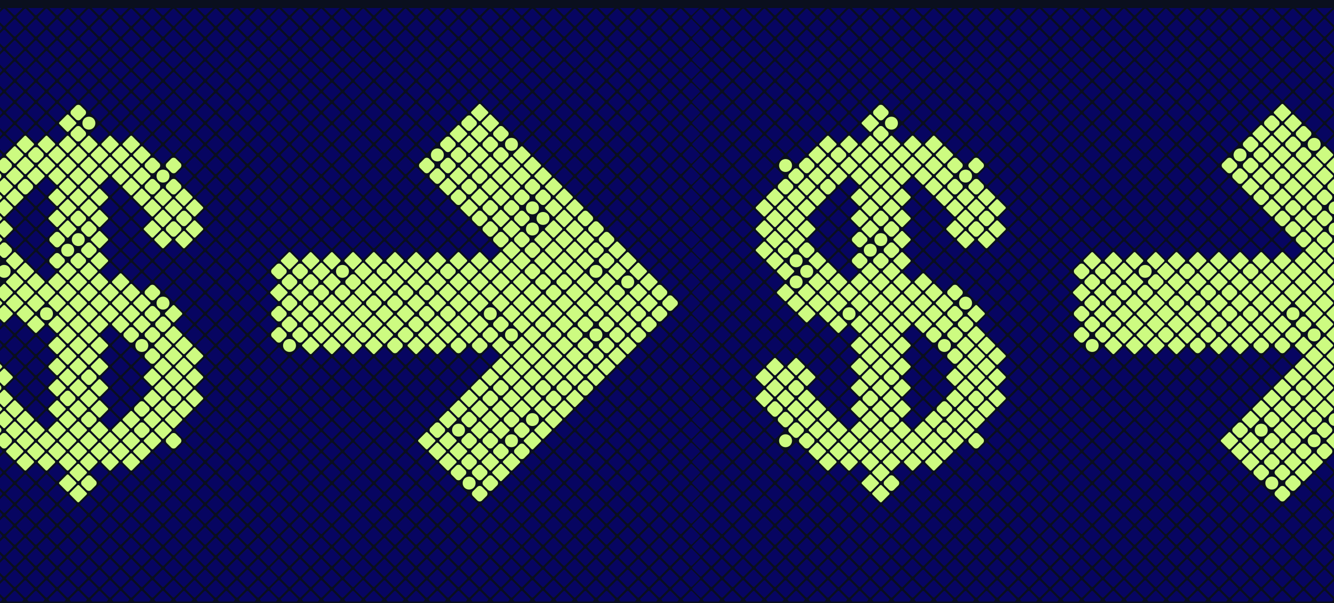
WHITEPAPER

Rethinking Fraud Prevention for Agentic Commerce

What breaks, what holds, and what's next



Sardine

AGENTIC PLATFORM TO
FIGHT FINANCIAL CRIME

Table of Contents

Agentic Commerce is already live. But your fraud strategy isn't ready for it.

For more than a year, Agentic Commerce has dominated industry headlines. Yet transaction volumes remain modest, early adopters drive most activity, and many teams assume there is still time to prepare.

That assumption may not hold through 2026.

As AI agents move from API experiments to in-browser autonomy, the structural foundations of eCommerce are shifting beneath us. Traditional fraud defenses are losing their grip: behavioral data is degrading, authentication is breaking, and verification is becoming dangerously fragmented. In this new landscape, merchants keep the liability while losing the visibility they rely on to manage it.

In the world of Agentic Commerce, fraud strategies are still full of gaps. Teams are identifying the problems faster than they can build the solutions.

For fraud, payments, and risk leaders who cannot afford to wait for clarity, this paper:

- Maps the structural vulnerabilities already visible in agentic flows
- Identifies the fraud typologies most likely to emerge next
- Pinpoints where existing controls degrade while highlighting where they still hold.
- Outlines how merchants, issuers, PSPs, and acquirers should begin recalibrating their defenses now.

Agentic Commerce may scale faster than expected. The question is whether your strategy would scale with it.

Executive summary

Agentic commerce introduces a new way for transactions to be initiated: by autonomous systems acting on behalf of users. These agents interact directly with merchants, often across multiple intermediaries, and execute purchases without real-time user involvement. While the underlying payment mechanisms remain familiar, the transaction chain introduces new intermediaries and becomes less directly observable. This changes what can be verified, by whom, and at what point in the flow. The payment rails remain the same; however, the trust assumptions do not.

In conventional eCommerce, verification and risk assessment rely on a relatively coherent set of signals: device context, behavioral continuity, network attributes, and user authentication events that merchants and issuers can independently observe. Agentic flows fragment this visibility. Signals are often missing, abstracted, or attributed to intermediaries rather than end users. Therefore, while existing fraud controls continue to block unsophisticated attacks, their effectiveness becomes uneven across different agentic implementations.

At the same time, automation itself becomes legitimate. Agents are expected to transact at scale and with speed, often across multiple merchants.

> In contrast to historical assumptions where automated behavior was treated as inherently suspicious, Agentic Commerce requires merchants to tolerate patterns that would previously have triggered intervention. This creates transition risk: **controls must be relaxed selectively, while adversaries face no such obligation.**

Agentic systems also expand the attack surface. Fraud can originate upstream through compromised AI applications, misused agent accounts, or delegated permissions that exceed user intent. In these cases, the transaction may appear valid at every observable checkpoint. The agent is authenticated, the payment is authorized, and the protocol behaves as designed, yet the outcome is still harmful.

Therefore, **defending Agentic Commerce cannot rely on a single control or actor.**

Effective defense requires a layered approach:

- Reinforcing traditional verification where possible
- Introducing standardized ways to understand and verify agents themselves
- Applying risk-based decisioning that explicitly accounts for agentic behavior

This paper outlines how these layers interact, where current defenses degrade gracefully, and why closing the remaining gaps will require coordination across merchants, issuers, networks, and AI providers alike.

01

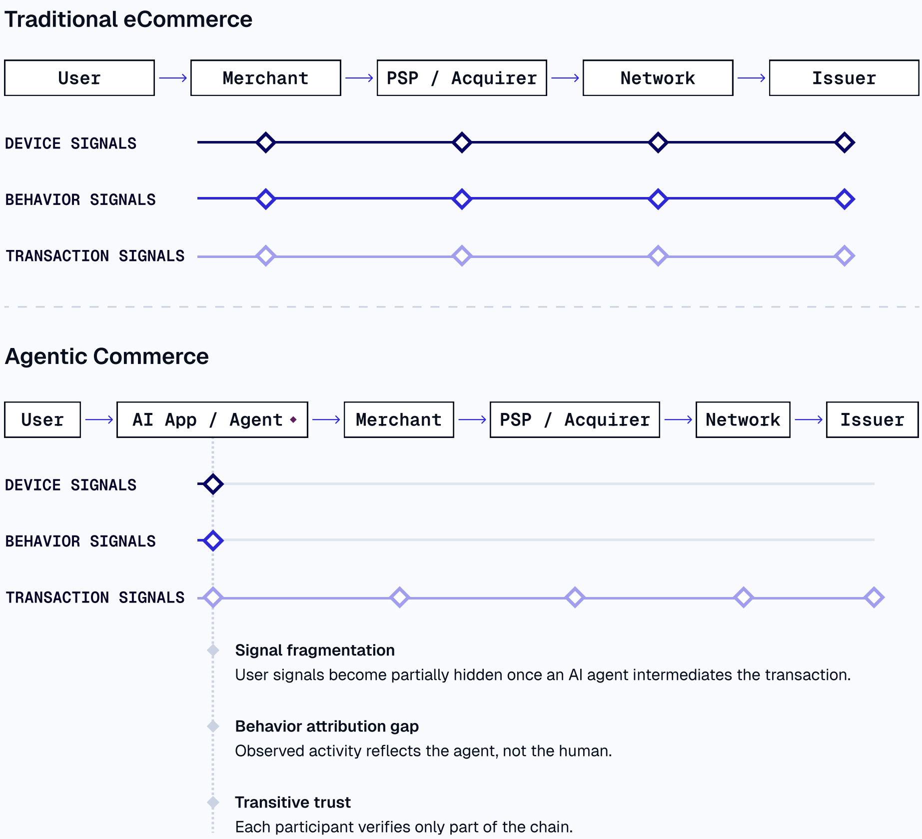
Start here:
What is Agentic Commerce?

