployment traces out the shift from Type A to Type B in real time.

In the next section, we examine how the foresight of model decay, that is, the PSPs' awareness that fraud models lose effectiveness over time, affects their detection intensities across different market structures.

III.B PSP Losses and Detection Intensities

We solve the model in a competitive setting with N PSPs that simultaneously choose detection intensities to minimize fraud losses, taking into account the strategic interaction through

aggregate model decay.[14]

Our formulation follows the attack-defense literature in modeling the PSP's objective as minimizing expected losses rather than maximizing profits (Goyal and Vigier, 2014; Maheshri and Mastrobuoni, 2021)[15]. In the payment system context, PSPs compete for transaction volume and set fees to cover costs. Under Bertrand competition with homogeneous consumers, fees are pinned down by a zero-profit condition and the PSP's problem reduces to minimizing the sum of fraud losses and detection costs. Modeling the objective as profit maximization with an endogenous fee yields identical equilibrium detection intensities; the loss-minimization formulation is therefore without loss of generality and keeps the exposition focused on the fraud detection decision.

We assume that the transaction volume $M = 1$ is evenly split across PSPs, so each PSP processes $1/N$ transactions per period. The loss function of PSP i is:

$$L_i = \frac{1}{N}\left(\phi_{i,0} + \phi_{i,1}\right)\gamma\alpha V + \frac{1}{2}I_{i,0}^2, \tag{9}$$

where $\phi_{i,0} = 1 - I_{i,0}$ and $\phi_{i,1} = 1 - I_{i,0} + (1-\theta)^2\alpha V \bar{I}_0 I_{i,0}$, and $\bar{I}_0 = \frac{1}{N}\sum_{i=1}^{N} I_{i,0}$. The first term is the expected fraud cost borne by PSP i over both periods, where $\gamma \in [0,1]$ is the liability share. In this section we treat γ as exogenous; the next section derives the optimal γ set by a regulator. The second term is the quadratic cost of detection investment at $t = 0$.

Proposition 1. *In a symmetric equilibrium with N identical PSPs, the equilibrium detection intensity is:*

$$I_N^* = \frac{2\gamma\alpha V}{N + \gamma\alpha^2 V^2(1-\theta)^2\left(1 + \frac{1}{N}\right)}.$$

The equilibrium detection intensity I_N^* is increasing in N if and only if $N < \gamma^{1/2}\alpha V(1-\theta)$.

The number of PSPs affects equilibrium detection intensity through two opposing mechanisms. First, as N increases, each PSP's share of total fraud losses declines, reducing the marginal benefit of detection investment: the market share effect tends to lower investment. Second, each PSP ignores the effect of its own detection intensity on aggregate model decay, which imposes a negative externality on rivals: the model decay externality tends to raise investment. When N is below the threshold $\gamma^{1/2}\alpha V(1-\theta)$, the externality effect dominates and I_N^* rises with N. Above this threshold, the market share effect prevails and detection intensity falls monotonically with market fragmentation.

[14]Internet Appendix G presents the case with a monopolistic PSP.

[15]This is the natural objective in security settings where the primary decision concerns how much to invest in deterrence, not how to price a service.

Detection intensity is also increasing in θ: when Type B fraud is more detectable, fraudsters have weaker incentives to switch types, which mitigates second-period model decay and strengthens ex ante detection incentives.

This overinvestment result differs from Hauswald and Marquez (2006), who show that banks overinvest in screening technology to steal business from competitors. In their setting, overinvestment is driven by a business-stealing externality; in ours, it is driven by a model decay externality. PSPs overinvest not to attract business, but because they do not internalize how stronger detection intensities accelerate fraudster adaptation market-wide.

IV Welfare and Optimal Liability

This section derives the socially optimal detection level and characterizes the liability rule that decentralizes it. The social planner minimizes total fraud costs plus detection costs, taking into account the market-wide effect of detection investment on fraudster adaptation.

IV.A Socially Optimal Detection Investment

The social planner chooses a common detection intensity I_0 to minimize the sum of expected fraud costs across both periods and the total cost of detection investment:

$$\min_{I_0} W = \left[\frac{1}{N}\sum_{i=1}^{N}\phi_{i,0} + \frac{1}{N}\sum_{i=1}^{N}\phi_{i,1}\right]\alpha V + \sum_{i=1}^{N}\frac{1}{2}I_{i,0}^2, \tag{10}$$

where $\phi_{i,0} = 1 - I_{i,0}$, $\phi_{i,1} = 1 - I_{i,0} + (1-\theta)^2\alpha V\bar{I}_0 I_{i,0}$, and $\bar{I}_0 = \frac{1}{N}\sum_{i=1}^{N} I_{i,0}$. The first term is total expected fraud losses across both periods; the second is the total cost of detection investment. Taking first-order conditions with respect to I_0 yields the socially optimal detection intensity:

$$I_W^* = \frac{2\alpha V}{N + 2(1-\theta)^2\alpha^2 V^2}. \tag{11}$$

Proposition 2. *The socially optimal detection intensity I_W^* satisfies:*

1. *$I_W^* > I_N^*$ when $\theta = 1$, $\gamma < 1$, and $N > 1$.*
2. *$I_W^* < I_N^*$ when $\gamma = 1$ and $N > 1$.*

The first inequality reflects a moral hazard problem: when $\gamma < 1$, PSPs do not fully internalize fraud losses and underinvest relative to the social optimum. The second inequality reflects the model decay externality: under full liability each PSP bears its own fraud costs but ignores the effect of its detection intensity on the fraud distribution faced by rival PSPs.

Competition therefore induces overinvestment relative to what a social planner would choose. We assume that the regulator could not directly set optimal investment level I_W^*, because this requires observing detection intensities that are typically private information. We therefore consider an implementable alternative: setting the liability share γ to align private incentives with the social optimum.

IV.B Optimal Liability Rule

The allocation of liability has become a central instrument of payment system regulation. Recent initiatives in the UK and EU have raised PSP liability for authorized push payment fraud, aiming to strengthen detection incentives. Formally, the regulator chooses $\gamma \in [0, 1]$ to minimize total fraud costs, internalizing the effect of liability on equilibrium detection investment.

Proposition 3. *The welfare-maximizing liability share is:*

$$\gamma^* = \frac{N^2}{N^2 + (1-\theta)^2\alpha^2V^2(N-1)} \leq 1.$$

For $N > 1$, γ^* is increasing in N for $N > 2$:

$$\frac{\partial \gamma^*}{\partial N} > 0.$$

When model decay is relevant ($\theta < 1$) and the market is competitive ($N > 1$), the regulator sets $\gamma^* < 1$. The intuition is straightforward: full liability induces overinvestment because PSPs do not internalize how their detection effort accelerates fraudster adaptation market-wide. The corrective factor $\gamma^* < 1$ reduces private detection incentives until the decentralized equilibrium aligns with the social optimum. As N increases, the model decay externality intensifies and the optimal liability share rises toward one: with many PSPs, each individual PSP's contribution to aggregate adaptation is small, so the overinvestment distortion weakens and less correction is needed.

Figure 6 illustrates the welfare function as a function of γ for alternative values of N and θ, confirming that the welfare-maximizing γ^* lies strictly below one whenever $\theta < 1$ and $N > 1$.

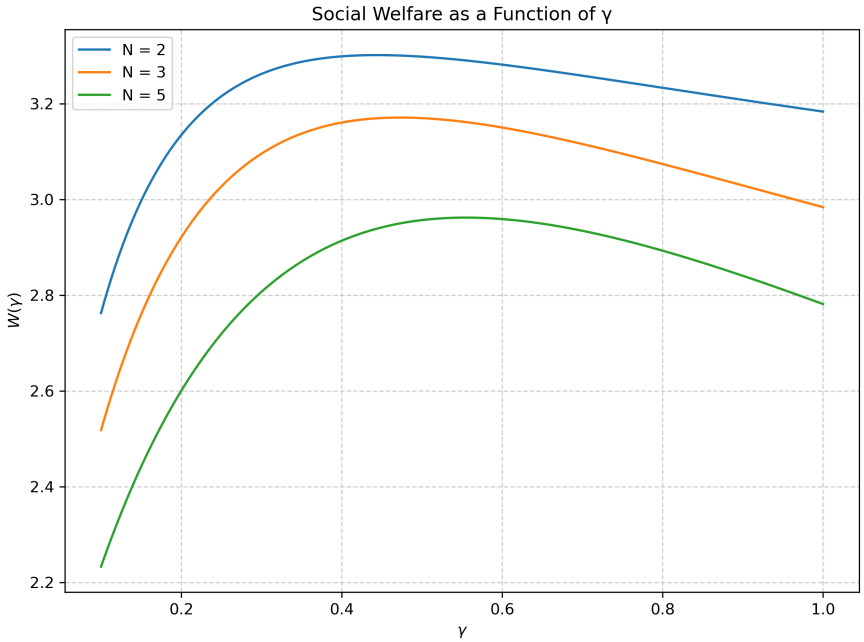


Figure 6
Optimal Liability is Partial

Notes: Total fraud and detection costs as a function of the PSP liability share γ for alternative values of N and θ. Baseline parameter values are reported in Table 6.

V Detection Reinvestment

So far we have assumed that PSPs deploy a detection technology at $t = 0$ and maintain it unchanged through both periods. In practice, PSPs can respond to model decay by reinvesting at $t = 1$: deploying a next-generation model trained on post-adaptation data, introducing new authentication protocols such as biometric verification or behavioral analytics, or upgrading transaction monitoring systems to detect emerging fraud patterns. All of these responses share the same economic structure: a second investment that restores detection effectiveness at a cost.

The convergence ratio evidence in Section II.D.3 motivates this extension directly. The finding that years since deployment is the strongest predictor of adaptation speed implies that detection performance deteriorates gradually but persistently, making reinvestment increasingly valuable for older, well-understood technologies. This accumulation of decay generates the ratchet effect we characterize below: reinvestment restores effectiveness but simultaneously accelerates fraudster adaptation, raising the future return to yet further investment.

We relax the assumption that PSPs cannot reinvest at $t = 1$. At $t = 1$, each PSP chooses between maintaining its existing detection technology, which yields a fraud success rate of

$\phi_{i1}^{\text{old}} = 1 - I_{i,0} + (1-\theta)^2 \alpha V \bar{I}_0 I_{i,0}$, or reinvesting at intensity I_{i1} at a quadratic cost $\frac{1}{2} I_{i1}^2$, which restores the fraud success rate to $\phi_{i1}^{\text{new}} = 1 - I_{i1}$.

The loss function of PSP i under reinvestment is:

$$L_i^{\text{new}} = \frac{1}{N} (2 - I_{i0} - I_{i1}) \gamma \alpha V + \frac{1}{2} \left(I_{i0}^2 + I_{i1}^2 \right), \tag{12}$$

where the first term captures expected fraud losses over both periods under the renewed detection technology, and the second term is the total cost of detection investment at $t = 0$ and reinvestment at $t = 1$.

Proposition 4 (PSP Reinvestment Rule)**.** *Define the reinvestment rule r_{PSP} as the reduction in losses from reinvesting at $t = 1$ relative to maintaining the existing technology:*

$$r_{PSP} = \frac{A}{N^2} - \frac{4A}{ND} + \frac{4\gamma A \alpha^2 V^2 (1-\theta)^2}{ND^2} + \frac{2A}{D^2}, \tag{13}$$

where

$$D = N + \gamma \alpha^2 V^2 (1-\theta)^2 \left(1 + \frac{1}{N} \right), \quad A = \gamma^2 \alpha^2 V^2.$$

A PSP prefers to reinvest if and only if $r_{PSP}(N) > 0$. Reinvestment is more likely when the number of PSPs N is small, the liability share γ is high, and model decay is more severe (low θ).

Reinvestment becomes optimal when the decay in detection effectiveness is sufficiently severe to justify the cost of a second investment round. The convergence ratio evidence in Section II.D.3 supports this directly: years since deployment is the strongest predictor of adaptation speed ($\rho = +0.59$), confirming that detection performance deteriorates persistently over time and that the return to reinvestment grows with technology age. Reinvestment, however, generates a ratchet effect. Each round of investment accelerates adaptation, making the next round necessary.

Proposition 5 (Reinvestment and the Social Optimum)**.** *Let $r_{PSP}(N)$ denote the individual PSP's reinvestment rule and $r_W(N)$ the corresponding social planner's reinvestment rule. Then:*

$$r_{PSP}(N) > r_W(N) \quad \text{if } N > 1 \text{ and } \gamma = 1.$$

PSPs reinvest more frequently than is socially optimal.

The intuition is analogous to the overinvestment result of Proposition 1. Under full liability, each PSP bears its own fraud losses but ignores how its reinvestment accelerates adaptation market-wide, imposing a negative externality on rival PSPs. The social planner

internalizes this externality and therefore requires a larger deterioration in detection performance before reinvestment is warranted. PSPs, by contrast, reinvest at lower levels of model decay than the social planner would choose. The optimal liability rule depends on whether PSPs reinvest in equilibrium. If $r_{\text{PSP}}(N) \leq 0$, PSPs do not reinvest and the optimal liability rule is γ^* from Proposition 3. If $r_{\text{PSP}}(N) > 0$, PSPs reinvest, and the reinvestment itself eliminates the model decay externality by restoring detection accuracy. In this case the regulator can set $\gamma = 1$: full liability provides correct incentives for both the initial investment and reinvestment, and the ratchet effect is the constrained-optimal outcome rather than a distortion to be corrected.

VI Detection and Market-Wide Fraud Reduction

Detection affects market outcomes through two channels. The first is the direct channel modeled above: higher detection intensity reduces the success rate of individual fraud attempts. The second is a market-wide deterrence channel: when detection is sufficiently intense, some fraudsters exit the market entirely, reducing the volume of fraud attempts in subsequent periods. We now incorporate this second channel.

In the baseline model, a single fraudster is active in both periods. We relax this assumption and consider an economy with F fraudsters. A fraction $(1 - \lambda \bar{I}_0)$ of fraudsters remain active into the second period, where $\lambda > 0$ governs the sensitivity of exit to aggregate detection intensity. Higher detection therefore reduces second-period fraud volume through two reinforcing mechanisms: lower success rates for active fraudsters and fewer active fraudsters. The loss function of PSP i becomes:

$$L_i = \frac{1}{N}\left[\phi_{i,0} + (1 - \lambda \bar{I}_0)\phi_{i,1}\right]\gamma \alpha V + \frac{1}{2} I_{i,0}^2, \tag{14}$$

where $\phi_{i,0} = 1 - I_{i,0}$ and $\phi_{i,1} = 1 - I_{i,0} + (1-\theta)^2 \alpha V \bar{I}_0 I_{i,0}$. The term $(1 - \lambda \bar{I}_0)$ captures the reduction in second-period fraud volume due to fraudster exit.