Start here: What is Agentic Commerce?

Agentic Commerce refers to payment flows where the initiator of a transaction is an autonomous system (i.e. AI agent) acting on behalf of a user, rather than the user directly. The defining shift is not in how payments are processed, but in who initiates them, when, and through which interfaces.

Traditional vs. Agentic transaction visibility

AI agents introduce a new intermediary that fragments the signals used for fraud detection.



Agentic Commerce introduces an intermediary between the user and the merchant, reducing the visibility merchants and issuers rely on to verify identity, behavior, and intent.

In practice, most agentic transactions today are Merchant Context Protocol (MCP)-based. AI agents connect to merchant systems programmatically, retrieve product and pricing information, and execute purchases as structured API requests rather than through a browser session. From the merchant's perspective, these transactions often resemble backend service calls more than traditional checkouts. While in-browser agentic flows are emerging as a way for agents to navigate websites like humans, they remain secondary to MCP-based integrations in early production usage.

A natural question follows: **if Agentic Commerce is already live, can fraud and payments teams simply observe the data and user behavior to inform their strategy?** At first glance, this seems reasonable. Transactions are occurring, agents are transacting, and early patterns are visible. However, we believe the answer is no, or at the very least requires significant caution.

First, we expect the dominant technical manifestation of Agentic Commerce to evolve. /1

While MCP-based integrations account for most observable activity today, we view this as a reflection of early architectural convenience rather than a long-term design. As agentic experiences mature and move closer to mainstream consumer adoption, we expect a growing share of transactions to shift toward in-browser agentic flows, where agents operate within environments originally designed for human users.

This transition would materially change how transactions are initiated, what signals are available, and how intent is expressed. Meaning, behaviors observed in MCP-heavy traffic today may not generalize to future agentic flows.

Second, we believe that the current user population is highly unrepresentative. /2

Early Agentic Commerce usage appears concentrated among early adopters and power users, both legitimate and adversarial, who are willing to experiment with new capabilities, tolerate friction, and operate at the edges of what the ecosystem currently supports.

This population is structurally different from the broader consumer base that agentic interfaces are likely to reach over time. As access broadens and tooling becomes more seamless, we expect both user intent and behavior to shift in ways that are difficult to infer from early data.

Taken together, these factors suggest that material changes in flow design, user composition, and behavior are still ahead. As a result, we believe it's risky to draw strong conclusions or to lock in long-term fraud strategies based on the limited data available today. Early observations are useful for surfacing immediate failure modes, but insufficient for predicting steady-state risk.

This uncertainty matters. Even as Agentic Commerce becomes operational, foundational questions persist regarding where vulnerabilities will emerge and which fraud typologies will dominate. There is also no consensus on how responsibility will fragment across actors or how fraud teams should adapt their defenses. **The gap between what is visible now and what is likely to emerge next is larger than it may initially appear.**

Bridging that gap requires examining the structural shifts already underway and identifying where new fraud risk is most likely to emerge as verification and visibility change.

02

Agentic Commerce's inherent vulnerabilities

Agentic Commerce's inherent vulnerabilities

Before considering how fraudsters could exploit Agentic Commerce flows, we must first examine how these flows diverge from traditional eCommerce. These points of divergence often become points of failure because fraud defenses are optimized for human behavior. Fraudsters likely already search for these vulnerabilities, as their attacks are typically preceded by extensive penetration testing of new payment architectures.

This complexity increases because AI companies are newcomers to the payments security ecosystem. It's safe to assume there is an inherent gap between their current capabilities and the mature fraud defenses utilized by banks and fintechs. Because integrating a new participant into a multi-party flow requires complex new protocols, the ecosystem will be hard-pressed to iron out early kinks before fraudsters find them.

Part of the complexity comes from the fact that adding a new actor to the transaction flow isn't just about adding vertical complexity, but a horizontal one too. Card transactions already introduce many vulnerabilities for the number of players that can take up the different roles in the flow. Differences in protocols between schemes, issuers, acquirers, and merchants create a varied environment that will inevitably grow more complex with the introduction of AI providers.

This variety ensures that certain actor combinations will result in less secure flows that fraudsters can quickly uncover. Similarly, patching these vulnerabilities does not necessarily end an attack. Instead, it often initiates a migration toward the next most vulnerable flow in the ecosystem.

Degrading data integrity and its impact on fraud detection

As we touched upon earlier, Agentic Commerce can technically manifest in two ways: in-browser agentic flows and MCP-based flows. Both flow types are expected to significantly influence the data we can collect on the user, and how accurately we can interpret it.

Data integrity challenges

In-browser agentic flows would manipulate user data in one of two ways. The first scenario would present a redacted user data, as in the cases of headless or semi-headless agentic flows. In the table below, you can see how it'll affect the user data you collect:

SIGNAL	NORMAL BROWSER	HEADLESS / SEMI-HEADLESS
`navigator.plugins`	Non-empty	Empty
`navigator.mimeTypes`	Non-empty	Empty
`navigator.languages`	Realistic list	Single value / default
`navigator.hardwareConcurrency`	Plausible (4–16)	Fixed low value
`navigator.deviceMemory`	Present	Missing or fixed
`window.chrome` **(Chrome)**	Fully populated	Partial / missing
Media codecs support	Broad	Narrow / inconsistent

But missing data represents the best case as some in-browser flows might rely on mediated or remote browser sessions. This would mean that not only data would be missing, but also that present data might be misrepresenting the real user. Similarly, MCP-based agentic flows would likely be either missing data completely (e.g., useragent), using corrupted values (e.g., default, common values), or even the wrong values (e.g., the agent App's).

Needless to say that each of these scenarios (missing, corrupted, or wrong data values) would cause a significant disruption of any fraud prevention system. Whether it's based on manual reviews, rules, ML models, or AI agents, every fraud prevention defense is built first and foremost on the data feeding it. Degraded data quality will inevitably cause degraded performance.

We also need to keep in mind that user data is often just the base layer on which advanced fraud prevention systems are built, and so the degradation doesn't stop there. If the IP address that's passed to the merchant is incorrect, how can its location be compared to the user's card, address, or phone number? If device data is generic, how can a merchant effectively link it to known bad actors or calculate velocity? Such data issues could cause compounding failures that help fraudsters remain undetected while simultaneously increasing false positive rates.

The erosion of behavioral context and user intent

As described above, processing agentic payments prevents us from reliably "touching" the user's device or the user themselves. This creates a fundamental attribution gap because any "biometric" data collected reflects the agent's behavior rather than the human's identity. Signals such as typing speed, scrolling patterns, mouse movements, input capitalization, and touch dynamics lose their value as authentication factors in this environment.

But the disruption doesn't stop there. Beyond biometrics, agentic commerce obscures critical contextual and environmental signals. Data points such as activity timestamps, clickstream patterns, decision velocity, and precise geolocation are foundational to risk assessment because they facilitate:

- **Passive verification:** Establishing behavioral consistency to reduce the need for active re-authentication.
- **Intent analysis:** Identifying patterns associated with malicious actors.
- **Bot detection:** Distinguishing between human users and automated scripts.

However, in an agentic environment, these signals are stripped of their diagnostic value. Because bot activity is no longer a definitive indicator of fraud, the underlying rules and models become significantly less accurate or entirely defunct.

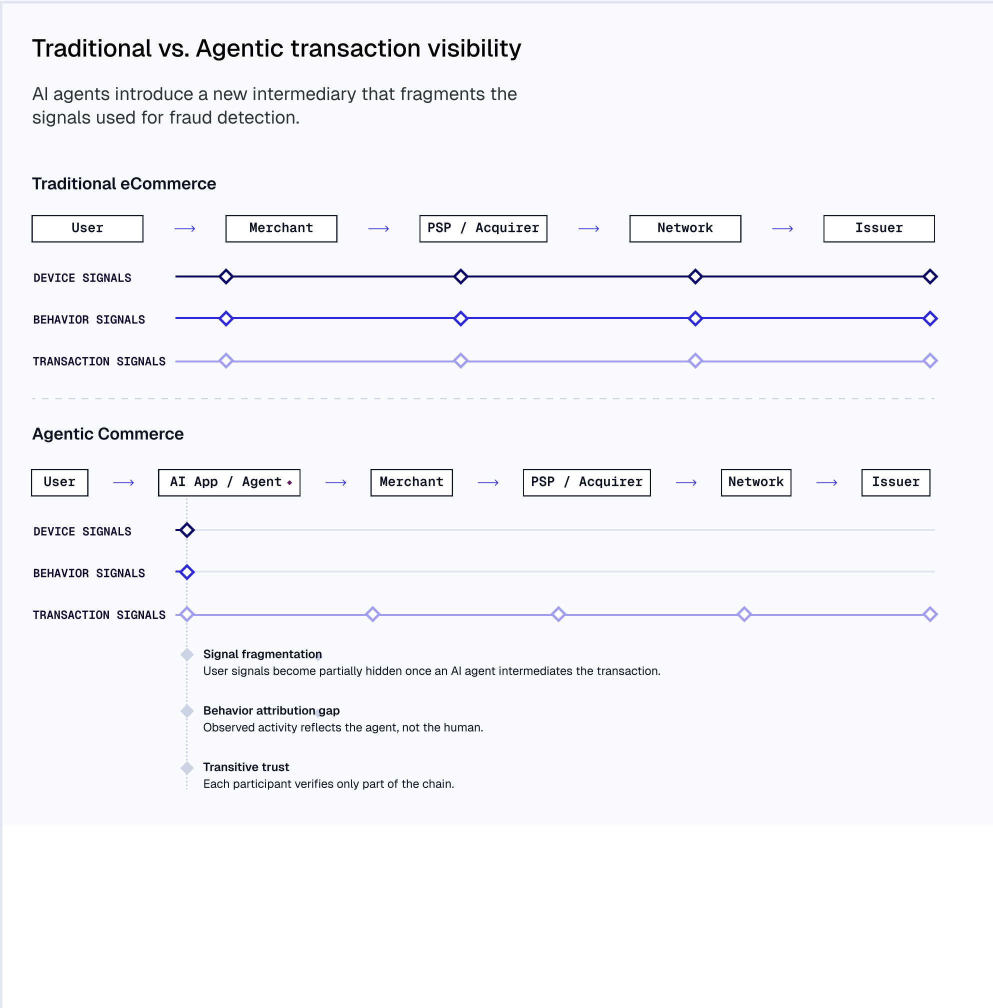
Agentic friction distorts downstream performance metrics

Agentic behavior distorts the performance metrics that fraud and payments teams rely on to understand user behavior and system health. Declines, authentication failures, and abandoned checkouts take on a different meaning when actions are executed by an AI agent rather than by a human, yet they are still recorded and acted upon in the same way.

For example, an AI agent may retry purchases across multiple merchants, payment methods, or price points in rapid succession as part of an optimization process. What looks like exploratory behavior from the agent's perspective can manifest as repeated declines, partial authorizations, or abandoned carts on the payments side. These events feed directly into issuer and network risk models, as well as merchant-side controls, increasing the likelihood of subsequent declines when the user later attempts to place payments.

Authentication flows amplify this effect. Failed or delayed Two-Factor Authentication (2FA) challenges, whether due to timing mismatches, asynchronous agent behavior, or user unavailability, may be logged as failed verifications rather than deferred ones. Over time, this can condition issuers and merchants to treat the user as higher risk, even though the failures were caused by flow misalignment rather than malicious intent.

The result is a feedback loop where agentic behavior degrades the user's payment profile over time, increasing friction and decline rates in subsequent, fully legitimate, human-initiated transactions. This risk is subtle, accumulative, and largely invisible to issuers and merchants alike, unless explicitly modeled for.



Repeated agent retries can unintentionally generate risk signals across the payment ecosystem, creating a feedback loop that increases friction and declines for legitimate users.

Defensive obsolescence and control gaps

Setting aside data integrity issues, another defense vulnerability might arise from the unavailability of certain risk controls in agentic flows. These can be the result of conscious design choices as well as architectural oversight.

The incompatibility of human-centric defenses

Traditionally, companies implemented anti-bot controls to filter out all non-human traffic. But we are already seeing this shift; if we now accept bots as legitimate users, what's the use of CAPTCHA challenges? This evolution requires a **rethinking of how, where, and when we implement these challenges across different transaction flows.**

Simultaneously, **we must rethink our approach to rate limiting.** For high-traffic merchants specifically, it is plausible to expect that traditional velocity checks will begin to break down once agentic commerce becomes a mainstream flow. This vulnerability applies to both network security, such as IP ranges, and fraud prevention, including both IP ranges and device signatures.

The risk here isn't that such measures will be discarded entirely, leaving the system open to known attack vectors. Rather, these controls must undergo an extended period of live calibration to remain effective as we learn how this new population behaves. Such a transition carries inherent risk; the system becomes noisier, changes are more frequent, and fraudsters tend to move faster than their targets.

Verification flows must also evolve. Traditional 2FA methods, such as OTP and 3DS, are time-bound because they assume a person is aware of and reactive to the transaction process. As these actions are delegated to AI agents, it's reasonable to expect the reaction time to such challenges will dramatically change.

Given that all actors would be interested in retaining high approval rates, and assuming that 2FA would subsequently change to enable it, the question becomes: how much fraud is going to slip through the cracks? Will it even be possible to implement such changes in all 2FA flows?