Proposition 6 (Optimal Liability Share with Fraudster Exit)**.** *Let γ^* denote the welfare-maximizing liability share. Then:*

1. *γ^* is strictly increasing in θ: when Type B fraud is more detectable, model decay weakens and the overinvestment distortion diminishes, so the regulator requires a higher liability share to maintain detection incentives.*

2. *γ^* is increasing in λ: a stronger fraudster exit response amplifies the public good problem, requiring a higher liability share to correct underinvestment.*

The optimal liability rule balances two opposing forces. A stronger fraudster exit response (λ) favors a higher liability share to counteract free-riding. More severe model decay (low θ) favors a lower liability share to restrain overinvestment. The welfare-maximizing γ^* trades off these competing distortions, and its comparative statics reflect whichever externality is quantitatively more important in the relevant parameter region.

VII Platform Liability and the Free-Rider Problem

Payment fraud increasingly originates on digital platforms. A Reuters investigation based on internal Meta documents revealed that the company internally projected earning approximately $16 billion annually from fraudulent advertisements, roughly 10% of total revenue, while serving around 15 billion high-risk scam ads per day (Horwitz, 2025).[16] Internal documents further estimated that Meta's platforms were involved in approximately one-third of all successful scams in the United States. The documents show that Meta required 95% certainty before banning fraudulent advertisers, and established a revenue guardrail of $135 million as the maximum it was willing to forgo to crack down on suspicious ads. An internal anti-fraud team developed a screening method that halved scam prevalence from China in testing; the team was subsequently disbanded after the revenue implications were flagged. In 2023, Meta platforms were linked to 54% of reported scam incidents in the UK.[17] Yet platforms face no statutory obligation to reimburse fraud victims ($\gamma_{\text{platform}} = 0$), while PSPs bear the full regulatory burden of reimbursement. This asymmetric allocation of liability creates a free-riding problem: platforms benefit from PSP detection efforts without internalizing any share of fraud losses, and as the Meta evidence illustrates, explicitly weigh the revenue from fraudulent activity against the cost of enforcement. Our model provides a formal rationale for why zero platform liability generates this outcome and characterizes the liability rule needed to correct it.

We formalize this intuition using the two-agent framework. A fraudulent transaction requires two sequential failures: the platform fails to detect or remove the fraudulent solicitation, and the PSP fails to flag the resulting payment. Let $I_{PSP} \in [0, 1]$ denote the platform's investment in fraud prevention, removing scam accounts, flagging suspicious content, and warning users, and $I_{plat} \in [0, 1]$ the PSP's investment in transaction screening. A fraud attempt succeeds with probability $(1 - I_{PSP})(1 - I_{plat})$, so each party's effort generates a positive externality for the other. We set $\theta = 1$ throughout this section to isolate the

[16] Meta disputed the characterization, describing the 10% figure as a rough and overly-inclusive estimate. https://www.reuters.com/investigations/meta-is-earning-fortune-deluge-fraudulent-ads-documents-show-2025-11-06/

[17] https://www.psr.org.uk/information-for-consumers/unmasking-how-fraudsters-target-uk-consumers-in-th

free-riding mechanism, abstracting from model decay.[18] The loss functions of the PSP and platform are:

$$L_{\text{PSP}} = \gamma_{PSP} \cdot \alpha V 2(1 - I_{PSP})(1 - I_{plat}) + \frac{1}{2} I_{PSP}^2, \tag{15}$$

$$L_{\text{plat}} = \gamma_{plat} \cdot \alpha V 2(1 - I_{PSP})(1 - I_{plat}) + \frac{1}{2} I_{plat}^2, \tag{16}$$

where $\gamma_{PSP}, \gamma_{plat} \in [0, 1]$ are the liability shares borne by the PSP and the platform respectively.

Proposition 7 (Optimal Liability for PSPs and Platforms)**.** *The welfare-maximizing liability parameters are:*

$$\gamma_{PSP}^* = \gamma_{plat}^* = 1.$$

Full liability for both the PSP and the platform, combined with a regulatory fee equal to total expected fraud losses, maximizes welfare and resolves the free-riding problem.

The intuition underlying Proposition 7 is straightforward. Because fraud requires both detection efforts to fail, each party's investment is complementary to the other's. Under any $\gamma_{plat} < 1$, the platform does not fully internalize the social cost of fraud originating on its service and underinvests in prevention. Under $\gamma_{plat} = 0$, the platform's first-order condition is independent of fraud losses entirely, generating zero voluntary investment in fraud prevention. Full liability for both parties aligns private incentives with the social optimum. The regulatory fee ensures that total reimbursement does not over-compensate consumers when both parties bear full liability simultaneously.

The current regulatory landscape departs sharply from this prescription. Platforms in the UK, EU, and US face no statutory reimbursement obligation for payment fraud originating on their services, despite generating a majority of fraud incidents. The Online Safety Act (UK, 2023) imposes content moderation duties on platforms but stops short of financial liability for fraud losses. Our results show that extending reimbursement obligations to platforms would improve welfare by internalizing the externality that zero-liability platforms currently impose on PSPs and consumers.

The same logic applies within the payment chain itself. Every transaction involves a sending PSP, which screens outgoing payments, and a receiving PSP, which vets incoming account holders through KYC checks. Because a fraudulent transaction requires both screening efforts to fail, the two PSPs are in exactly the same strategic position as the PSP-platform pair above. Proposition 7 therefore applies directly: the welfare-maximizing

[18] The general case with $\theta < 1$, which combines free-riding and model decay externalities, is characterized in Internet Appendix I.

rule sets $\gamma^*_{sending} = \gamma^*_{receiving} = 1$ for both sending and receiving PSPs, supplemented by a regulatory fee. The UK PSR's October 2024 APP Fraud regulations mandate a 50/50 reimbursement split ($\gamma_{sending} = \gamma_{receiving} = 0.5$), which our analysis shows is insufficient: each PSP bears only half the marginal cost of a fraudulent transaction, leaving the remainder as an externality on its counterpart and generating underinvestment by both. The general N-agent framework in Internet Appendix I extends this analysis to arbitrary combinations of PSPs, platforms, and other agents, characterizing the optimal liability rule in each case.

VIII Cross-Market Fraud Displacement

The baseline model assumes a single payment chain in which the PSP detects both fraud types. In practice, fraud reaches consumers through two distinct channels requiring fundamentally different detection technologies. In credit card fraud, a fraudster uses stolen card details to make unauthorized payments. The PSP can detect anomalous transaction patterns and block the payment; no platform is involved. In eBay fraud, a fraudster lists a non-existent item, collects payment from the buyer, and disappears. eBay can detect and remove fraudulent listings and suspend suspicious sellers; the PSP that processed the payment saw only a routine transaction and had no way to detect the deception upstream. We model this specialization and study how liability allocation shapes detection incentives.

The economy has two domains, each processing 1/2 of all transactions. The PSP domain is targeted by type A fraud, which succeeds with probability $1 - I_{\text{PSP}}$. The platform domain is targeted by type B fraud, which succeeds with probability $1 - I_{\text{plat}}$. At $t = 0$, a mass 1/2 of type A fraudsters targets the PSP domain and a mass 1/2 of type B fraudsters targets the platform domain.

At the end of period 0, fraudsters observe detection intensities and may switch domains at a convex cost. A type A fraudster chooses the share $p \in [0, 1]$ remaining in the PSP domain to maximize:

$$\Pi_F^A = \frac{1}{2}\alpha V\Big[p(1 - I_{\text{PSP}}) + (1 - p)(1 - I_{\text{plat}})\Big] - \frac{1}{2}(1 - p)^2, \tag{17}$$

yielding $p^* = 1 - \frac{1}{2}\alpha V(I_{\text{PSP}} - I_{\text{plat}})$. Symmetrically, a type B fraudster chooses the share $q \in [0, 1]$ remaining in the platform domain to maximize:

$$\Pi_F^B = \frac{1}{2}\alpha V\Big[q(1 - I_{\text{plat}}) + (1 - q)(1 - I_{\text{PSP}})\Big] - \frac{1}{2}(1 - q)^2, \tag{18}$$

yielding $q^* = 1 - \frac{1}{2}\alpha V(I_{\text{plat}} - I_{\text{PSP}})$. When $I_{\text{PSP}} > I_{\text{plat}}$, type A fraudsters migrate toward the

platform domain ($p^* < 1$) and type B fraudsters migrate toward the PSP domain ($q^* < 1$).

Period 0 fraud rates are $\phi_0^{\text{PSP}} = (1 - I_{\text{PSP}})$ and $\phi_0^{\text{plat}} = (1 - I_{\text{plat}})$. After adaptation, period 1 fraud rates are:

$$\phi_1^{\text{PSP}} = (p^* + 1 - q^*)(1 - I_{\text{PSP}}), \qquad \phi_1^{\text{plat}} = (q^* + 1 - p^*)(1 - I_{\text{plat}}). \tag{19}$$

Each domain bears its liability share of total fraud losses across both periods:

$$L_{\text{PSP}} = \gamma_{\text{PSP}} \cdot \frac{1}{2}\alpha V \cdot \left(\phi_0^{\text{PSP}} + \phi_1^{\text{PSP}}\right) + \frac{1}{2} I_{\text{PSP}}^2, \tag{20}$$

$$L_{\text{plat}} = \gamma_{\text{plat}} \cdot \frac{1}{2}\alpha V \cdot \left(\phi_0^{\text{plat}} + \phi_1^{\text{plat}}\right) + \frac{1}{2} I_{\text{plat}}^2. \tag{21}$$

The PSP minimizes L_{PSP} with respect to I_{PSP}, taking I_{plat} as given. The first-order condition is:

$$\frac{1}{2}\gamma_{\text{PSP}}\alpha V \left[-2 + \frac{1}{2}\alpha V \left(1 - I_{\text{plat}}\right)\right] + I_{\text{PSP}} = 0,$$

yielding the PSP's best-response function:

$$I_{\text{PSP}}^*(I_{\text{plat}}) = \frac{1}{2}\gamma_{\text{PSP}}\alpha V \left(2 - \frac{1}{2}\alpha V(1 - I_{\text{plat}})\right).$$

Symmetrically, the platform's best-response function is:

$$I_{\text{plat}}^*(I_{\text{PSP}}) = \frac{1}{2}\gamma_{\text{plat}}\alpha V \left(2 - \frac{1}{2}\alpha V(1 - I_{\text{PSP}})\right).$$

Solving the two best-response functions simultaneously yields the Nash equilibrium detection intensities I_{PSP}^* and I_{plat}^* as functions of γ_{PSP} and γ_{plat}. Substituting each agent's best-response function into the other and solving simultaneously yields the Nash equilibrium detection intensities:

$$I_{\text{PSP}}^* = \frac{\gamma_{\text{PSP}}\alpha V(4 - \alpha V)(4 + \gamma_{\text{plat}}\alpha^2 V^2)}{16 - \gamma_{\text{PSP}}\gamma_{\text{plat}}\alpha^4 V^4}, \tag{22}$$

$$I_{\text{plat}}^* = \frac{\gamma_{\text{plat}}\alpha V(4 - \alpha V)(4 + \gamma_{\text{PSP}}\alpha^2 V^2)}{16 - \gamma_{\text{PSP}}\gamma_{\text{plat}}\alpha^4 V^4}, \tag{23}$$

where the denominator $16 - \gamma_{\text{PSP}}\gamma_{\text{plat}}\alpha^4 V^4 > 0$ under the maintained assumption $\alpha V < 1$. Both equilibrium intensities are positive and interior. Under the symmetric case $\gamma_{\text{PSP}} =$

$\gamma_{\text{plat}} = \gamma$, the two expressions coincide and reduce to:

$$I^* = \frac{\gamma\alpha V(4 - \alpha V)}{4 - \gamma\alpha^2 V^2}. \tag{24}$$

Each agent's equilibrium investment is increasing in its own liability share. Importantly, I^*_{PSP} is also increasing in γ_{plat} and vice versa: higher platform liability raises platform investment, which reduces fraud in the platform domain, pushing type B fraudsters toward the PSP domain and raising the PSP's marginal return to detection. This strategic complementarity distinguishes the cross-market model from the baseline, where PSP investments are strategic substitutes. Each domain's investment is increasing in its own liability share and independent of the other's. However, a higher γ_{PSP} raises I^*_{PSP}, which induces fraudster migration toward the platform domain, increasing the platform's period 1 losses. This is the cross-market displacement mechanism: higher detection in one domain displaces fraud toward the other rather than eliminating it, the cross-market analogue of the within-market Whac-a-Mole documented in Section II.D.

Proposition 8 (Optimal Liability with Cross-Market Displacement)**.** *The welfare-maximizing liability parameters satisfy:*

$$\gamma^*_{PSP} = \gamma^*_{plat} = \frac{2}{(2 - \alpha V)(1 + \alpha V)} < 1$$

for $\alpha V \in (0, 1)$. Full liability induces overinvestment: each domain ignores how its detection displaces fraud toward the other domain.

The result parallels the baseline model: in both cases $\gamma^* < 1$ corrects overinvestment, but through different mechanisms. In the baseline, overinvestment arises because each PSP ignores how its detection accelerates fraudster adaptation within the payment chain. Here, overinvestment arises because each domain ignores how its detection displaces fraud across domains. The two externalities reinforce each other when both are present, strengthening the case for partial liability. Under zero platform liability ($\gamma_{\text{plat}} = 0$), the platform invests nothing and fraud migrates freely toward the unprotected domain, generating welfare losses increasing in αV. The optimal response is not to impose full liability on the platform but to set $\gamma^*_{\text{plat}} < 1$ at the level that balances detection incentives against displacement costs.

IX Conclusion

This paper studies payment fraud detection in a dynamic environment where fraudsters adapt strategically to detection technologies, causing model performance to decay over time.

We provide the first direct micro-level evidence of criminal adaptation using CrimeBB, one of the largest publicly available underground forum datasets. The release of ChatGPT in November 2022 triggered a near-doubling of within-creator AI–fraud co-discussion (+97%), with technically sophisticated creators driving early adoption and less skilled creators subsequently broadening participation, consistent with a general-purpose enabling shock that lowers skill barriers. Stacked event studies across 27 anti-fraud technology deployments reveal that countermeasure discussion grows monotonically after deployment while vulnerable-method discussion is unaffected, the empirical signature of the Whac-A-Mole mechanism. These findings fill a gap in the economics of crime literature: while prior work establishes that criminals respond to incentives through spatial reallocation (Maheshri and Mastrobuoni, 2021) and cross-type substitution (Draca et al., 2019), direct evidence of criminal adaptation has been absent because method development leaves no trace in administrative crime records.

We embed these empirical regularities in a model of payment fraud detection. The model restricts the general framework of Section II.A to two fraud types and endogenises PSP detection investment, yielding three main results. First, when PSPs do not bear fraud losses, detection is underprovided relative to the social optimum. Second, under full liability in competitive markets, PSPs overinvest: each PSP ignores how its detection intensity accelerates fraudster adaptation market-wide, imposing a negative externality on rivals. The optimal liability share is therefore strictly below one and increasing in the number of PSPs. Third, when PSPs can reinvest in detection technology, competition generates a ratchet effect: reinvestment restores effectiveness but accelerates adaptation, raising the future return to yet further investment. PSPs reinvest more frequently than is socially optimal under full liability, and the regulator must account for this when designing liability rules.

The paper further shows that fraud prevention is a joint responsibility across multiple agents whose efforts are complementary. When platforms bear no reimbursement obligation despite originating the majority of fraud solicitations, they free-ride on PSP detection efforts, systematically undermining prevention incentives. Our N-agent framework shows that the welfare-maximizing liability rule sets full liability for all agents whose detection efforts are complementary, supplemented by a regulatory fee, regardless of whether the agents are sending and receiving PSPs, PSPs and platforms, or other combinations. The UK PSR's 50/50 reimbursement split between sending and receiving PSPs falls short of this prescription, and the absence of platform liability represents the largest remaining regulatory gap.

You are right. You cannot prevent ratcheting through liability rules alone, and claiming you can overstates the result. The model shows ratcheting is individually rational, so the policy instrument is not prevention but correction through liability calibration. Revise to

two instruments and be honest about what the model delivers:

Taken together, the results imply that effective fraud regulation requires two complementary instruments: liability rules calibrated to the model decay externality rather than treating detection investment as static, and liability extension to platforms whose zero-liability status generates systematic underinvestment in fraud prevention. Future work could use proprietary transaction data to estimate adaptation rates in deployed fraud detection systems and calibrate optimal liability rules empirically.

References

Aho, A. V. and M. J. Corasick (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM 18*(6), 333–340.

Apple Inc. (2014). Apple Pay transforms mobile payments with iPhone. Press release.

Bayram, F., B. S. Ahmed, and A. Kassler (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems 245*, 108632.

Bhalerao, R., M. Katariya, B. Penber, S. Dashevskyi, J. Mugan, P. Neves, S. Le Blond, and J. van der Merwe (2019). Mapping the underground: Supervised discovery of cybercrime supply chains. In *Proceedings of the 2019 APWG Symposium on Electronic Crime Research*, pp. 1–16. IEEE.

Board of Governors of the Federal Reserve System (2020). Synthetic identity fraud in the U.S. payment system. Technical report, Federal Reserve.

Bolton, R. J. and D. J. Hand (2002). Statistical fraud detection: A review. *Statistical Science 17*(3), 235–255.

Carletti, E., A. Leonello, and R. Marquez (2023). Loan guarantees, bank underwriting policies and financial stability. *Journal of Financial Economics 149*(2), 260–295.

CipherTrace (2019). Cryptocurrency AML and sanctions compliance: 2019 status report. Technical report, CipherTrace.

Cloudflare (2019). Bot management: Detect and prevent automated attacks. Technical report, Cloudflare.

Coalition for Content Provenance and Authenticity (2022). Content provenance and authenticity specification v1.0. Technical report, C2PA.

Cresci, S., R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 963–972. ACM.

Dimmock, S. G., W. C. Gerken, and N. P. Graham (2018). Is fraud contagious? coworker influence on misconduct by financial advisors. *Journal of Finance 73*(3), 1417–1450.

Draca, M., T. Koutmeridis, and S. Machin (2019). The changing returns to crime: Do criminals respond to prices? *Review of Economic Studies 86*(3), 1228–1257.

Egan, M., G. Matvos, and A. Seru (2019). The market for financial adviser misconduct. *Journal of Political Economy 127*(1), 233–295.

EMVCo (2017). 3-D secure protocol and core functions specification, version 2.1. Technical report, EMVCo LLC.

Erel, I., L. H. Stern, C. Tan, and M. S. Weisbach (2021). Selecting directors using machine learning. *Review of Financial Studies 34*(7), 3226–3264.

European Parliament and Council (2015). Directive 2015/2366 on payment services in the internal market (PSD2). Technical report, European Union.

Featurespace (2018). ARIC: Anomaly recognition and intelligent control platform. Technical report, Featurespace.

Financial Action Task Force (2021). Updated guidance for a risk-based approach to virtual assets and virtual asset service providers. Technical report, FATF.

Foley, S., J. R. Karlsen, and T. J. Putniņš (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *Review of Financial Studies 32*(5), 1798–1853.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022). Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance 77*(1), 5–47.

Gao, J., J. Pacelli, J. Schneemeier, and Y. Wu (2020). Dirty money: How banks influence financial crime. Working paper. Available at `https://ssrn.com/abstract=3722342`.

Google Chrome Team (2020). SameSite cookie changes in Chrome 80. web.dev.

Goyal, S. and A. Vigier (2014). Attack, defence, and contagion in networks. *Review of Economic Studies 81*(4), 1518–1542.

Gretzky, K. (2018). EvilGinx 2.0. GitHub.

Griffin, J. M. and K. Mei (2024). How do crypto flows finance slavery? the economics of pig butchering. Working paper, University of Texas at Austin. Available at `https://ssrn.com/abstract=4742235`.

GSMA (2016). Remote SIM provisioning architecture. Technical report, GSM Association.

GSMA (2017). SS7 and diameter interconnect security: Guidelines and mitigations. Technical report, GSM Association.

Hauswald, R. and R. Marquez (2006). Competition and strategic information acquisition in credit markets. *Review of Financial Studies 19*(3), 967–1000.

Hayashi, F., T. Moore, and R. J. Sullivan (2015). The economics of retail payments security. *Federal Reserve Bank of Kansas City Economic Review 100*(2), 21–66.

Horwitz, J. (2025, November). Meta is earning a fortune on a deluge of fraudulent ads, documents show. Reuters.

Hunt, T. (2013). Have I been pwned. Online database.

ISO/IEC (2017). ISO/IEC 30107-3:2017 information technology – biometric presentation attack detection – part 3: Testing and reporting. *International Organization for Standardization*.

Jumio (2018). Digital identity verification: Trends in fintech and digital onboarding. Technical report, Jumio.

Korzhyk, D., V. Conitzer, and R. Parr (2011). Security games with multiple attacker resources. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Volume 22, pp. 273–279.

Laxminarayan, R. and G. M. Brown (2001). Economics of antibiotic resistance: A theory of optimal use. *Journal of Environmental Economics and Management 42*(2), 183–206.

Maheshri, V. and G. Mastrobuoni (2021). The race between deterrence and displacement: Theory and evidence from bank robberies. *Review of Economics and Statistics 103*(3), 547–562.

Meiklejohn, S., M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage (2013). A fistful of bitcoins: Characterizing payments among men with no names. *Proceedings of the 2013 Internet Measurement Conference*, 127–140.

Microsoft (2020). Detecting deepfake videos: Video Authenticator tool. Technical report, Microsoft.

Mondal, S. and P. Bours (2017). A study on continuous authentication using keystroke dynamics. *Proceedings of the International Workshop on Information Security Applications*, 210–221.

Parlour, C. A., U. Rajan, and H. Zhu (2022). When fintech competes for payment flows. *Review of Financial Studies 35*(11), 4985–5024.

Pastrana, S., D. R. Thomas, A. Hutchings, and R. Clayton (2018). CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1845–1854. International World Wide Web Conferences Steering Committee.

Piskorski, T., A. Seru, and J. Witkin (2015). Asset quality misrepresentation by financial intermediaries: Evidence from the RMBS market. *Journal of Finance 70*(6), 2635–2678.

Portnoff, R. S., S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson (2017). Tools for automated analysis of cybercriminal markets. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 657–666. International World Wide Web Conferences Steering Committee.

Prove (2019). Phone-centric identity verification platform. Technical report, Prove.

Rajan, U., A. Seru, and V. Vig (2015). The failure of models that predict failure: Distance, incentives, and defaults. *Journal of Financial Economics 115*(2), 237–260.

Recorded Future (2020). Recorded Future threat intelligence platform. Technical report, Recorded Future.

Stop Scams UK (2023). The impact of AI on fraud and scams. Technical report, Stop Scams UK.

Stripe (2017). Stripe Radar: Machine learning fraud detection for payments. Technical report, Stripe.

Tambe, M. (2011). *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.

Tian, E. (2023). GPTZero: Detecting AI-generated text. Online tool.

Vastel, A., P. Laperdrix, and W. Rudametkin (2020). FP-Scanner: The privacy implications of browser fingerprinting. *Proceedings of the 16th ACM Web Conference*, 1653–1663.

Vives, X. and Z. Ye (2025). Information technology and lender competition. *Journal of Financial Economics 163*, 103957.

von Ahn, L., M. Blum, N. J. Hopper, and J. Langford (2003). CAPTCHA: Using hard AI problems for security. *Advances in Cryptology – EUROCRYPT 2003*, 294–311.

Who Pays for Payment Fraud? Detection and Liability Rules under Strategic Fraudster Adaptation
Internet Appendix

A Variable Definitions

Table A1
Variable Definitions

Variable	Definition
Panel A: Dependent Variables (Post Level)	
1(GenAI)	Binary indicator equal to one if the post contains at least one generative AI term (e.g., ChatGPT, GPT, LLM, deepfake, Stable Diffusion).
1(Fraud)	Binary indicator equal to one if the post contains at least one fraud-related term (e.g., phishing, carding, ransomware, credential stuffing).
1(GenAI × Fraud)	Binary indicator equal to one if the post contains both a generative AI term and a fraud-related term.
1(Fraud Method)	Binary indicator equal to one if the post mentions a specific fraud method or tool (e.g., phishing kit, fullz, CVV, combo list).
1(GenAI × Method)	Binary indicator equal to one if the post contains both a generative AI term and a specific fraud method term.
Panel B: Stacked Event-Study Outcomes (Creator-Month Level)	
vuln_{ijt}	Binary indicator equal to one if creator i posted at least one vulnerable-method term associated with technology j in month t. Constructed from a dictionary of 178 unique terms describing the attack vector targeted by each anti-fraud technology.
ctr_{ijt}	Binary indicator equal to one if creator i posted at least one countermeasure term associated with technology j in month t. Constructed from a dictionary of 270 unique terms describing circumvention techniques for each anti-fraud technology.
Panel C: Creator-Level Characteristics	
After_t	Binary indicator equal to one for months at or after December 2022 (the month of ChatGPT's public release).
Pre-period posts	Total number of posts by creator i in the pre-period (November 2021 through November 2022). Used in logarithmic form, $\log(1 +$ pre-period posts), as a measure of creator experience.
$\log(1 + \text{sophistication posts})$	Natural logarithm of one plus the number of pre-period posts by creator i containing at least one term from the 212-term technical sophistication dictionary. Measured entirely before December 2022.
$\text{TechSoph}^{\text{hi}}$	Binary indicator equal to one if the creator's pre-period sophistication post count is at or above the median among creators with at least one sophistication post.

Countermeasure dictionaries are reported in Internet Appendix D. The technical sophistication dictionary is reported in Internet Appendix E. The full list of anti-fraud technology events and their deployment dates is in Internet Appendix C.

B Wald Tests

Table B2
Pairwise $\hat{\beta}$ Comparison — Full Sample

DV_1	DV_2	$\Delta\hat{\beta}$	SE	z	p
1(AI)	**1**(Fraud)	0.1055	(0.0150)	7.04	$0.0000 * **$
1(AI)	**1**(AI×Fr)	0.0706	(0.0100)	7.03	$0.0000 * **$
1(AI)	**1**(AI×M)	0.0822	(0.0098)	8.37	$0.0000 * **$
1(Fraud)	**1**(AI×Fr)	−0.0350	(0.0131)	−2.67	$0.0076 * **$
1(Fraud)	**1**(Method)	−0.0049	(0.0138)	−0.36	0.7211
1(Method)	**1**(AI×M)	−0.0185	(0.0079)	−2.35	$0.0189 * *$
1(AI×Fr)	**1**(AI×M)	0.0116	(0.0066)	1.76	0.0786∗

Notes: Full sample, Creator FE specification. Wald test under independence: $z = (\hat{\beta}_1 - \hat{\beta}_2)/\sqrt{\text{SE}_1^2 + \text{SE}_2^2}$. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C Anti-Fraud Technology Descriptions

This appendix documents the 27 anti-fraud technologies deployed between 2013 and 2023, organized into 9 fraud categories. Each technology is classified as either *primary* (core defense) or *secondary* (complementary defense) within its category. The deployment dates mark significant milestones such as product launches, regulatory effective dates, or major industry adoption signals. These technologies are used in the stacked event study (Section II.D).

C.1 Carding Fraud

Device Fingerprinting Maturation (Primary). Deployment Date: 2018-01-18. Device fingerprinting has evolved from a niche security technique into an industry-standard fraud detection mechanism. Fingerprinting algorithms combine passive browser and device attributes (canvas rendering, WebGL results, font availability, screen resolution, timezone, and plugin enumeration) to create a unique identifier that persists across sessions. The key deployment date marks the LexisNexis acquisition of ThreatMetrix for $817 million, which consolidated device fingerprinting as a core enterprise fraud capability (Vastel et al., 2020).

reCAPTCHA v3 Invisible Scoring (Secondary). Deployment Date: 2018-10-29. Google's reCAPTCHA v3 shifted from explicit user challenges to invisible, behavioral scoring of user interactions. Rather than asking users to solve puzzles, reCAPTCHA v3 silently

analyzes mouse movements, keystroke patterns, and session behavior to assign a risk score used to decide whether to challenge or allow transactions (von Ahn et al., 2003).

ML Fraud Scoring at Scale (Secondary). Deployment Date: 2016-06-01. Machine learning fraud scoring platforms like Sift Science and Feedzai transitioned from experimental projects to production systems at tier-1 financial institutions. These systems ingest hundreds of features and output real-time fraud probability scores. The June 2016 date reflects Sift Science's Series C funding round, which coincided with the first large-scale deployments at major banks (Bolton and Hand, 2002).

C.2 Account Takeover Fraud

Behavioral Biometrics Scaling (Primary). Deployment Date: 2017-06-01. Behavioral biometrics systems authenticate users based on unique behavioral patterns including keystroke dynamics, mouse movements, touch pressure, and typing rhythm. BioCatch pioneered continuous authentication that monitors user behavior throughout a session rather than just at login. The June 2017 deployment date reflects BioCatch's Series B funding and concurrent contracts with tier-1 banks (Mondal and Bours, 2017).