Integration gaps in multi-party agentic flows

Architecturally, Agentic Commerce presents a fragmented landscape even before launch. This fragmentation comes from the split between API-based and in-browser flows, security differences between AI apps, different implementations on the PSP level, and multiple security protocols from payment schemes. The flows are already fractured when considering card payments alone, not to mention the broader complexity of the payment ecosystem.

Given this complexity, and the industry's rush to enable the technology, unintentional security gaps inevitably emerge. While the granular details remain difficult to forecast, the primary vulnerability stems from the missing or partial integration of fraud prevention systems across these flows.

These systems can be internal systems such as rules engine, fraud models, in-house dashboards, or certain actions and restrictions. They also encompass external systems, such as fraud vendors, 3DS servers, and data providers. All these components are tied together with a spaghetti-like system of APIs and data pipelines. Such systems are sensitive to changes and carry the risk of critical omissions when teams set out to integrate Agentic Commerce.

We also expect these same pipelines to move lower-integrity data. The cascading data enrichment flows here can easily cause both rules and models to completely fail rather than simply degrade in performance.

For example, let's assume our best-performing rule focuses on cases where the IP location is far removed from the user's shipping address. In MCP-based flows, this data would likely degrade so badly that we'd have to turn it off for that segment. The result would be that risk controls could not only be mis-integrated for Agentic Commerce, but they could also miss crucial analytical components even if they are in place.

Does this mean Agentic Commerce is broken by default or that fraudsters find it easier to attack merchants? No; in most cases and flows, the system works well and in a similar capacity to regular payment flows. However, fraudsters only need one target with a sub-optimal environment to cause major damage. A fragmented environment inevitably creates these opportunities.

Agent spoofing creates a new attack surface for fraudsters

AI agents that shop on behalf of a user add a new link in the payment chain. As such, they represent a new target for fraudsters to test, attack, and exploit. This risk manifests in two flavors: the disguise of malicious AI agents as legitimate ones, and taking over someone's AI application to use their legitimate agent for malicious activities.

When it comes to the first case, it's worth noting that a report from Radware's Cyber Threat Intelligence team confirms that some AI agents (namely Claude and Grok) are easily spoofable.[1]

It's safe to assume that even if these vulnerabilities are unpatched today, they may very well be resolved in the near future. But as we know, no system stays 100% secure over time. Our working assumption, especially given the fragmented nature of the ecosystem, is that fraudsters will find new, even if costlier, ways to spoof AI agents. As new AI providers offer Agentic Commerce functionalities, they will continue to present new ways for security issues to manifest.

The extended "spoofable" surface area doesn't end there. With Agentic Commerce, fraudsters can now target the user's AI application for account takeover (ATO) fraud rather than focusing solely on their wallet, merchant, or email accounts. By taking over an account and changing the shipping address or email for digital goods, an attacker is free to commit fraud.

To an extent, this scenario is even more dangerous than the first one. When an agent is spoofed, the merchant still comes in contact with it and can detect the agent is compromised. But when the AI application itself is taken over, merchants do not expect to get much helpful data from the AI providers and must trust them to keep their system secure.

This bears some concerns. *Are AI applications secured to the same standards as wallets or even email accounts? Are they experienced in detecting, mitigating, and securing against evolving, sophisticated ATO attacks? Do they all uphold the same quality in their defenses, or can fraudsters pick on the weak link among them?* One would need to adopt an optimistic worldview to answer all these questions positively.

Regulation limbo increases pressure on merchants

Agentic Commerce receives headlines almost on a daily basis. But there's one party that has been absent from the discussion so far: the regulator. Strangely enough, despite global regulators being quite active in recent years when it comes to fraud, AML, and scam concerns, none have yet clarified their stance on Agentic Commerce.

This matters for a number of reasons. **The first is the AI providers, who are now a new, integrated part of the payment flow, and their efforts to secure it.** Specifically, we expect AI providers to place strong user and card authentication when the user opts-in to the Agentic Commerce feature on the AI app. This process uses third-party vendors, each with their own standards, capabilities, and vulnerabilities.

AI providers, though, are highly incentivized to allow users to transact through their app, especially when they are under no regulatory pressure to mitigate fraud. This isn't just about fraud loss liability, but also about card scheme programs such as VAMP that hold acquirers and merchants accountable to high fraud rates.

Another consideration to keep in mind is that **while merchants are liable for losses, there's no incentive or guarantee for AI providers to share relevant data with them.** A "barely pass" KYC score on the AI application's side should be a critical datapoint for merchants to include in their payment risk assessment, but they are rarely aware of it. Merchants certainly aren't aware if a user has failed to add five different cards to their AI application before they finally managed to verify one.

The risk of ATO is as significant as the threat of stolen identities and cards. Merchants are already feeling squeezed when upstream ATO cases, such as with the user's wallet or bank account, manifest as chargebacks they are liable for. With Agentic Commerce, AI application ATO creates another possible vulnerability that they cannot control or assess, yet are still financially responsible for.

The final issue involves agent mistakes and friendly fraud. While it's hard to forecast how users will end up incorporating Agentic Commerce in their daily lives, it's reasonable to expect that human-in-the-loop payments would still make up a large chunk of overall volumes. However, it's also reasonable to expect that as AI-powered, autonomous payments grow in number, so will cases of agent mistakes.

These mistakes could come in many forms and for many reasons. Maybe the user decides the shipping date is too late, or that they are better off with a different model of the item purchased. Maybe the agent subscribes by mistake to a monthly plan, rather than a one-time purchase. Perhaps it accidentally buys a band's new album instead of getting tickets to their show, or an item localized for the wrong region. And these scenarios haven't even considered cases of hallucinations.

As time goes by, we should expect to see more user complaints regarding "accidental" purchases by AI agents. It's entirely reasonable to believe that even if AI providers and merchants offer flexible return or refund policies, a part of these cases will manifest as chargebacks. That is exactly what we see already today in cases of *buyer remorse*: the user either claims friendly fraud or, in extreme cases, commits chargeback fraud by denying the payment altogether. While regulation doesn't dictate different liability structures, merchants absorb these losses.

03

New fraud for a new commerce era

New fraud for a new commerce era

Keeping in mind the inherent vulnerabilities, it's important to remember that while Agentic Commerce relies on existing financial rails, it has the potential to completely overhaul the shopping experience and its user journey. This new user journey, which brings its own unique vulnerabilities, is in effect widening the surface of attack of eCommerce products by introducing new possible attack vectors.

But the exact appearance of these attack vectors remains to be seen. Agentic Commerce is in its infancy; even if we predict how attacks look initially, the ever-evolving nature of fraud ensures those same attacks look different over time.

Still, it's worth going through this exercise if only to show that these new attack vectors are not limited to a narrow use-case and enable fraudsters to operate in several different ways. By combining our fraud domain expertise and deep understanding of fraudsters' MOs, assessing the new user journeys and their vulnerabilities, and looking at early Agentic Commerce fraud cases, we can draw a reliable picture of what threats are likely to emerge first.

However, we have to keep in mind that fraudsters are a creative adversary and they will find numerous ways to exploit agentic flows, as they invariably do when new technology enters the market. For that reason, we urge you to treat these likely threats not as an exhaustive list, but as a way to inform and shape your own fraud threat assessment.