Real-Time Phishing Proxies (Offensive) (Secondary). Deployment Date: 2018-07-01. EvilGinx 2.0 represented a major advancement in reverse-proxy phishing tools that transparently intercept HTTPS traffic and enable real-time MFA bypass. Unlike earlier phishing proxies, EvilGinx 2.0 supported transparent session hijacking, allowing attackers to obtain valid authentication cookies and bypass two-factor authentication (Gretzky, 2018).

Credential Monitoring Services (Secondary). Deployment Date: 2013-12-04. Have I Been Pwned (HIBP), launched in December 2013, aggregates breached credential datasets and enables users and organisations to check whether their credentials appear in known breaches. For account takeover prevention, credential monitoring serves as an early warning system that allows preemptive defense through MFA enforcement or forced password resets (Hunt, 2013).

C.3 Banking Fraud

Open Banking APIs Replacing Screen-Scraping (Primary). Deployment Date: 2018-01-13. The Payment Services Directive 2 (PSD2), effective January 13, 2018, mandated that European banks provide open APIs for third-party access to account data. Open APIs

provide authenticated, standardized access with granular permission scoping and audit trails, eliminating the need for banks to expose customer credentials to third parties (European Parliament and Council, 2015).

Behavioral Analytics in Banking Apps (Secondary). Deployment Date: 2019-01-01. Featurespace's ARIC platform applies machine learning and behavioral analytics specifically to mobile and online banking applications. The January 2019 date reflects when ARIC had achieved deployments at 50+ banks globally, transitioning from pilot programs to mainstream adoption (Featurespace, 2018).

Device Fingerprinting in Banking (Secondary). Deployment Date: 2018-01-18. Banks adopted ThreatMetrix fingerprinting to identify whether a user is accessing their account from a recognized device and to flag access from suspicious devices. The LexisNexis acquisition of ThreatMetrix coincided with rapid adoption by major banks (Vastel et al., 2020).

C.4 Identity Fraud

ML Identity Verification Platforms (Primary). Deployment Date: 2019-01-01. Platforms such as Jumio, Onfido, and Socure introduced machine learning-based identity verification combining document analysis, liveness detection, and biometric matching. The January 2019 date marks when these platforms achieved critical mass adoption by major fintech companies, neobanks, and cryptocurrency exchanges (Jumio, 2018).

Liveness Detection in KYC (Secondary). Deployment Date: 2020-06-01. Liveness detection technology verifies that a presented facial image is from an actual live human rather than a spoof. iProov and FaceTec developed algorithms that challenge users to perform specific actions while capturing facial biometrics. The June 2020 date reflects the certification wave when multiple vendors achieved ISO 30107-3 compliance (ISO/IEC, 2017).

Synthetic Identity Detection (Secondary). Deployment Date: 2021-01-01. Socure's Sigma platform applies machine learning to detect inconsistencies in identity attributes that suggest synthetic construction. The January 2021 date reflects the Federal Reserve's publication of a white paper on synthetic identity fraud, which triggered widespread adoption of detection tools (Board of Governors of the Federal Reserve System, 2020).

C.5 Payment Systems Fraud

3D Secure 2.0 Rollout (Primary). Deployment Date: 2019-04-01. EMVCo 3-D Secure 2.0 replaces static password challenges with risk-based, dynamic authentication in e-commerce transactions. The April 2019 date reflects when Visa and Mastercard began enforcing 3DS 2.0 adoption in their mandate timelines (EMVCo, 2017).

Network Tokenisation (Apple Pay) (Secondary). Deployment Date: 2014-10-20. Apple Pay, launched October 20, 2014, pioneered network tokenization by substituting unique, single-use tokens for card numbers in each transaction. The iPhone's Secure Enclave stores the cryptographic keys required to generate tokens, preventing malware or network attackers from accessing raw card data (Apple Inc., 2014).

Real-Time Fraud Scoring APIs (Secondary). Deployment Date: 2018-01-01. Payment processors deployed real-time fraud scoring APIs enabling merchants to receive fraud risk assessments within milliseconds of transaction submission. Stripe Radar and Adyen's risk engine exemplify this approach. The January 2018 date reflects mass adoption by mid-market and enterprise merchants (Stripe, 2017).

C.6 SIM and Telecom Fraud

Phone-Centric Identity APIs (Primary). Deployment Date: 2020-01-01. Prove (formerly Payfone) developed identity verification APIs leveraging mobile network operators' core databases to verify phone number ownership. By January 2020, Prove had integrated its APIs into major bank authentication flows, enabling passwordless login based on phone number verification (Prove, 2019).

SS7 Monitoring Deployment (Secondary). Deployment Date: 2018-06-01. The GSMA published technical guidelines for detecting and mitigating SS7 protocol abuse in June 2018. Attackers exploit SS7 vulnerabilities to intercept SMS-based OTP messages. The June 2018 date reflects when the GSMA formally published guidance and carriers began deploying monitoring infrastructure (GSMA, 2017).

eSIM Technology Adoption (Secondary). Deployment Date: 2018-09-21. Apple's iPhone XS, launched September 21, 2018, was the first flagship phone to offer eSIM dual-SIM capability. eSIM reduces SIM swap fraud by making it significantly harder for attackers to transfer a victim's SIM to their own device (GSMA, 2016).

C.7 AI Content Detection Fraud

AI Content Detection (Primary). Deployment Date: 2023-01-02. GPTZero, launched in January 2023, was the first widely-used public tool for detecting AI-generated text. GPTZero and competing tools address fraud scenarios where attackers generate synthetic content at scale (Tian, 2023).

Deepfake Detection Tools (Secondary). Deployment Date: 2020-09-01. Microsoft released the Video Authenticator tool in September 2020 to detect manipulated or deepfake video content. Deepfake detection is relevant to identity fraud because attackers use deepfake video to bypass liveness detection in KYC workflows (Microsoft, 2020).

Content Provenance Standards (Secondary). Deployment Date: 2022-01-26. The Coalition for Content Provenance and Authenticity (C2PA) published specification v1.0 on January 26, 2022, enabling digital assets to carry cryptographically-signed metadata establishing their origin and edit history (Coalition for Content Provenance and Authenticity, 2022).

C.8 Circumvention and Evasion Fraud

Advanced Bot Management (Primary). Deployment Date: 2019-09-01. Cloudflare Bot Management, reaching general availability in September 2019, uses behavioral analysis, JavaScript fingerprinting, TLS fingerprinting, and machine learning to distinguish legitimate users from automated attackers (Cloudflare, 2019).

Browser Fingerprint Hardening (Secondary). Deployment Date: 2020-01-01. Google Chrome 80, released in January 2020, implemented SameSite cookie restrictions and reduced the entropy of exposed browser fingerprinting signals, restricting third-party cookie transmission by default and limiting access to certain screen properties (Google Chrome Team, 2020).

Threat Intelligence Platform Maturation (Secondary). Deployment Date: 2020-09-17. Recorded Future's IPO on September 17, 2020 signalled the maturity of dark web and threat intelligence monitoring as a core enterprise security capability. Threat intelligence platforms monitor underground forums to identify emerging fraud schemes and provide early warning of attack campaigns (Recorded Future, 2020).

C.9 Cryptocurrency Fraud

Blockchain Analytics at Scale (Primary). Deployment Date: 2017-02-01. Chainalysis emerged as a leading blockchain analytics platform with Series A funding in February 2017, concurrent with securing contracts with US law enforcement. Chainalysis applies machine learning and graph analysis to blockchain transaction histories to identify illicit activity and track stolen funds (Meiklejohn et al., 2013).

Travel Rule Implementation (Secondary). Deployment Date: 2021-06-01. The Financial Action Task Force (FATF) updated its guidance on virtual assets in June 2021, formalizing the travel rule requirement that cryptocurrency exchanges report transaction counterparties to comply with AML regulations (Financial Action Task Force, 2021).

Exchange KYC Standardisation (Secondary). Deployment Date: 2019-09-01. Binance announced in September 2019 that it would require complete KYC verification for all users, triggering industry-wide standardisation of KYC requirements across major exchanges (CipherTrace, 2019).

D CrimeBB Keyword Dictionaries

This appendix reports the complete keyword dictionaries used to score the CrimeBB corpus against the 27 anti-fraud technologies described in Appendix C. For each technology, we construct two curated dictionaries: *vulnerable-method* (Vuln) terms describe the attack vector the technology targets, and *countermeasure* (Ctr) terms describe techniques to circumvent the technology. Across all 27 technologies, there are 189 vulnerable-method terms and 298 countermeasure terms (443 distinct terms after deduplication). Each term is matched case-insensitively against the full text of each CrimeBB post using the Aho-Corasick algorithm (Aho and Corasick, 1975), which constructs a finite-state automaton from all dictionary terms and matches them in a single linear-time pass over the corpus.

Carding Fraud.

Technology	Stream	Keywords
Device fingerprinting maturation	Vuln	basic carding, beginner carding, bin checker, card checker, card tester, cc checker, fresh socks, gate checker, live cc, luhn check, magnetic stripe, no proxy, simple method, track 2, track2
	Ctr	anti-detect, antidetect, antidetect browser, browser fingerprint, canvas fingerprint, canvas spoof, device emulat, device id spoof, fingerprint spoof, gologin, iovation, linken sphere, multilogin, residential proxy, threat metrix, threatmetrix, user agent spoof, webgl fingerprint
reCAPTCHA v3 invisible scoring	Vuln	auto checker, automated card, bot card, brute force, card checker, mass check, scripted check
	Ctr	2captcha, anti-captcha, anticaptcha, captcha bypass, captcha farm, captcha solv, captcha token, hcaptcha, invisible captcha, recaptcha bypass, recaptcha v3
ML fraud scoring at scale	Vuln	basic method, beginner method, easy card, no check site, noob method, simple cashout
	Ctr	feedzai, forter, fraud detect, fraud score, machine learning, risk engine, risk score, riskified, sift science, signifyd, velocity check

Account Takeover Fraud.

Technology	Stream	Keywords
Behavioral biometrics scaling	Vuln	brute force, checker config, combo list, config, cracked account, credential stuffing, fresh log, hit combo, open bullet, openbullet, password list, sentry mba, sentrymba, simple login, storm
	Ctr	2fa bypass, behavioral, biocatch, cookie steal, cookie theft, evil ginx, evilginx, humaniz, mfa bypass, modlishka, mouse emulat, otp bot, real-time phish, realtime phish, reverse proxy phish, session hijack, session token, token steal, typing pattern

Technology	Stream	Keywords
Real-time phishing proxies	Vuln	basic phish, clone site, fake login, phishing kit, phishing page, scam page, static phish
	Ctr	evilginx, mitm phish, modlishka, muraena, necrobrowser, real-time phish, relay phish, reverse proxy, transparent proxy
Credential monitoring services	Vuln	database dump, leaked database, old breach, old combo, password reuse, stale log
	Ctr	exclusive combo, fresh combo, fresh log, haveibeenpwned, hibp, not leaked, private combo, private database, unchecked combo, unshared

Banking Fraud.

Technology	Stream	Keywords
Open banking APIs	Vuln	aggregator exploit, bank account hack, bank drop, bank login, cash out, manual transfer, online banking hack, screen scrap, web inject, wire transfer
	Ctr	api exploit, api token, behavioral analytics, consent phish, featurespace, oauth, open banking, randomis, session replay, token hijack, transaction pattern
Behavioral analytics in banking	Vuln	bank login, manual transfer, rdp bank, remote access, teamviewer, vnc bank
	Ctr	anomaly detect, behavioral analytics, featurespace, pattern evasion, randomis, risk engine, session fingerprint, slow transfer, transaction velocity
Device fingerprinting in banking	Vuln	bank login, basic access, direct login, no spoof
	Ctr	anti-detect, antidetect, browser fingerprint, device id, device spoof, emulat, threatmetrix

Identity Fraud.

Technology	Stream	Keywords
ML identity verification	Vuln	counterfeit id, fake birth, fake driver, fake id, fake license, fake passport, forged document, fullz, novelty id, paper id, photoshop id, scan, ssn, template
	Ctr	3d mask, ai generated id, camera inject, deepfake id, face swap, injection attack, iproov, jumio, kyc bypass, liveness, liveness bypass, onfido, selfie bypass, socure, synthetic identity, verification bypass, virtual camera
Liveness detection in KYC	Vuln	photo spoof, printed photo, screenshot selfie, selfie photo, static image
	Ctr	3d mask, camera inject, deepfake, face swap, iso 30107, liveness bypass, liveness detect, manycam, obs virtual, presentation attack, virtual camera
Synthetic identity detection	Vuln	authorized user, cpn, credit privacy number, credit profile, piggybacking, synthetic identity, tradeline
	Ctr	behavioral link, cross reference, device associ, identity graph, link analysis, socure sigma, synthetic detect

Payment Systems Fraud.

Technology	Stream	Keywords
3D Secure 2.0 rollout	Vuln	3ds1, mcsc bypass, no 3d secure, no 3ds, no verification, non vbv, payment intercept, replay attack, static password, vbv bypass
	Ctr	3d secure 2, 3d secure bypass, 3ds2, 3ds2 bypass, challenge bypass, emv 3ds, frictionless, low value exempt, merchant bypass, non sca merchant, sca exempt, tra exempt
Network tokenisation	Vuln	card number, magnetic stripe, pan theft, raw card data, stored card
	Ctr	apple pay exploit, device token, google pay exploit, network token, token, tokeniz, virtual card
Real-time fraud scoring APIs	Vuln	basic card, high velocity, mass charge, simple cashout

Technology	Stream	Keywords
	Ctr	adyen risk, cool down, drip charge, fraud score, risk score, slow charge, stripe radar, velocity limit

SIM and Telecom Fraud.

Technology	Stream	Keywords
Phone-centric identity APIs	Vuln	call carrier, carrier social, number port, port out, prepaid sim, sim clone, sim jacker, sim swap, social engineer carrier
	Ctr	carrier insider, corrupt employee, diameter, esim, esim exploit, insider, payfone, prove bypass, silent sms, sim farm, ss7, ss7 exploit, telesign, voip bypass
SS7 monitoring	Vuln	signaling attack, ss7 attack, ss7 exploit, ss7 intercept
	Ctr	cat2 bypass, diameter security, gtp attack, roaming exploit, signaling firewall, ss7 firewall
eSIM technology	Vuln	blank sim, physical sim, sim card theft, sim clone
	Ctr	esim exploit, esim hack, esim profile, qr code sim, remote sim

AI Content Detection Fraud.