Third-party fraud

When it comes to third-party fraud, we do not expect fraudsters to re-invent the wheel as the payment method itself stays the same. However, given fraudsters' perception of Agentic Commerce as a new and suboptimally defended flow, we expect fraudsters to heavily invest in testing and exploiting it.

Running stolen cards through AI agents

This attack vector would be designed to bypass AI agent authentication. Fraudsters could use available AI apps to sign up with stolen, synthetic, or mule identities to run stolen cards against eCommerce websites. Attacks could be as simple as running multiple cards against the same account and identity, or as advanced as orchestrating attacks where each card would be matched with a dedicated account and identity.

Presumably, fraudsters would need to go through identity verification on the AI application side, which they wouldn't have to do if they shopped with the merchant directly. But given the data and controls issues discussed in the previous section, **fraudsters might identify specific merchants and stolen goods that would be worth the extra cost on their side.**

Upstream account takeovers

Getting access to an AI application account that has its true owner signed-up for Agentic Commerce features is probably the easiest way in which fraudsters could monetize. They wouldn't need to authenticate anything: the identity, card, or AI agent. Once they get access to the AI app, nothing but a human-in-the-loop 2FA challenge would stop them.

For this reason, we expect AI apps to become prime targets for ATO, credential stuffing, and phishing attacks. This challenges not only AI providers to secure their systems with new standards, but also the victims' own awareness and caution. Victims are more likely to accidentally share credentials for their AI application rather than for their bank account, even if the former has Agentic Commerce permissions.

AI agent spoofing

Despite multiple parties, such as MasterCard, Visa, and Google, publishing their own protocols for Agentic Commerce bot authentication, our working assumption should always be that persistent, sophisticated fraudsters can bypass these defenses. However, we shouldn't discount loopholes found in the early days of Agentic Commerce as AI providers, security teams, and eCommerce platforms all rush to implement new protocols that are unlikely to be standardized quickly.

Spoofing AI agents would enable fraudsters to run stolen cards without the need to authenticate their identity. Unfortunately, it's also safe to assume that there would be little-to-no risk controls monitoring their actions, as verified entities are often given full permissions to act on a platform, despite fraud cases still emerging from this population.

Merchant fraud

One wonders what agentic-optimized eCommerce sites could look like in the future, from UI design to agentic user journeys. Will retailers optimize their shops to cater to both humans and AI agents, or are they going to opt to duplicate their stock over to an agentic-only website? Presumably, as merchants adapt to Agentic Commerce, so too will the fraudsters pretending to be ones. In doing so, they target a new victim persona: AI agents.

Merchant website spoofing

GenAI has caused an explosive growth in merchant website spoofing. It's now easier and faster than ever to copy a legitimate merchant's website and recreate it from scratch with the same look and feel. So far, fraudsters have used such tactics to fool victims into giving away their login credentials for use in ATO attacks. This MO is prevalent in the FI and Fintech space, and to a lesser extent on eCommerce marketplaces as well.

But now, those tactics can be used against AI agents as well for a different purpose. Fraudsters could try to "fool" AI agents into placing purchases on their website in ways that a human would detect right away. While AI agents could be configured to avoid such websites through agentic security protocols that control the domains they visit, the possibility of fraudsters finding creative ways to lure agents into transactions for non-existent goods is still concerning.

Merchant scams

Another way for fraudsters to take advantage of AI agents is by managing to obfuscate the exact nature of what they're selling, exactly as they do with human users. Examples include hidden subscription payments, misrepresented products such as a picture of an iPhone instead of a real one, and counterfeit goods.

It is to be expected that fraudsters will test and optimize techniques to fool non-human-in-the-loop AI agents. This may involve aggressive pricing strategies, specific user journey flows, or textual manipulation of website content. Unless AI shopping agents are trained specifically to detect and avoid such scams, an unlikely development in the near term, it's reasonable to expect them to perform at or even below human levels in this regard.

First-party fraud

First-party fraud is the largest and fastest growing loss driver for eCommerce merchants today. Will Agentic Commerce make this even worse? Unless you include "honest" bot mistakes, the answer is not necessarily. However, there are a couple of typologies we think are worth looking at more closely.

Friendly fraud

We discussed in a previous section how agent mistakes will manifest naturally and introduce friction in Agentic Commerce. It's reasonable to expect users taking advantage of this phenomena to initiate chargebacks or disputes on payments placed via AI agents. Buyer's remorse now has a new outlet: the claim that the agent acted without the user's full consent.

Furthermore, instances of true friendly fraud will likely become more prevalent. Given the amount of digital freedom permitted nowadays, young children could likely access their parents' desktop AI applications much more easily than their bank account or physical card.

Policy abuse

With AI agents, it'll be easier than ever for fraudsters to buy goods at scale and in real-time. Arguably, this capability is not new, but it will become cheaper and easier to bypass traditional bot detection methods.

Fraudsters will be able to set up dropshipping shops at scale with ease and reopen them quickly when a marketplace closes them down. Similarly, promo abuse will also become easier and cheaper to commit. Bot scalping would become much harder to stop as both legitimate and malicious AI shopping agents race to be the first ones to purchase event tickets or grab a limited edition fashion drop.

04

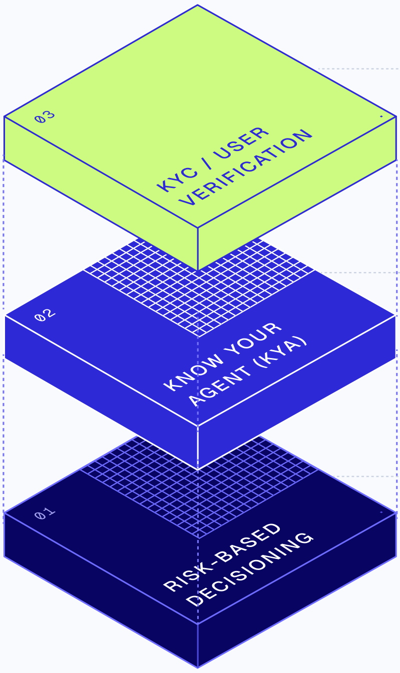
Rethinking fraud prevention for agentic commerce

Rethinking fraud prevention for agentic commerce

Despite the novelty of Agentic Commerce, the underlying payment infrastructure remains unchanged. Transactions still run on existing card rails, authorization decisions are still made by issuers, and liability still sits largely with merchants. From a distance, this continuity creates the impression that existing fraud defenses can be extended rather than rethought.

In practice, however, Agentic Commerce alters the structure of the transaction flow in ways that materially weaken traditional defensive assumptions. Here are a few of the gaps, and the fraud prevention strategies designed to mitigate them.

A layered defense for Agentic Commerce



03

Purpose
- Confirm the user's identity
- Link the user to a payment method

Limitations
- Cannot detect compromised AI application
- Does not verify intent at the moment of purchase

02

Purpose
- Verify that the AI agent is legitimate
- Confirm the agent is authorized to act for a user

Limitations
- Cannot prevent fraud using stolen cards
- Authorized agents can still act maliciously

01

Purpose
- Detect suspicious transactions
- Identify fraud patterns and abuse

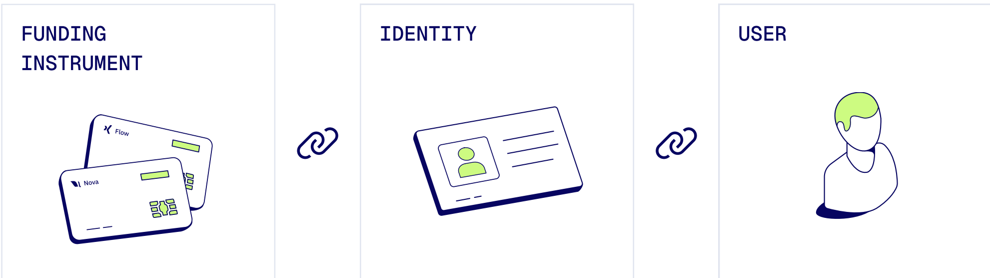
Limitations
- Depends on data quality
- Agentic flows may degrade key signals

Layer 1: KYC and user verification

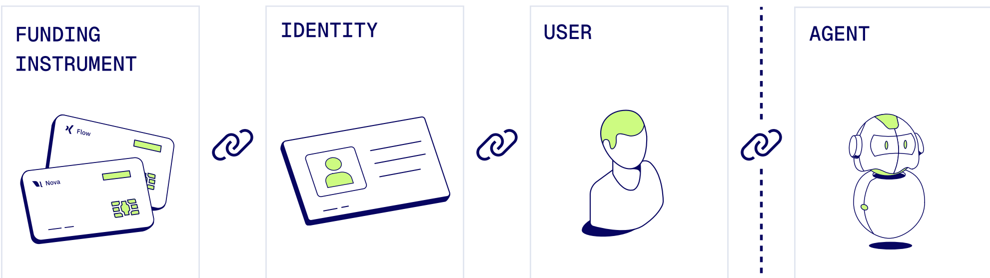
Verification, which has historically been the foundation of digital fraud prevention, will become distributed across more actors, executed at different times, and informed by partial and often degraded data. None of this will render verification obsolete, but it will mean that it can no longer serve as a complete or self-sufficient defense.

The defensive challenge then shifts from proving identity at a single point in time to managing risk across a fragmented, multi-actor system whose components do not share the same visibility or incentives. Therefore, understanding how fraud prevention must evolve will require a prior understanding of how the verification chain itself has fundamentally changed.

In traditional eCommerce, verification follows a relatively stable and well-understood structure. A human user establishes an identity, binds that identity to a funding instrument, and demonstrates control over both. Whether through explicit authentication challenges, passive behavioral analysis, or historical trust, the system is designed to answer a small set of core questions: who is acting, are they authorized to use this card, and does their behavior resemble legitimate intent?



Agentic Commerce, on the other hand, introduces a new actor to this chain: the AI agent that acts on the user’s behalf. This agent may operate within a browser, through an MCP, and potentially across multiple merchants and sessions, often without direct human involvement at the moment of purchase. As a result, the verification chain effectively extends like this:



This additional link has outsized defensive implications. While AI providers authenticate users and provision agents, merchants interact primarily with the agent rather than the human, and issuers continue to make authorization decisions based on card-level signals that may no longer cleanly reflect user intent. Each participant verifies a portion of the chain, but no participant verifies it in full.

Crucially, trust becomes transitive while liability does not. Merchants are expected to accept agentic transactions initiated by users they did not authenticate, based on identities they cannot directly assess, while remaining fully liable for downstream fraud and disputes. This asymmetry does not break verification outright, but it does change where and how it fails. Weaknesses are less likely to emerge inside individual systems than at the boundaries between actors, flows, and implementations.

This limitation becomes more acute as identity itself becomes easier to manufacture or compromise. GenAI significantly lowers the cost of creating synthetic or deepfaked identities capable of passing basic onboarding checks. Even when initial verification is strong, AI apps introduce a new upstream ATO vector: compromising the AI application can grant delegated purchasing authority without triggering any merchant-side identity failure. In these cases, verification succeeds exactly as designed and still permits fraud to occur.

For this reason, consider re-verification in every agentic flow session and not just as a one-time formality. In practice, this means lowering the threshold for step-up authentication that demonstrates funding instrument ownership, particularly mechanisms such as 3DS. The objective is twofold: to increase confidence that the cardholder is present and consenting, and to blunt the impact of upstream account takeovers that merchants cannot otherwise detect.

However, this measure is inherently expensive. Introducing human-in-the-loop authentication increases friction, and authentication flows like 3DS are known to turn shoppers away. Each of these effects independently reduces conversion and, when aggregated, can be prohibitive if applied bluntly.

As a result, effective deployment of such risk controls depends less on whether step-up verification is available in-flow, and more on how flexibly it can be applied. Highly configurable controls, whether they be adjustable by flow, merchant segment, risk signal, or learning phase, allow teams to observe emerging fraud patterns and respond without stunting growth. In an environment this fluid, being forced to choose between "always on" and "always off" is itself a structural risk. Instead, merchants should prepare to adjust their risk appetite in real time.

Layer 2: Introducing Know Your Agent (KYA)

If merchants and issuers cannot directly verify the full user chain, the minimum requirement becomes understanding who (or what) is acting on the user's behalf. Know Your Agent (KYA) emerges from this need.

The KYA layer has four main responsibilities:

- **Identify that the user is in fact a bot**
- **Verify that the bot is authorized to act on behalf of a verified user**
- **Detect any agent spoofing attempts**
- **Verify the bot is acting within its declared scope of authority**

While this is sound in theory, the current ecosystem's fragmentation makes standardizing implementation difficult. In the previously mentioned Radware report, it's clear that each of the five leading AI providers chose to implement their own security and agent verification protocols. Even if they were to all be highly secured (which is not currently the case), it is practically impossible for PSPs and merchants to maintain pace with disparate standards, nuances, and update schedules. Consequently, merchants and their payment providers will be hard pressed to keep their KYA layer updated and secure.

For that reason, Visa and Mastercard both unveiled their respective Agentic Commerce protocols last year. Both protocols will implement **Web Bot Auth**, an authentication standard developed by the Internet Engineering Task Force (IETF).[2] At a high level, Web Bot Auth allows an AI agent to cryptographically prove that it is a legitimate agent issued by a known provider and authorized to act on behalf of a user. Rather than inferring automation from behavior or traffic patterns, merchants will receive an explicit, tamper-resistant signal that the request originates from a declared agent.

This cryptographic verification makes it significantly harder for adversarial actors to spoof legitimate agents and, more importantly, will introduce a standardized way to identify and verify agents across providers and platforms. However, with the earliest version of Web Bot Auth expected in April 2026, it's unlikely that we will see meaningful adoption across the ecosystem before the end of the year.

The good news is that both large schemes are introducing similar protocols based on the same standard. This means that while it will take some time, the industry is on a set course for eliminating today's core vulnerability.

At the same time, traditional bot detection will not suddenly become obsolete in an agentic world. On the contrary, it will become even more critical as AI-powered attackers gain the ability to closely imitate human interaction patterns with unprecedented precision. Advanced bot attacks can now deliberately avoid declaring themselves as agents, blending into legacy flows designed for human users and bypassing agent-specific verification entirely.

In practice, this means that bot detection and KYA are complementary rather than substitutive. KYA governs declared agents, while bot detection remains critical for identifying undeclared automation infiltrating traditional flows.