Technology	Stream	Keywords
AI content detection	Vuln	ai generated, ai phishing, ai voice clone, chatgpt phish, chatgpt scam, copy paste gpt, deepfake, gpt email, gpt social engineer
	Ctr	ai detector bypass, bypass filter, custom model, dan prompt, fraudgpt, gptzero bypass, humaniz, jailbreak, paraphras, pentagpt, quillbot, uncensored model, undetectable ai, unfiltered llm, watermark remov, wormgpt
Deepfake detection tools	Vuln	ai face, ai voice, deepfake, face swap, voice clone
	Ctr	anti deepfake, deepfake detect, deepware, gan artifact, sensity, undetectable deepfake, video authenticat
Content provenance standards	Vuln	ai generated document, ai generated image, synthetic media

Technology	Stream	Keywords
	Ctr	c2pa, content credentials, content provenance, exif remov, metadata strip, watermark remov

Circumvention and Evasion Fraud.

Technology	Stream	Keywords
Advanced bot management	Vuln	basic bot, basic proxy, curl, headless browser, headless chrome, mechanize, phantomjs, rotate proxy, selenium, simple scrip
	Ctr	2captcha, akamai bypass, anti-captcha, anticaptcha, bot protect, captcha solv, cloudflare bypass, cloudflare challenge, datadome bypass, imperva bypass, ja3, ja3 spoof, kasada bypass, perimeterx bypass, playwright stealth, puppeteer stealth, residential proxy, tls fingerprint, undetected chromedriver, undetected_chromedriver
Browser fingerprint hardening	Vuln	cross site cookie, third party cookie, tracking cookie
	Ctr	cookie bypass, first party set, privacy sandbox, samesite bypass, storage partition
Threat intelligence maturation	Vuln	clearnet forum, open forum, public listing
	Ctr	counter intelligence, digital shadows, fed, flashpoint, honeypot, intel 471, invite only, law enforcement, le monitoring, operational security, opsec, private forum, recorded future

Cryptocurrency Fraud.

Technology	Stream	Keywords
Blockchain analytics at scale	Vuln	bitcoin launder, bitcoin tumbl, btc cash out, btc mixer, clean bitcoin, direct transfer, exchange cashout, no kyc exchange, unregulated exchange

Technology	Stream	Keywords
	Ctr	atomic swap, chain analysis, chain hop, chainalysis, coinjoin, cross chain, crystal blockchain, dex, elliptic, mixer, monero, privacy coin, tornado cash, traced, tumbler, untraceable, wasabi wallet, xmr, zcash
Travel rule implementation	Vuln	anonymous transfer, no kyc, p2p crypto, peer to peer, unregulated exchange
	Ctr	bridge, compliance avoid, decentrali, defi, fatf bypass, non custodial, self custody, travel rule
Exchange KYC standardisation	Vuln	anonymous exchange, exchange without id, no kyc exchange, unverified account
	Ctr	binance verified, bisq, buy kyc account, coinbase verified, dex, exchange account, kyc bypass, localbitcoins, localmonero, paxful, sell verified, verified account

E Technical Sophistication Dictionary

The technical sophistication measure described in Section II.B.5 is based on a curated dictionary of 212 terms spanning 13 categories of advanced technical knowledge. Categories 11 (Deepfake/Liveness Bypass) and 12 (AI-Generated Content for Fraud) are excluded from the sophistication measure because they overlap with the generative AI dependent variables. All terms are matched case-insensitively as substrings against post content.

Table E12
Technical Sophistication Dictionary (212 terms, 13 categories)

Cat.	Category Name	N	Terms
1	Exploitation & Vulnerability Research	34	zero day, 0-day, cve, proof of concept, poc, privilege escalation, remote code execution, rce, buffer overflow, heap spray, rop chain, return oriented, use after free, memory corruption, heap corruption, sql injection, cross site scripting, xss, path traversal, local file inclusion, lfi, command injection, server side request, ssrf, deserialization, dll inject, dll hijack, process injection, process hollow, shellcode, sandbox escape, type confusion, exploit chain, authentication bypass

Continued on next page

Cat.	Category Name	N	Terms
2	Malware Development & Architecture	32	rootkit, bootkit, uefi, kernel driver, kernel module, ring-0, crypter, obfuscat*, polymorphic, metamorphic code, dropper, stager, reverse shell, bind shell, meterpreter, c2 server, command and control, cobalt strike, reflective inject, thread hijack, amsi bypass, etw patch, edr bypass, av bypass, fully undetectable, fileless malware, botnet stub, persistence mechanism, ransomware payload, remote access trojan, infostealer, keylogger
3	Reverse Engineering & Low-Level	18	reverse engineer*, decompil*, disassembl*, ida pro, ghidra, ollydbg, hex editor, memory dump, windows api, ntapi, syscall, windows kernel, lsass, mimikatz, dcsync, token impersonation, lolbins, dll sideload
4	Network Attacks & Infrastructure	31	port scan, nmap, packet sniff, man in the middle, mitm, arp spoof, arp poisoning, dns poison, ssl strip, lateral movement, credential dump, pass the hash, kerberoast, active directory, eternalblue, metasploit, msf, msfvenom, hashcat, rainbow table, session hijack, cookie hijack, packet injection, ja3, nat traversal, winrm, ss7, signaling attack, diameter attack, gtp attack, roaming exploit

Continued on next page

Cat.	Category Name	N	Terms
5	Reverse Proxy Phishing & AiTM	13	evilginx, modlishka, muraena, necrobrowser, transparent proxy, real-time phish*, aitm, adversary in the middle, reverse proxy phish, relay phish, oauth abuse, consent phishing, token hijack
6	Stealer Logs & Credential Infrastructure	9	stealer log, redline log, raccoon log, fresh combo, private combo, combo list, log market, cookie stealer, token replay
7	Evasion & OPSEC	19	opsec, operational security, tails os, proxy chain, runtime detection, scantime fud, heuristic detection, antidetect, browser fingerprint, multilogin, gologin, puppeteer stealth, undetected chromedriver, cloudflare bypass, akamai bypass, imperva bypass, ja3 spoof, device associ*, behavioral link
8	CAPTCHA Solving & Bot Infrastructure	6	captcha solv*, 2captcha, anticaptcha, captcha bypass, recaptcha bypass, captcha farm
9	OSINT & Recon	10	osint framework, subdomain enumeration, google dorking, breach database, metadata extraction, certificate transparency, social media enumeration, dns enumeration, email reconnaissance, theharvester

Continued on next page

Cat.	Category Name	N	Terms
10	Specific Tools & Frameworks	12	setoolkit, evilginx2, openbullet, hydra brute, sqlmap, aircrack, impacket, burp suite, wireshark, bluetoothctl, hcitool, blue hydra
13	Crypto Privacy & Obfuscation	10	monero, xmr, tornado cash, wasabi wallet, coinjoin, chain hop, crypto mixer, btc mixer, non custodial, self custody
14	Account & Device Manipulation	9	buy kyc account, sell verified, synthetic identity, manufactured identity, complex synthetic, esim hack, esim exploit, qr code sim, remote sim provision
15	Advanced Payment Evasion	9	velocity limit, velocity evasion, pattern evasion, drip charge, slow charge, warm card, mimic legit, avs match, low and slow

Note: Categories 11 (Deepfake/Liveness Bypass) and 12 (AI-Generated Content for Fraud) are excluded because they overlap with the generative AI dependent variables. An asterisk (*) denotes a prefix match. Terms such as `cve` and `xss` are matched as whole words only to avoid false positives.

Table E13
Description of Twitter Account Features

Variable	Description
Bot (0/1)	Equal to 1 if the Twitter account is a bot, and 0 otherwise.
Traditional Bot (0/1)	Equal to 1 if the Twitter account is a traditional spambot, and 0 otherwise.
Social Bot (0/1)	Equal to 1 if the Twitter account is a social bot, and 0 otherwise.
Followers (#)	Number of followers the account has.
Account in lists (#)	Number of public Twitter lists the account has been added to by other users.
Default profile (0/1)	Equal to 1 if the account uses the default Twitter profile, and 0 otherwise.
Default profile image (0/1)	Equal to 1 if the account uses the default Twitter profile image, and 0 otherwise.
Verified (0/1)	Equal to 1 if the account is verified, and 0 otherwise.
Has URL (0/1)	Equal to 1 if the account lists a URL, and 0 otherwise.

Table E14
Fraud Types and Associated Keywords

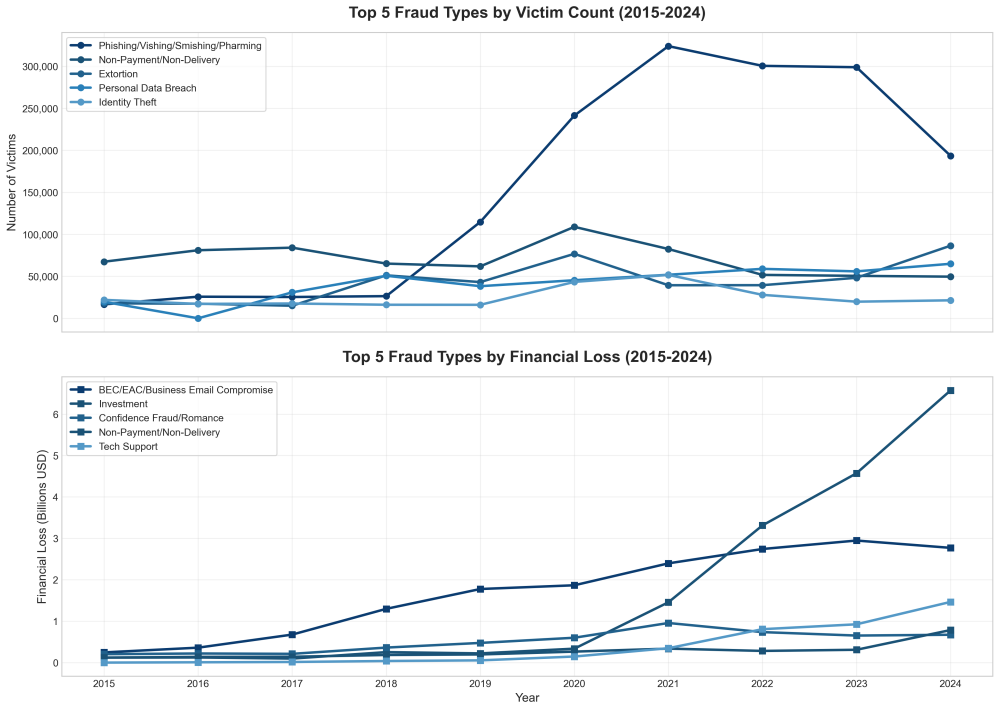
Fraud Type	Keywords
Freelance / Job Scams	upwork, fake client, fake job, overpayment, paypal invoice, fake invoice, advance payment, training fee, payoneer, job offer scam, remote job, work from home, freelance scam, freelance job scam, phone call interview
Payment Method Scams	gift card, western union, money transfer, bank transfer, paypal, cashapp, venmo, wire transfer, check, fake check
Marketplace Scams	ebay, amazon, fake listing, non-delivery, counterfeit, refund, return scam, fake buyer, payment upon delivery, tracking scam
Phishing Scams	phishing, link, credential, login, account takeover, stolen credentials, password reset
Identity / Account Scams	impersonation, fake id, irs, police, business email compromise, dmv
Investment / Financial Scams	cryptocurrency, crypto, fake coin, nft, ponzi, high return, pump-and-dump, investment scam, fake exchange, pyramid scheme, forex, binary options, day trading, get rich quick
Romance / Dating Scams	romance scam, catfish, dating app, long distance, soulmate, inheritance, stranded abroad, love scam, pig butchering, blackmail
Helpdesk / Tech Support Scams	tech support, fake tech support, remote access, remote desktop, virus, malware, popup, scareware
Impersonation (AI / Deepfake)	deepfake, voice clone, face swap, synthetic, ai generated, deepfake blackmail, deepfake impersonation
Website Copycat Scams	fraudulent website, copycat, clone, fake website, lookalike, mimic, spoof
Charity Scams	charity scam, fake charity, donation scam, disaster relief scam, philanthropy scam, fundraising scam, go fund me scams, crowdfunding scam
Unwanted / Robocalls	robocall, robocalling, robocalls, robocaller

F Corroborating Evidence

F.1 Changes in Fraud Type Share over Time

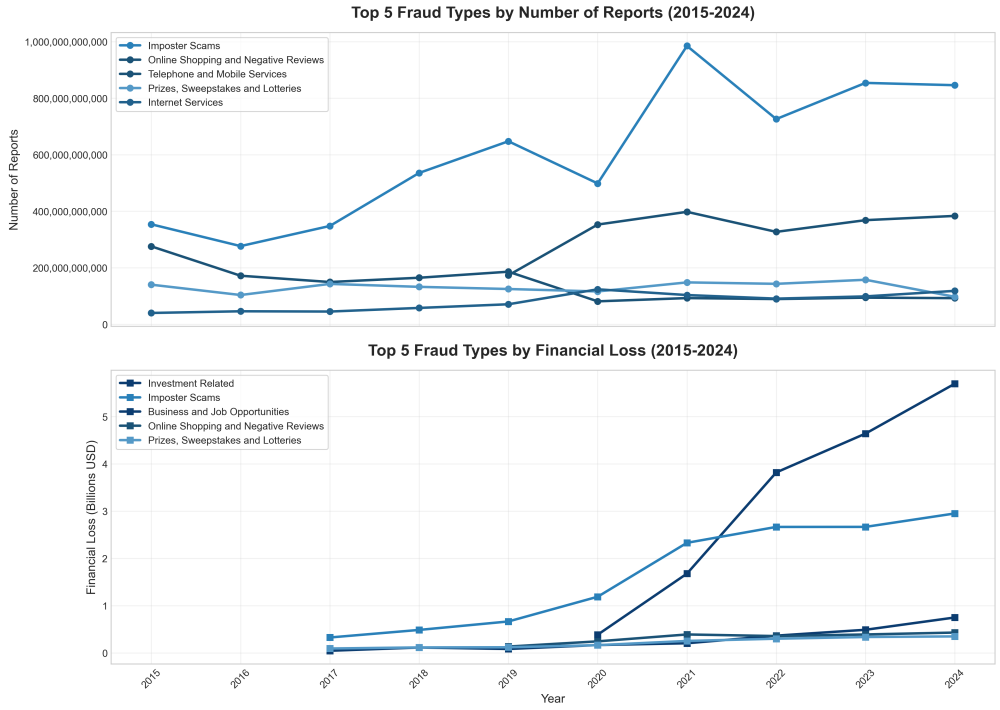
Before turning to micro-level evidence on the mechanism, we document that fraud-type composition changes substantially over time at the aggregate level. Figure F1 shows the five most common fraud types in the United States by complaint volume (top panel) and financial losses (bottom panel) over 2015–2024. The composition shifts markedly: phishing and impersonation fraud grow steadily by complaint count, while investment fraud dominates financial losses and rises sharply after 2020. The types generating the highest losses are not the same types generating the highest complaint volume, reflecting heterogeneity in V across fraud types. Figure F2 corroborates these patterns using FTC Sentinel data.