However, even when an agent explicitly identifies itself and passes KYA, transaction safety is not assured. As outlined earlier, authentication establishes the legitimacy of the actor, not the legitimacy of intent. Authorized agents can execute fraud using stolen payment credentials, operate through compromised AI application accounts, or exploit merchant policies in ways that remain valid at the protocol level. In these scenarios, the agent is functioning exactly within its declared authority, but not toward a legitimate outcome.

This means that KYA reduces agent spoofing risk and raises the cost of certain attacks, but it does not prevent abuse by authorized entities. Trust without ongoing assessment creates blind spots, particularly in systems where agents can act quickly, repeatedly, and asynchronously.

The result is that KYA, like traditional identity verification, cannot act as a fully independent defense mechanism. The most crucial piece remains the ability to make risk-based decisions grounded in assessment of user intent.

Layer 3: Risk-based decisioning in the age of agents

When verification methods reach their limit, organizations must turn to risk-based decision making. By assessing the user's intent and the financial exposure they bring, eCommerce merchants and financial institutions can decide whether to accept a payment without verifying the user. This has always been at the heart of fraud management, and it will not change with the rise of Agentic Commerce.

Companies achieve this through manual reviews, expert-written rules, machine learning models, and increasingly, with AI agents.

The challenge is that these methods rely on a single prerequisite to perform well: data integrity. However, significant data integrity issues will become a hallmark of agentic commerce flows, affecting both in-browser and MCP-based architectures.

Segmenting Agentic Commerce

Agentic Commerce flows will still rely partially on the existing data layer, provided they utilize dedicated fraud scoring and strategy layers trained on similarly constrained data. This segmented flow should be executed in parallel with, yet distinctly from, the standard eCommerce payments flow. Fraud teams will need to calibrate this decisioning workflow to account for the known data integrity issues, ensuring performance remains stable and optimized for the specific requirements of agentic payments.

Segmentation is also critical for monitoring. Teams should expect erratic behavioral patterns as users adopt technologies based on new foundational LLM releases, platform partnerships, and marketing campaigns. These spikes in volume will have the potential to appear almost overnight. By isolating agentic traffic, organizations will ensure that the volatility of early-stage adoption does not contaminate the baseline performance metrics of their core business.

Similarly, fraud will continue to behave with its inherent erraticism. The main danger will lie in failing to notice an emerging threat in what may initially be a small portion of total traffic. This is particularly critical in cases where fraud exploits a specific and novel defense loophole. Without dedicated monitoring and alerting for these segments, organizations will likely dismiss such activity as statistical noise until the damage is already irreparable.

Beyond classification, organizations must address the challenge of intent: how to distinguish the objectives of legitimate AI agents from those of malicious ones. This represents a whole new domain of which we have little data and even less practical experience. This knowledge gap will become even more pronounced as future behavior will have differed dramatically from current observations. To date, Agentic Commerce has been shaped primarily by early adopters, power users, and MCP-based flows, but these precedents will offer little guidance for the broader ecosystem to come.

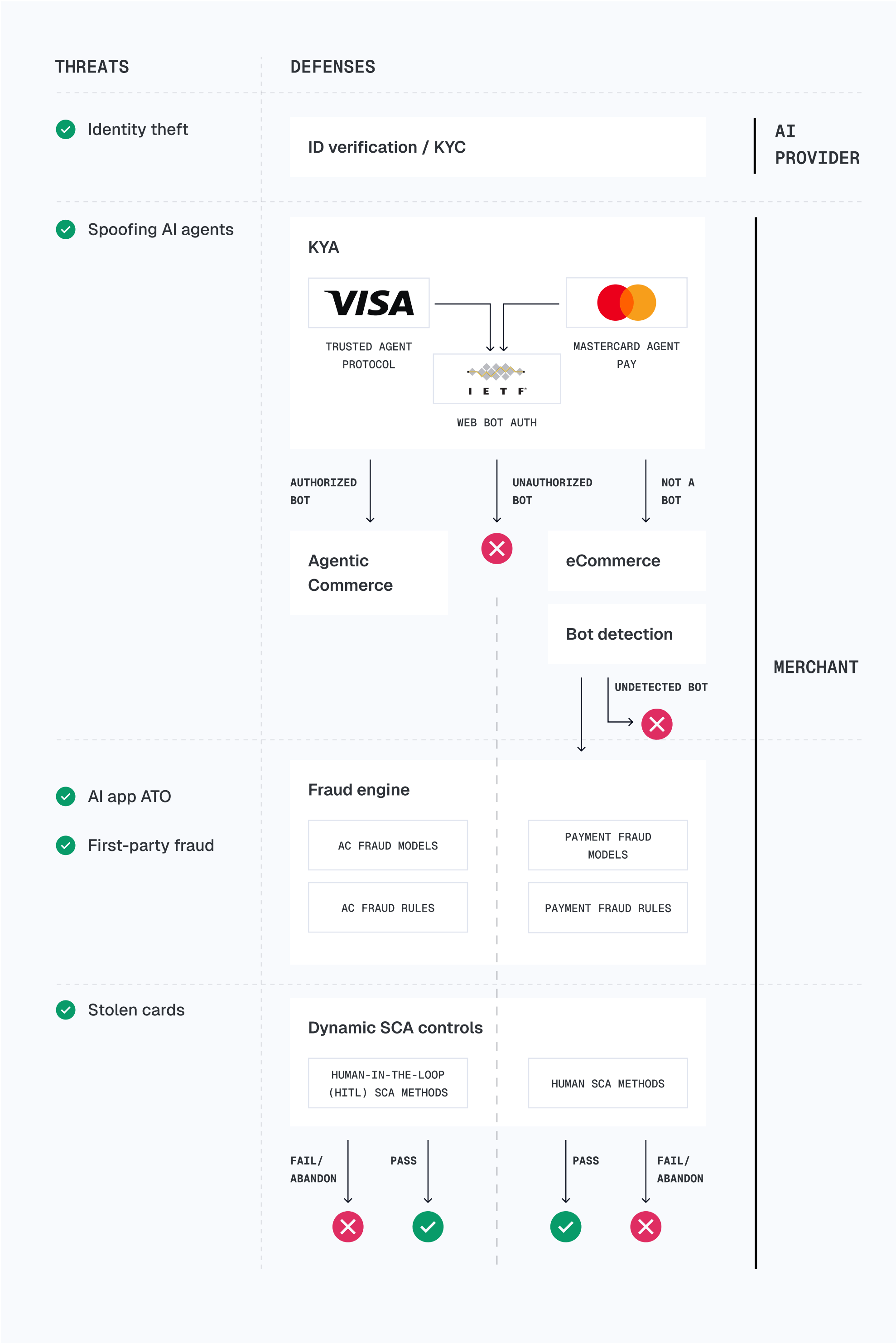
> Given the assumption that we in the industry have entered uncharted waters, monitoring Agentic Commerce separately and closely will be the only way to maintain visibility into a shifting threat landscape. Fraud teams will need to learn on the fly how to define legitimate versus malicious intent typologies and how to reliably connect those behaviors to the established customer identity.

Industry-specific considerations

Different actors along the transaction journey prioritize specific risks, maintain disparate levels of data visibility, and focus on distinct defensive measures. Here's what those focus areas look like for merchants, acquirers, PSPs, and issuers.