Figure F1
US Fraud Trends (FBI IC3)



Notes: Evolution of the five most common internet fraud categories in the United States by complaint volume (top panel) and financial losses (bottom panel). Source: FBI Internet Crime Complaint Center (IC3).

Figure F2
US Fraud Trends (FTC Sentinel)



Notes: Evolution of the five most common internet fraud categories in the United States by complaint volume (top panel) and financial losses (bottom panel). Source: FTC Sentinel.

These aggregate shifts have multiple potential sources, including the expansion of digital payment channels, changes in consumer behavior, and technological change. Most relevant for the present paper is the technology channel: anti-fraud technologies raise θ_k for targeted fraud types, inducing substitution toward harder-to-detect alternatives. Hayashi et al. (2015) provide direct evidence for this mechanism, documenting that card-not-present fraud rose sharply and became the dominant source of card fraud losses in the United Kingdom, France, and Canada in the years immediately following EMV chip card adoption, as fraudsters substituted away from counterfeit card-present fraud toward online channels where chip authentication offered no protection. Table F15 reports the underlying fraud statistics across four jurisdictions.

We also document aggregate fraud-type substitution using approximately 500,000 timestamped submissions from Reddit's /r/Scams community over 2012–2024, constructed by filtering monthly Reddit archives for content posted on the forum. Fraud types are identified using type-specific keywords such as "pump-and-dump" and "binary options" for investment fraud (see Table E14 for the complete list).

Figure F3 plots the distribution of scam topics over time. The composition shifts

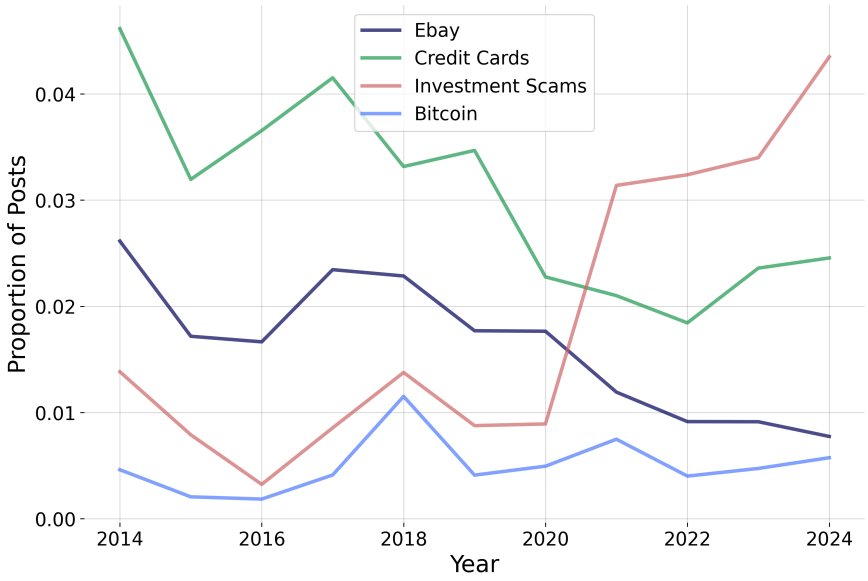
Table F15
Fraud Statistics by Jurisdiction

Panel A: Fraud per 100,000 inhabitants				
Year	US	EU	UK	China
2017	153.8	481.6	1087.5	24.6
2018	163.1	498.7	1175.1	28.1
2019	199.0	536.2	1305.7	38.9
2020	248.9	564.6	1393.8	50.7
2021	312.3	601.6	1697.0	46.5
2022	331.6	619.7	1922.0	38.4
2023	349.3	–	2054.8	40.4
2024	330.1	–	2073.2	–
Panel B: Fraud as Share of Total Crime				
Year	US	EU	UK	China
2017	0.04	0.16	0.12	0.03
2018	0.04	0.17	0.12	0.04
2019	0.05	0.19	0.13	0.06
2020	0.05	0.21	0.15	0.08
2021	0.07	0.22	0.16	0.07
2022	0.06	0.21	0.17	0.06
2023	0.06	–	0.19	0.07
2024	0.06	–	0.19	–

Notes: Fraud cases per 100,000 inhabitants and fraud as share of total reported crime in the US, EU, UK, and China from 2017 to 2024. Sources: US: FBI UCR; EU: Eurostat; UK: ONS Crime in England and Wales; China: National Bureau of Statistics.

markedly: discussions of credit card and eBay-related scams decline in prominence while investment fraud and cryptocurrency schemes rise sharply. Whether individual fraudsters switch tactics or the composition of active fraudsters changes over time is immaterial for the argument: in either case, the aggregate distribution of fraud types shifts over time, generating the composition-driven decay of equation (3).

Figure F3
Evolution in Fraud Types

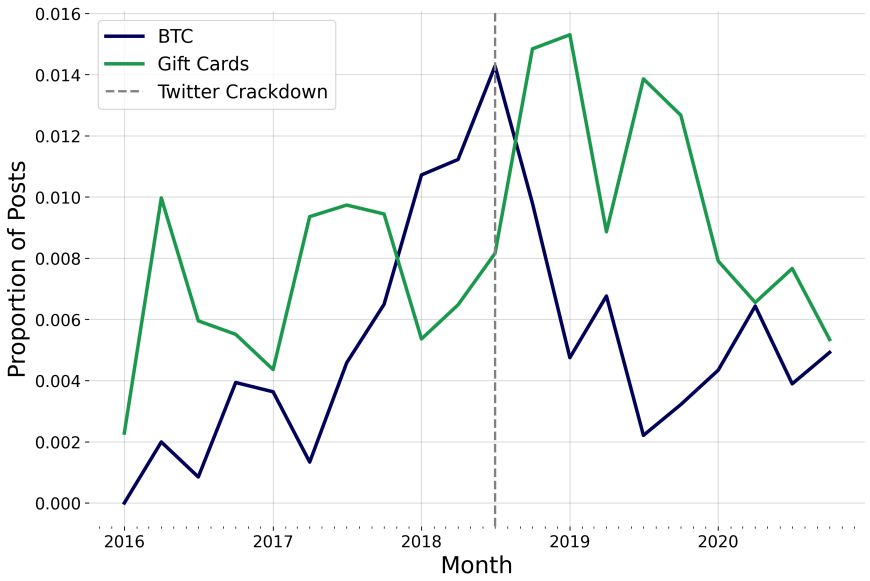


Notes: Distribution of scam topics over time using approximately 500,000 submissions from Reddit's /r/Scams community (2012–2024). Fraud types are identified using type-specific keywords (see Table E14).

Figure F4 focuses on Bitcoin-related scams and illustrates the targeted-shock mechanism of Section II.A. Following a rapid increase in cryptocurrency fraud on Twitter in early 2018, the platform cracked down on accounts engaged in such activity.[19] Bitcoin scam discussions on Reddit fell sharply in 2018, consistent with the technology effect in equation (3). However, fraudulent activity did not disappear: the data indicate substitution toward alternative scam types, including gift card fraud, consistent with the composition effect.

[19]See: tinyurl.com/FraudTwitter.

Figure F4
Twitter Crackdown Effect on Bitcoin Scam Discussion



Notes: Monthly count of Bitcoin-related scam submissions on Reddit's /r/Scams. The vertical line marks Twitter's crackdown on cryptocurrency scam accounts in early 2018.

F.2 Detection Model Decay

We provide direct evidence of composition-driven detection model decay using the Cresci et al. (2017) Twitter bot dataset, which contains 3,474 human-verified genuine accounts and 4,086 bot accounts spanning several years, including 2,631 traditional bots and 1,455 social bots. Traditional spambots are simple automated accounts that disseminate scams or promotional content. Social bots are more sophisticated, combining seemingly normal activity with coordinated amplification, and resemble genuine accounts more closely: 22 percent of social bots use the default profile setting, compared to only 4 percent of traditional bots and 30 percent of genuine users. Table F16 reports summary statistics.

We train three classifiers, logistic regression, random forest, and neural network, on accounts from 2007–2011 (2,327 accounts: 1,549 bots, 778 genuine) using a 70%/30% train-test split, and evaluate out-of-sample performance on accounts created in 2011–2014. Performance metrics are averaged across 10 random draws of 500 accounts per year to reduce sampling variability.

Table F17 documents clear model decay: all three classifiers exhibit sharp declines in accuracy and recall as social bots become more prevalent over time. To isolate composition-driven decay from confounds such as covariate drift, label noise, or overfitting, Table F18

Table F16
Summary Statistics: Twitter Bot Data

Variable	N	Mean	SD	Genuine Mean	Traditional Bots Mean	Social Bots Mean
Bot (0/1)	7,560	0.540	0.498			
Traditional Bot (0/1)	7,560	0.192	0.394			
Social Bot (0/1)	7,560	0.348	0.476			
Followers (#)	7,560	1,631.63	37,045.09	1,393.22	1,737.49	2,009.46
Account in lists (#)	7,560	21.07	367.75	19.50	31.07	6.73
Default profile (0/1)	7,560	0.193	0.395	0.297	0.041	0.222
Default profile image (0/1)	7,560	0.003	0.059	0.004	0.005	0.001
Verified (0/1)	7,560	0.001	0.038	0.003	0.000	0.000
Has URL (0/1)	7,560	0.463	0.499	0.364	0.744	0.189

Notes: Key account features for bots and non-bots from the Cresci et al. (2017) Twitter bot dataset. Variable definitions are in Table E13.

reports performance on test datasets that hold the overall bot share fixed at 66.6% while varying the proportion of traditional versus social bots from 75% to 25%. Performance declines monotonically as the share of social bots increases, confirming that composition change, not covariate drift, is the primary driver, consistent with the mechanism in equation (3).

Table F17
Model Performance Over Time

	Metric	2007–2010	2011	2012	2013	2014
Observations / Bots	Observations	500	500	500	500	500
	Bots (count / %)	333 (66.60%)	145 (29.00%)	265 (53.00%)	227 (45.40%)	220 (44.00%)
	Social Bots (count / %)	16.40 (3.28%)	100.90 (20.18%)	240.30 (48.06%)	205.90 (41.18%)	169.90 (33.98%)
Logistic Regression	Accuracy	0.71	0.41	0.54	0.52	0.55
	Recall (Bot)	0.99	0.48	0.65	0.59	0.60
	F1-score (Bot)	0.82	0.94	0.81	0.77	0.78
Random Forest	Accuracy	0.75	0.60	0.42	0.46	0.48
	Recall (Bot)	0.82	0.52	0.35	0.33	0.33
	F1-score (Bot)	0.81	0.75	0.29	0.29	0.29
Neural Network	Accuracy	0.72	0.43	0.55	0.53	0.55
	Recall (Bot)	0.97	0.49	0.65	0.59	0.60
	F1-score (Bot)	0.82	0.94	0.79	0.75	0.75

Notes: Average performance metrics across 10 random draws of 500 accounts per year. Models trained on the 2007–2010 sample.

Table F18
Model Performance Under Varying Social Bot Shares

	Metric	Traditional Bots = 0.75	Traditional Bots = 0.50	Traditional Bots = 0.25
Logistic Regression	Accuracy	0.69	0.68	0.66
	Recall (Bot)	0.84	0.82	0.80
	F1-score (Bot)	0.79	0.77	0.76
Random Forest	Accuracy	0.46	0.44	0.42
	Recall (Bot)	0.37	0.34	0.31
	F1-score (Bot)	0.48	0.44	0.42
Neural Network	Accuracy	0.68	0.66	0.65
	Recall (Bot)	0.81	0.79	0.77
	F1-score (Bot)	0.77	0.76	0.75

Notes: Model performance under varying proportions of traditional bots. The overall bot share is fixed at 66% across all test samples; only the composition of traditional versus social bots varies.

G Monopoly

In this section we analyze the benchmark case with a single PSP ($N = 1$). The loss function reduces to:

$$L = (\phi_0 + \phi_1)\,\gamma\alpha V + \frac{1}{2}I_0^2, \tag{25}$$

where $\phi_0 = 1 - I_0$ and $\phi_1 = 1 - I_0 + (1-\theta)^2\alpha V I_0^2$. The monopolist's optimal detection intensity coincides with the socially optimal level when $\gamma = 1$:

$$I_W^* = \frac{2\alpha V}{1 + 2(1-\theta)^2\alpha^2 V^2}, \tag{26}$$

because the monopolist fully internalizes the effect of model decay on its own fraud losses. This contrasts with the competitive case, where each PSP ignores the market-wide impact of its detection intensity on fraudster adaptation. Most payment sectors are competitive, which motivates our focus on the competitive case in the main text. While a monopolistic PSP internalizes the fraud externality, a public central bank digital currency (CBDC) operated by a central bank could achieve the same outcome without the deadweight loss associated with monopoly pricing.

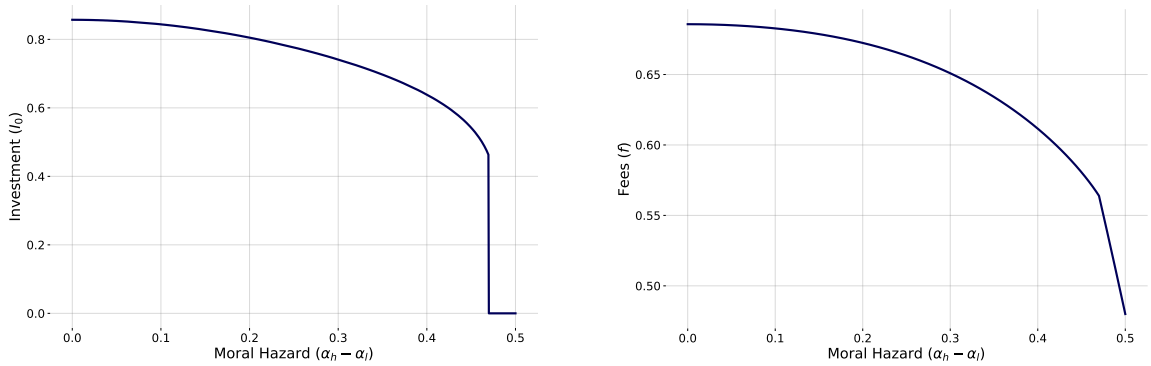
H Consumer Moral Hazard

One concern with shifting liability from consumers to PSPs is that consumers become less vigilant. Given detection intensity I_0, a consumer selects vigilance level c to minimize ex-

pected fraud losses net of vigilance costs:

$$\min_{c} \; \phi_0(1-\gamma)\left(\alpha_H - (\alpha_H - \alpha_L)c\right)V + g(c), \tag{27}$$

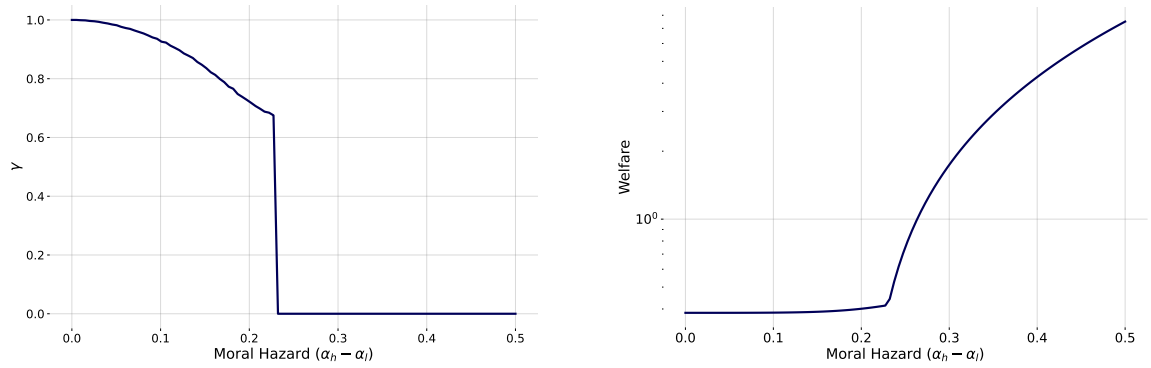
where α_H and α_L are the fraud exposure rates under low and high consumer vigilance respectively, and $g(c)$ is a convex cost of vigilance. Consumer vigilance decreases when PSPs invest more in detection, when PSPs bear more of the fraud costs (higher γ), and when the private gains from vigilance fall. Moral hazard creates a trade-off for the regulator: a higher γ strengthens PSP incentives but weakens consumer vigilance. The optimal γ^* depends on the relative effectiveness of the two efforts.



(a) Detection intensity I_0 **(b)** Equilibrium fraud losses

Figure H5
Detection Intensity and Fraud Losses with Consumer Moral Hazard

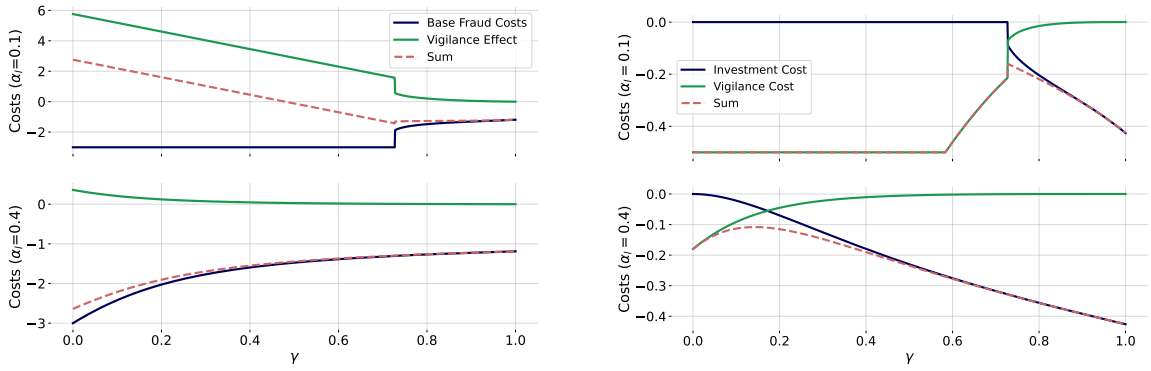
Notes: Panel (a) plots PSP detection intensity I_0 and panel (b) plots equilibrium fraud losses as a function of the consumer moral hazard parameter. Higher moral hazard reduces consumer vigilance, altering the PSP's optimal detection decision.



(a) Optimal γ **(b)** Welfare

Figure H6
Optimal Liability and Welfare with Consumer Moral Hazard

Notes: Panel (a) plots the optimal liability share γ and panel (b) plots welfare as a function of the consumer moral hazard parameter.



(a) Fraud Costs **(b)** Investment and Vigilance Costs

Figure H7
Liability Share and Costs

Notes: Panel (a) plots total fraud costs and panel (b) plots PSP investment and consumer vigilance costs as a function of the liability share γ.

I N-Agent Detection Game

This appendix develops the general N-agent framework that nests the baseline model, the PSP-platform application of Section VII, and the sending-receiving PSP extension as special cases. There are two types of agents: N_1 agents of type 1 (e.g., sending PSPs or PSPs) and N_2 agents of type 2 (e.g., receiving PSPs, platforms, or consumers). Each type-1 agent serves a fraction $1/N_1$ of transactions and each type-2 agent serves a fraction $1/N_2$ of transactions.

Let I_{1j} denote the detection investment of type-1 agent j and I_{2k} that of type-2 agent k, with averages $\bar{I}_1 = \frac{1}{N_1}\sum_j I_{1j}$ and $\bar{I}_2 = \frac{1}{N_2}\sum_k I_{2k}$.

I.1 Fraudster's Problem

Given average investments $\bar{I}_1$ and $\bar{I}_2$, a type-A fraud attempt succeeds with probability $(1-\bar{I}_1)(1-\bar{I}_2)$ and a type-B attempt succeeds with probability $(1-\theta_1\bar{I}_1)(1-\theta_2\bar{I}_2)$, where $\theta_1, \theta_2 \in [0,1]$ capture the relative detectability of type-B fraud by each agent type. The fraudster chooses a share $p \in [0,1]$ of type-A attacks to maximize:

$$\Pi_F(p; \bar{I}_1, \bar{I}_2) = \Big[p(1-\bar{I}_1)(1-\bar{I}_2) + (1-p)(1-\theta_1\bar{I}_1)(1-\theta_2\bar{I}_2)\Big]\alpha V - \frac{1}{2}(1-p)^2. \tag{28}$$

The first-order condition yields the fraudster's optimal share of type-A attacks:

$$p^*(\bar{I}_1, \bar{I}_2) = 1 - \alpha V\, \Delta(\bar{I}_1, \bar{I}_2), \tag{29}$$

where $\Delta(\bar{I}_1, \bar{I}_2) := (1-\theta_1\bar{I}_1)(1-\theta_2\bar{I}_2) - (1-\bar{I}_1)(1-\bar{I}_2)$ measures the differential success rate of type-B relative to type-A fraud. Higher average detection intensity by either agent type increases Δ and induces the fraudster to shift toward type-B attacks, generating the composition effect formalized in Section II.A.

I.2 Loss Functions and Best Responses

The loss function of a representative type-1 agent j is:

$$L_{1j} = \frac{\gamma_1 \alpha V}{N_1}\Big[(1-I_{1j})(1-\bar{I}_2) + \phi_{1j}(I_{1j}, \bar{I}_1, \bar{I}_2)\Big] + \frac{1}{2}I_{1j}^2, \tag{30}$$

where ϕ_{1j} captures the second-period fraud rate accounting for the fraudster's optimal reallocation toward type-B attacks, and $\gamma_1 \in [0,1]$ is the liability share borne by type-1 agents. The loss function of a representative type-2 agent k is defined symmetrically with γ_2 replacing γ_1. Each agent minimizes its loss function taking the other type's average investment as given.

Lemma 1 (Symmetric Best-Response Functions)**.** *Under symmetry within types, the best-*

response functions of type-1 and type-2 agents are:

$$I_1^* = f_1(I_2) := \frac{C_1\big[2(1-I_2) - K_1 b(I_2)\big]}{1 + C_1 K_1 a(I_2)}, \tag{31}$$

$$I_2^* = f_2(I_1) := \frac{C_2\big[2(1-I_1) - K_2 b(I_1)\big]}{1 + C_2 K_2 a(I_1)}, \tag{32}$$

where $C_1 = \frac{\gamma_1 \alpha V}{N_1}$, $C_2 = \frac{\gamma_2 \alpha V}{N_2}$, $a(x) = (1-\theta_1) + (\theta_1\theta_2 - 1)x$, $b(x) = (1-\theta_2)x$, $K_1 = \alpha V(1 + \frac{1}{N_1})\big((1-I_2) - \theta_1(1-\theta_2 I_2)\big)$, *and* $K_2 = \alpha V(1 + \frac{1}{N_2})\big((1-I_1) - \theta_2(1-\theta_1 I_1)\big)$.

The best-response functions reflect three channels. The direct channel captures the marginal return from investing against type-A fraud. The free-riding channel reflects strategic substitutability: as one type increases investment, the other can reduce its own effort and still achieve the same fraud success rate. The model decay channel partially counteracts free-riding: higher investment by one type induces fraudster substitution toward type-B attacks, which raises the marginal return to investment by the other type if its model retains effectiveness against type-B fraud. The relative magnitudes of θ_1 and θ_2 determine which channel dominates in equilibrium.

I.3 Equilibrium Existence

Proposition 9 (Symmetric Nash Equilibrium)**.** *There exists a symmetric Nash equilibrium* $(I_1^*, I_2^*) \in [0,1]^2$ *in which all type-1 agents choose* I_1^* *and all type-2 agents choose* I_2^**. For sufficiently small* V*, this equilibrium is unique.*

Proof. Define the best-response mapping $F : [0,1]^2 \to [0,1]^2$ by $F(I_1, I_2) = (f_1(I_2), f_2(I_1))$. The domain $[0,1]^2$ is compact and convex. The denominators $1+C_1K_1a(I_2)$ and $1+C_2K_2a(I_1)$ remain strictly positive for all $(I_1, I_2) \in [0,1]^2$ under the maintained parameter restrictions, so f_1 and f_2 are continuous on $[0,1]$, and F maps $[0,1]^2$ into itself. Existence of a fixed point $(I_1^*, I_2^*) = F(I_1^*, I_2^*)$ follows from Brouwer's fixed-point theorem. For sufficiently small V, the best-response functions exhibit a unique crossing, yielding a unique interior equilibrium. □

I.4 Optimal Liability Rules

The social planner minimizes total fraud losses plus detection costs, internalizing the effect of each agent's investment on the full fraud distribution including the fraudster's optimal reallocation. Under symmetric interior allocations $\bar{I}_1 = I_1$ and $\bar{I}_2 = I_2$, the planner's first-order conditions differ from private agents' conditions by the externality each agent imposes on the other through the model decay channel.

Proposition 10 (Optimal Liability Rules)**.** *For each type $i \in \{1, 2\}$, let $X_i = \alpha V H_i \Delta$ where $H_1 = (1 - I_2) - \theta_1(1 - \theta_2 I_2)$ and $H_2 = (1 - I_1) - \theta_2(1 - \theta_1 I_1)$. The liability parameter that aligns private and social incentives is:*

$$\gamma_i^* = \frac{2(1 - I_{-i}) - 2X_i}{2(1 - I_{-i}) + (1 + \frac{1}{N_i})X_i}. \tag{33}$$

Two cases follow directly. First, if $\theta_1 = \theta_2 = 1$ (no model decay), then $H_i = 0$, $X_i = 0$, and $\gamma_i^* = 1$ for both types. Second, if $\theta_1, \theta_2 < 1$ (model decay present) and $X_i > 0$, then $\gamma_i^* < 1$.

The no-decay case ($\theta_1 = \theta_2 = 1$) corresponds to the PSP-platform and sending-receiving PSP applications of Section VII, where free-riding is the dominant distortion and full liability corrects it. The decay case ($\theta_1, \theta_2 < 1$) nests the baseline competitive model of Section III.B, where the model decay externality induces overinvestment and the optimal liability share falls below one. The general formula characterizes the liability rule that balances these two opposing forces for arbitrary agent types, market structures, and decay parameters.

J Proofs

Proof of Proposition 1. The first-order condition for PSP i minimizing L_i with respect to $I_{i,0}$, using $\partial \bar{I}_0 / \partial I_{i,0} = 1/N$, is:

$$\frac{\gamma \alpha V}{N}\left[-2 + (1-\theta)^2 \alpha V \bar{I}_0 \left(1 + \frac{1}{N}\right)\right] + I_{i,0} = 0.$$

Imposing symmetry ($I_{i,0} = I_N^*$ for all i, so $\bar{I}_0 = I_N^*$) and solving gives:

$$I_N^* = \frac{2\gamma \alpha V}{N + \gamma \alpha^2 V^2 (1-\theta)^2 \left(1 + \frac{1}{N}\right)}.$$

Differentiating with respect to N:

$$\frac{dI_N^*}{dN} = -\frac{2\gamma \alpha V \left(1 - \frac{\gamma \alpha^2 V^2 (1-\theta)^2}{N^2}\right)}{\left[N + \gamma \alpha^2 V^2 (1-\theta)^2 \left(1 + \frac{1}{N}\right)\right]^2}.$$

This derivative is positive if and only if $\gamma \alpha^2 V^2 (1-\theta)^2 / N^2 > 1$, that is, $N < \gamma^{1/2} \alpha V (1 - \theta)$. ■

Proof of Proposition 2. *Part (i).* With $\theta = 1$: $I_W^* = 2\alpha V / N$ and $I_N^* = 2\gamma \alpha V / N$. Since $\gamma < 1$, $I_W^* > I_N^*$. *Part (ii).* With $\gamma = 1$ and $N > 1$: $I_N^* = 2\alpha V / [N + (1-\theta)^2 \alpha^2 V^2 (1 + 1/N)]$

and $I_W^* = 2\alpha V/[N + 2(1-\theta)^2\alpha^2V^2]$. The denominator of I_N^* is smaller than that of I_W^* if and only if $(1-\theta)^2\alpha^2V^2(1+1/N) < 2(1-\theta)^2\alpha^2V^2$, which simplifies to $1 + 1/N < 2$, that is, $N > 1$. Hence $I_N^* > I_W^*$. ■

Proof of Proposition 3. Setting $I_W^* = I_N^*$ and solving for γ: equating denominators gives $N + 2B = [N + \gamma B(1 + 1/N)]/\gamma$ where $B = (1-\theta)^2\alpha^2V^2$, which yields $\gamma(N + 2B) = N + \gamma B(1 + 1/N)$, hence $\gamma[N + 2B - B(1 + 1/N)] = N$, giving:

$$\gamma^* = \frac{N^2}{N^2 + (1-\theta)^2\alpha^2V^2(N-1)}.$$

Differentiating with respect to N:

$$\frac{\partial\gamma^*}{\partial N} = \frac{NB(N-2)}{\big(N^2 + B(N-1)\big)^2},$$

where $B = (1-\theta)^2\alpha^2V^2 > 0$. This is strictly positive for $N > 2$ and non-negative for $N \geq 2$. ■