Merchant fraud strategy for Agentic Commerce

Merchants should prepare to thwart novel third-party and first-party fraud vectors that will have emerged in their Agentic Commerce flows. Fortunately, they will be able to do so with a relatively high data integrity and tool availability compared to others in the ecosystem. The diagram below illustrates a fundamental design for integrating the defensive concepts outlined in this section.


THREATS
DEFENSES
Identity theft
ID verification / KYC
AI PROVIDER
Spoofing AI agents
KYA
VISA
TRUSTED AGENT PROTOCOL
MASTERCARD AGENT PAY
IETF
WEB BOT AUTH
AUTHORIZED BOT
UNAUTHORIZED BOT
NOT A BOT
Agentic Commerce
eCommerce
Bot detection
UNDETECTED BOT
MERCHANT
AI app ATO
First-party fraud
Fraud engine
AC FRAUD MODELS
PAYMENT FRAUD MODELS
AC FRAUD RULES
PAYMENT FRAUD RULES
Stolen cards
Dynamic SCA controls
HUMAN-IN-THE-LOOP (HITL) SCA METHODS
HUMAN SCA METHODS
FAIL/ ABANDON
PASS
PASS
FAIL/ ABANDON

Merchants Agentic Commerce preparedness check list:

- [ ] **Identify and segment agentic traffic:** Establish a reliable way to label agentic vs. human-initiated transactions and route agentic flows through separate monitoring and scoring pipelines.

- [ ] **Recalibrate verification:** Reassess 3DS and step-up strategy for agentic flows to avoid blanket enforcement. Teams should be ready to adjust step-up thresholds for agentic sessions.

- [ ] **Know Your Agent (KYA) readiness:** Verify that payment service providers have implemented the latest Agentic Commerce protocols from Visa and Mastercard. At the same time, maintain traditional bot detection to identify and manage undeclared automation.

- [ ] **Harden against upstream ATO:** Operate under the assumption that AI application accounts can and will be compromised. Tighten controls around shipping address changes, digital goods delivery, and velocity rules for agent-linked accounts.

- [ ] **Monitor first-party fraud drift:** Track dispute reason codes tied to agent errors. Merchants should update their Terms & Conditions to address delegation claims and mitigate friendly fraud.

- [ ] **Stress-test your stack:** Audit data pipelines for both MCP and in-browser agentic flows. Ensure all fraud vendors, rules, and machine learning models fire correctly within agentic segments.

Acquirer and PSP fraud strategy for Agentic Commerce

While acquirers and payment service providers (PSPs) should enable their merchant portfolios to accept Agentic Commerce payments using the most up-to-date scheme protocols, their core risk management interests lie elsewhere. These companies should focus on preventing merchant fraud broadly, and specifically on attacks that will have exploited Agentic Commerce as a primary vector.

Fortunately, payment companies don't necessarily need to implement new tools or technologies to achieve this. Instead, simply monitoring the share of volume attributed to Agentic Commerce for each merchant in a portfolio will surface merchants that focus heavily on that flow. When the ratio is exceptionally high in favor of agentic payments, a review of those merchants will likely capture the majority of merchant fraud cases.

PSP and Acquirer Agentic Commerce preparedness check list:

- [ ] **Track merchant agentic exposure:** Establish a baseline for agentic volume across the merchant portfolio. Organizations will have successfully identified and flagged outliers for enhanced review when agentic share deviates significantly from industry peers.

- [ ] **Strengthen onboarding and KYB due diligence:** Incorporate agentic strategy assessment into merchant onboarding. Evaluate business models likely to concentrate agentic traffic and assess whether associated fraud controls are sufficient for autonomous flows.

- [ ] **Detect merchant-side abuse:** Identify merchants disproportionately attracting agentic traffic and monitor for patterns indicative of scams. This includes flagging high dispute rates combined with an exceptionally high agentic volume share.

- [ ] **Incorporate scheme agentic protocols:** Implement Visa and Mastercard agentic authentication protocols within acquiring and processing infrastructure. This ensures proper validation and routing in compliance with updated network requirements.

Issuer fraud strategy for Agentic Commerce

While issuers do not appear to be directly liable to new losses derived from agentic payments, they occupy the most challenging position for tracking and management. The primary threat going forward will be a gradual erosion of risk models as AI agent behaviors are mistakenly attributed to human users. This could adversely affect acceptance rates and eventually degrade customer satisfaction and trust.

While there are currently no signs that the networks will clearly tag payments as Agentic Commerce, it does seem that schemes plan to share specific data insights with issuers. Whether these insights will suffice to correctly segment this flow, label it accurately, and protect risk models remains an open question.

Given this landscape, issuers should do their best to increase internal awareness of the topic, study data received from schemes as it becomes available, and diligently work toward a reliable Agentic Commerce payment label.

With that in place, issuers should deploy dedicated models and rules to control the risk in this segment, accounting for its particular data limitations and user behaviors. Specifically, issuers should segment labeled data originating from agentic actions separately from human-initiated data to mitigate the risk of adverse feedback loops.

Issuer Agentic Commerce preparedness check list:

- ☐ **Build an agentic transaction label:** Work with schemes to capture agent-initiated identifiers and establish reliable internal tagging. Organizations should integrate these signals into their ledgers to ensure clear visibility.

- ☐ **Segment models immediately:** Separate agentic-labeled data from human behavioral data to prevent the contamination of core authorization models. This ensures agent velocity will not have skewed the risk profiles of human cardholders.

- ☐ **Re-evaluate OTP / 2FA logic:** Analyze 3DS failure patterns caused by asynchronous agent behavior and monitor abandonment-induced risk score inflation.

- ☐ **Monitor profile degradation:** Track customers whose risk scores rise due to agent retries or failed challenges. Implement safeguards to prevent these anomalies from triggering downstream false positives in human-initiated transactions.

Agentic Commerce will not arrive all at once

It will expand gradually, through new integrations, shifting user behavior, and evolving protocols. But as with previous shifts in digital payments, the underlying dynamics will change before the volume does.

Fraud teams that treat agentic transactions as a small edge case may find themselves reacting too late. The signals they rely on are already changing, the actors in the transaction chain are multiplying, and the boundaries of responsibility are becoming less clear.

The challenge ahead goes beyond defending against new fraud typologies. Fraud prevention systems will need to operate in conditions where automation is legitimate, verification happens across multiple actors, and key signals may no longer be consistently visible.

In that environment, resilience will depend less on any single control and more on how quickly organizations can observe, adapt, and recalibrate their defenses as the ecosystem evolves.

References

1. **Radware: The AI Identity Dilemma: Malicious Bots in Disguise**
https://www.radware.com/security/threat-advisories-and-attack-reports/the-ai-identity-dilemma-malicious-bots-in-disguise/

2. **IETF: Web Bot Authentication Working Group**
https://datatracker.ietf.org/wg/webbotauth/about/

About Sardine

Sardine is the leading AI risk platform for fraud prevention, compliance, and credit underwriting, trusted by global enterprises in over 70 countries. Using device intelligence, behavior biometrics, AI and machine learning, Sardine stops fraud in real time, streamlines compliance, and unifies data across risk teams. Backed by world-class investors including Andreessen Horowitz, Activant, FIS, Visa, Experian, Moody's, and Geodesic Capital, Sardine is redefining risk management for the real-time economy.

Sardine

SARDINE.AI