Proof of Proposition 4. Under reinvestment, PSP i minimizes L_i^{new} over both I_{i0} and I_{i1}. The first-order conditions are $-(1/N)\gamma\alpha V + I_{i0} = 0$ and $-(1/N)\gamma\alpha V + I_{i1} = 0$, yielding $I_{i0}^* = I_{i1}^* = \gamma\alpha V/N$. Substituting the equilibrium detection intensity I_N^* into L_i^{old} and the optimal reinvestment intensities into L_i^{new}, the gain from reinvesting is $r_{\text{PSP}} = L_i^{\text{old}} - L_i^{\text{new}}$, which evaluates to:

$$r_{\text{PSP}} = \frac{A}{N^2} - \frac{4A}{ND} + \frac{4\gamma A\alpha^2V^2(1-\theta)^2}{ND^2} + \frac{2A}{D^2},$$

where $D = N + \gamma\alpha^2V^2(1-\theta)^2(1+1/N)$ and $A = \gamma^2\alpha^2V^2$. A PSP reinvests if and only if $r_{\text{PSP}} > 0$. The comparative statics $\partial r_{\text{PSP}}/\partial N < 0$, $\partial r_{\text{PSP}}/\partial\gamma > 0$, and $\partial r_{\text{PSP}}/\partial\theta > 0$ follow from differentiating r_{PSP} and noting that D is increasing in N and decreasing in γ and θ. ■

Proof of Proposition 5. Under $\gamma = 1$, the social planner's reinvestment gain r_W takes the same form as r_{PSP} but with D replaced by $D_W = N + 2(1-\theta)^2\alpha^2V^2$. For $N > 1$, $(1-\theta)^2\alpha^2V^2(1+1/N) < 2(1-\theta)^2\alpha^2V^2$, so $D < D_W$. Differentiating r with respect to D yields $\partial r/\partial D = 4A(D - N - 2(1-\theta)^2\alpha^2V^2)/(D^3N)$. At the equilibrium D, this is negative when $(1-\theta)^2\alpha^2V^2(1+1/N) < D - N$, which holds under the maintained parameter restrictions. Hence r is decreasing in D, and since $D < D_W$, we have $r_{\text{PSP}} > r_W$. ■

Proof of Proposition 6. We derive the optimal liability share by equating the private and social first-order conditions, then establish the comparative statics.

Step 1: Private equilibrium condition. From the loss function (14), the first-order condition for PSP i, using $\partial \bar{I}_0/\partial I_{i,0} = 1/N$ and imposing symmetry $I_{i,0} = \bar{I}_0 = I^*$, is:

$$\frac{\gamma \alpha V}{N}\left[(1-\lambda I^*)\left(-2+(1-\theta)^2 \alpha V I^*\left(1+\frac{1}{N}\right)\right)-\frac{\lambda \phi_{i,1}}{N}\right]+I^*=0, \tag{34}$$

where $\phi_{i,1} = 1 - I^* + (1-\theta)^2 \alpha V (I^*)^2$.

Step 2: Social planner's condition. The social planner minimizes:

$$W = \frac{1}{N}\sum_{i=1}^{N}\left[\phi_{i,0} + (1-\lambda \bar{I}_0)\phi_{i,1}\right]\alpha V + \sum_{i=1}^{N}\frac{1}{2}I_{i,0}^2.$$

Differentiating with respect to $I_{i,0}$ and imposing symmetry gives:

$$\alpha V\left[(1-\lambda I^*)\left(-2+2(1-\theta)^2 \alpha V I^*\right)-\frac{\lambda \phi_{i,1}}{N}\right]+I^*=0. \tag{35}$$

The social planner internalizes the full market-wide effect of detection on both the fraud success rate and the fraudster exit margin, reflected in the factor $2(1-\theta)^2 \alpha V I^*$ rather than $(1-\theta)^2 \alpha V I^*(1+1/N)$ in the private condition.

Step 3: Optimal γ^*. Define the private FOC as $G(\gamma, I^*) = 0$ and the social FOC as $H(I^*) = 0$. At the social optimum I^W satisfying $H(I^W) = 0$, the optimal γ^* is the value that makes $G(\gamma^*, I^W) = 0$. Dividing equation (34) by γ/N and equation (35) by 1, then equating the left-hand sides at $I^* = I^W$:

$$\frac{\gamma^*}{N}\left[(1-\lambda I^W)\left(-2+(1-\theta)^2 \alpha V I^W\left(1+\frac{1}{N}\right)\right)-\frac{\lambda \phi_{i,1}}{N}\right] = \left[(1-\lambda I^W)\left(-2+2(1-\theta)^2 \alpha V I^W\right)-\frac{\lambda \phi_{i,1}}{N}\right].$$

Solving for γ^*:

$$\gamma^* = N \cdot \frac{(1-\lambda I^W)\left(-2+2(1-\theta)^2 \alpha V I^W\right)-\lambda \phi_{i,1}/N}{(1-\lambda I^W)\left(-2+(1-\theta)^2 \alpha V I^W(1+1/N)\right)-\lambda \phi_{i,1}/N}. \tag{36}$$

When $\lambda = 0$, this reduces to $\gamma^* = N^2/[N^2 + (1-\theta)^2 \alpha^2 V^2 (N-1)]$ as in Proposition 3, confirming internal consistency.

Step 4: $\partial \gamma^*/\partial \theta > 0$. Let $\Phi(\gamma^*, \theta, \lambda) \equiv G(\gamma^*, I^W(\theta, \lambda)) = 0$ define γ^* implicitly. By the implicit function theorem:

$$\frac{\partial \gamma^*}{\partial \theta} = -\frac{\partial \Phi/\partial \theta}{\partial \Phi/\partial \gamma^*}.$$

The denominator $\partial\Phi/\partial\gamma^* > 0$ by the second-order condition. For the numerator, differentiating G with respect to θ: higher θ reduces $(1-\theta)^2$, which weakens the model decay term $(1-\theta)^2\alpha V I^*(1+1/N)$ in the private FOC. Since this term drives overinvestment, its reduction requires a higher γ^* to restore efficiency. Formally, $\partial\Phi/\partial\theta < 0$, giving $\partial\gamma^*/\partial\theta > 0$: when Type B fraud is more detectable, model decay is weaker, the overinvestment distortion diminishes, and the regulator raises γ^* toward one to maintain detection incentives.

Step 5: $\partial\gamma^*/\partial\lambda > 0$. Differentiating G with respect to λ: a higher λ strengthens the public good effect of detection through the $-\lambda\phi_{i,1}/N$ term, which appears symmetrically in both the private and social conditions. The net effect on γ^* depends on the relative magnitudes of the two externalities. Formally, differentiating equation (36) with respect to λ and evaluating at the equilibrium: the public good externality raises the social return to detection relative to the private return, requiring a higher γ^* to close the gap. Hence $\partial\gamma^*/\partial\lambda > 0$: a stronger fraudster exit response amplifies underinvestment pressure and the regulator responds by raising the liability share. ■

Proof of Proposition 8. *Step 1: Fraudster adaptation.* Given detection intensities I_{PSP} and I_{plat}, a type A fraudster chooses $p \in [0,1]$ to maximize:

$$\Pi_F^A = \frac{1}{2}\alpha V\Big[p(1-I_{\text{PSP}}) + (1-p)(1-I_{\text{plat}})\Big] - \frac{1}{2}(1-p)^2.$$

The first-order condition gives:

$$p^* = 1 - \frac{1}{2}\alpha V(I_{\text{PSP}} - I_{\text{plat}}).$$

Symmetrically, a type B fraudster chooses $q \in [0,1]$ to maximize:

$$\Pi_F^B = \frac{1}{2}\alpha V\Big[q(1-I_{\text{plat}}) + (1-q)(1-I_{\text{PSP}})\Big] - \frac{1}{2}(1-q)^2,$$

giving:

$$q^* = 1 - \frac{1}{2}\alpha V(I_{\text{plat}} - I_{\text{PSP}}).$$

Step 2: Fraud rates. Substituting p^* and q^*, the period 0 and period 1 fraud rates in

each domain are:

$$\phi_0^{\text{PSP}} = 1 - I_{\text{PSP}}, \qquad \phi_0^{\text{plat}} = 1 - I_{\text{plat}},$$
$$\phi_1^{\text{PSP}} = (p^* + 1 - q^*)(1 - I_{\text{PSP}}) = \Big(1 - \alpha V(I_{\text{PSP}} - I_{\text{plat}})\Big)(1 - I_{\text{PSP}}),$$
$$\phi_1^{\text{plat}} = (q^* + 1 - p^*)(1 - I_{\text{plat}}) = \Big(1 - \alpha V(I_{\text{plat}} - I_{\text{PSP}})\Big)(1 - I_{\text{plat}}).$$

Step 3: Social optimum. The social planner jointly minimizes:

$$W = \frac{1}{2}\alpha V\Big(\phi_0^{\text{PSP}} + \phi_1^{\text{PSP}} + \phi_0^{\text{plat}} + \phi_1^{\text{plat}}\Big) + \frac{1}{2}I_{\text{PSP}}^2 + \frac{1}{2}I_{\text{plat}}^2,$$

internalizing the effect of each domain's detection intensity on fraudster migration through p^* and q^*. Differentiating with respect to I_{PSP}, using $\partial p^*/\partial I_{\text{PSP}} = -\frac{1}{2}\alpha V$ and $\partial q^*/\partial I_{\text{PSP}} = \frac{1}{2}\alpha V$:

$$\frac{\partial W}{\partial I_{\text{PSP}}} = \alpha^2 V^2(I_{\text{PSP}} - I_{\text{plat}}) + I_{\text{PSP}} - \alpha V = 0.$$

Differentiating with respect to I_{plat}:

$$\frac{\partial W}{\partial I_{\text{plat}}} = \alpha^2 V^2(I_{\text{plat}} - I_{\text{PSP}}) + I_{\text{plat}} - \alpha V = 0.$$

Adding the two conditions eliminates the migration terms: $I_{\text{PSP}} + I_{\text{plat}} = 2\alpha V$. Subtracting gives $I_{\text{PSP}} = I_{\text{plat}}$. The unique social optimum is therefore:

$$I_{\text{PSP}}^W = I_{\text{plat}}^W = \alpha V.$$

The social optimum is symmetric even though the two agents are distinct: the planner equalizes detection across domains to eliminate fraudster migration incentives.

Step 4: Private first-order conditions. The PSP minimizes L_{PSP} over I_{PSP}, taking I_{plat} as given. Differentiating, using $\partial p^*/\partial I_{\text{PSP}} = -\frac{1}{2}\alpha V$ and $\partial q^*/\partial I_{\text{PSP}} = \frac{1}{2}\alpha V$:

$$\frac{\partial L_{\text{PSP}}}{\partial I_{\text{PSP}}} = \gamma_{\text{PSP}}\alpha^2 V^2 I_{\text{PSP}} - \frac{1}{2}\gamma_{\text{PSP}}\alpha^2 V^2 I_{\text{plat}} + I_{\text{PSP}} - \frac{1}{2}\gamma_{\text{PSP}}\alpha^2 V^2 - \gamma_{\text{PSP}}\alpha V = 0.$$

Rearranging gives the PSP's best-response function:

$$I_{\text{PSP}}^*(I_{\text{plat}}) = \frac{\gamma_{\text{PSP}}\alpha V\big(\alpha V I_{\text{plat}} + \alpha V + 2\big)}{2(1 + \gamma_{\text{PSP}}\alpha^2 V^2)}.$$

Symmetrically, the platform's best-response function is:

$$I^*_{\text{plat}}(I_{\text{PSP}}) = \frac{\gamma_{\text{plat}}\alpha V\big(\alpha V I_{\text{PSP}} + \alpha V + 2\big)}{2(1+\gamma_{\text{plat}}\alpha^2 V^2)}.$$

Solving the two best-response functions simultaneously yields the Nash equilibrium:

$$I^*_{\text{PSP}} = \frac{\gamma_{\text{PSP}}\alpha V(\alpha V+2)(3\gamma_{\text{plat}}\alpha^2V^2+2)}{3\gamma_{\text{PSP}}\gamma_{\text{plat}}\alpha^4V^4+4(\gamma_{\text{PSP}}+\gamma_{\text{plat}})\alpha^2V^2+4},$$

$$I^*_{\text{plat}} = \frac{\gamma_{\text{plat}}\alpha V(\alpha V+2)(3\gamma_{\text{PSP}}\alpha^2V^2+2)}{3\gamma_{\text{PSP}}\gamma_{\text{plat}}\alpha^4V^4+4(\gamma_{\text{PSP}}+\gamma_{\text{plat}})\alpha^2V^2+4}.$$

Under the symmetric case $\gamma_{\text{PSP}} = \gamma_{\text{plat}} = \gamma$, both expressions reduce to:

$$I^* = \frac{\gamma\alpha V(2+\alpha V)}{2+\gamma\alpha^2V^2}.$$

Step 5: Optimal γ^*. The optimal liability parameters make the Nash equilibrium coincide with the social optimum. Substituting $I_{\text{PSP}} = I_{\text{plat}} = \alpha V$ into the PSP's private first-order condition from Step 4:

$$\gamma^*_{\text{PSP}}\alpha^2V^2(\alpha V) - \frac{1}{2}\gamma^*_{\text{PSP}}\alpha^2V^2(\alpha V) + \alpha V - \frac{1}{2}\gamma^*_{\text{PSP}}\alpha^2V^2 - \gamma^*_{\text{PSP}}\alpha V = 0.$$

Collecting terms in γ^*_{PSP} and dividing by $\alpha V > 0$:

$$1 + \gamma^*_{\text{PSP}}\left(\frac{1}{2}\alpha^2V^2 - \frac{1}{2}\alpha V - 1\right) = 0.$$

Since $\frac{1}{2}\alpha^2V^2 - \frac{1}{2}\alpha V - 1 = -\frac{1}{2}(2-\alpha V)(1+\alpha V)$, solving gives:

$$\gamma^*_{\text{PSP}} = \frac{2}{(2-\alpha V)(1+\alpha V)}.$$

By symmetry of the two private first-order conditions under $I_{\text{PSP}} = I_{\text{plat}} = \alpha V$, the same calculation applies to the platform, giving $\gamma^*_{\text{plat}} = \gamma^*_{\text{PSP}}$.

Step 6: $\gamma^* < 1$. The condition $\gamma^* < 1$ requires $(2-\alpha V)(1+\alpha V) > 2$. Expanding: $2 + \alpha V - \alpha^2V^2 > 2$, which reduces to $\alpha V(1-\alpha V) > 0$. This holds for all $\alpha V \in (0,1)$. ■