EPISODE 28 | Forbes Analysis

Your Fraud Stack Was Built For Humans: Now What?

JULY 20, 2026

This transcript was auto-generated and may contain errors or inaccuracies.

Jeff Welcome to "What the F Happened? Fraud and Financial Crime Deconstructed," a DEFEND podcast where we break down what's actually happening across fraud scams, AML and financial crime.

Emily Each episode cuts through the noise to explain the tactics, trends, and real world impact behind the headlines so you're better prepared for what comes next. Let's get into it.

Emily So what actually happens to global banking when, um, like your digital identity is suddenly duplicated 10,000 times in a single afternoon?

Jeff Oh, it's a nightmare.

Emily Right? And by, you know, invisible assistants you didn't even know you hired.

Jeff Yeah. The entire financial security infrastructure, we spent, what, the last three decades building it just instantly fractures. We're literally entering a reality where the foundational math of cybersecurity just it simply stops working.

Emily It's wild. If you are someone who tracks the sort of, um, invisible architecture governing our digital lives, this is definitely an existential shift you need to have on your radar.

Jeff Absolutely.

Emily Today, we are analyzing a July 2026 framework published by Yinglian Xie

Jeff Our legacy financial fraud stack, I mean, it was built to verify human beings. Yes. But AI is actively, systematically breaking that model, like right now. And to understand the sheer scale of this fracture, we really have to look at the baseline assumption of legacy fraud detection.

Emily The old way of doing.

Jeff Things. Exactly. For years, the industry relied on this super rigid sequential logic gate. First, the system asks, is there a biological human behind this keyboard?

Emily Like, are you real?

Jeff Right. You pass that check and it moves to the second heavier gate, which is, is this human the actual account holder or a bad actor using stolen credentials?

Emily So it's basically check for humanity, then verify identity. It's exactly. Okay, let's unpack this. Because if the old system relied entirely on verifying a fixed human identity, we need to understand exactly why AI agents make that old math just, you know, totally obsolete. Yeah, because we are moving way past simple chat bots here.

Jeff Oh, way past. We're talking about agentic AI, right?

Emily Software systems that are given a goal, a budget, and like the actual autonomy to string together multiple actions across the internet without asking for permission at every single step.

Jeff And what's really fascinating here is how that autonomy destroys the mathematical constraints that these fraud engines rely on.

Emily How so?

Jeff Well, she outlines this escalation as basically a numbers game. Historically, verifying human behavior was computationally manageable because, well, humanity has a hard population cap, right?

Emily There are only so many of us.

Jeff Exactly. You design a system to monitor a finite number of people accessing a network, even as individuals accumulated multiple digital identities over the last decade.

Emily Like different email addresses or, you know, various social profiles.

Jeff Right? But the behavioral patterns were still tethered to a single slow moving biological entity, right. With, you know, limited hours in the day.

Emily Meaning one person can only type so fast or like click through so many web pages before they literally just need to go to sleep.

Jeff Exactly. And the fraud engine looks for anomalies in that biological speed limit, but a genetic AI removes the speed limit entirely, and it introduces this insane exponential multiplication effect.

Emily Which shatters the whole one to one relationship between a verify user and an action.

Jeff Completely shatters it. We're transitioning from a model of finite human actors to an infinite programmable volume of digital proxies.

Emily Okay, let me try to visualize this for a second. It's like, um, imagine you suddenly hired ten different personal assistants in a single day.

Jeff Okay.

Emily Yeah. And you just gave them all your wallet, right? Your passwords, a list of errands to run across town, and they are doing it all simultaneously.

Jeff A logistical nightmare.

Emily Totally. Like, one is paying your electric bill, another is rebalancing your four hundred one K. A third is, I don't know, bidding on concert tickets. And then by dinner time, you just fire all of them.

Jeff Just let them all go.

Emily Yeah. And then the next morning, you hire twenty brand new assistants to do entirely different tasks. Like how does a legacy banking system even begin to authenticate that kind of constant chaotic churn?

Jeff I mean, it doesn't that's the core vulnerability. She is highlighting the bank's fraud engine. It's looking for a stable device fingerprint.

Emily A recognized IP address, right.

Jeff A familiar typing cadence, but instead it gets hit by twenty ephemeral cloud based agents. And these agents, they're spun up via API. They execute a highly specific microtask in literally milliseconds, and then they're retired.

Emily So the identity is essentially self-destructing after every transaction.

Jeff Precisely. Which destroys the foundational assumption of our current authentication infrastructure. It was built entirely around that one to one relationship.

Emily That makes the whole concept of a login session completely meaningless.

Jeff Oh, completely.

Emily Because if my temporary bot is routing through an AWS server in Ohio to pay a bill while another bot I authorize is using like a virtual machine in London to buy a software license. The bank's system is just going to flag both as massive anomalies.

Jeff And that forces the bank into a terrible position because they either block everything, which totally cripples the user.

Emily Everyone gets mad.

Jeff Exactly. Or they loosen their security parameters to accommodate all these agents, which basically leaves the vault door wide open for actual fraud.

Emily Wow. So when that one to one identity rule breaks down, the security checks built around that rule just become massive liabilities.

Jeff Yeah, you essentially have a moving target.

Emily So if we can't definitively lock down who is making the request because you know, the who is just a temporary piece of code acting on my behalf, right? Then that implies we have to like, abandon identity as the primary shield, right. And shift the entire architecture toward authorization.

Jeff And that is the absolute crux of the paradigm shift Datavisor is advocating for here. Okay. Because the industry's initial knee jerk reaction has been to try and force agents to declare themselves as agents.

Emily Right? Like wearing a little digital nametag.

Jeff Exactly. We're seeing proposals for frameworks where AI operates with a sort of digital license plate. But knowing an agent is an agent, it only addresses a tiny fraction of the problem.

Emily Because it seems like that just replaces one broken identity check with another.

Jeff Pretty much.

Emily Yeah, like if a bot announces it's a bot, the bank still doesn't know if I actually gave it permission to, you know, drain my savings account.

Jeff That is the vital distinction. The new ultimate question replacing the legacy model is no longer asking who is making the request, but rather, is this specific request legitimate and authorized.

Emily And what are its absolute limits?

Jeff Exactly. The identity of the actor matters far less than the cryptographic proof of authorization.

Emily Okay, here's where it gets really interesting because she introduces this architectural concept called the bounding box.

Jeff Yes, the bounding box.

Emily Right. So if authorization is just the digital key that gets the AI through the front door of your bank account, the bounding box represents like the rules of the house. Once it's inside.

Jeff That's a great way to put it. It dictates which rooms you can enter, what you can touch, and when. You absolutely must ask for permission.

Emily So it's a shift from trusting the entity to constraining the entity.

Jeff Exactly. Without a bounding box, saying yes to an AI agent is equivalent to handing a stranger a blank check, and just hoping they have a strong moral compass.

Emily And hope is definitely not a viable cybersecurity strategy.

Jeff Definitely not. The bounding box has to exist before the agent ever takes a single action, because once an autonomous system starts executing, human reaction, times are just way too slow to intervene.

Emily Okay. And the text actually provides a perfect, honestly terrifying, real world example of what happens when that bounding box is missing.

Jeff Oh, I know the one you're talking about.

Emily Yeah. There's this cautionary tale about a senior AI research director at meta. Right. And this is someone who literally specializes in AI safety and alignment. Like she understands the architecture better than almost anyone on the planet.

Jeff She knows what she's doing exactly.

Emily But she deployed her own personal AI agent to clean up her emails, and it proceeded to permanently delete her entire inbox.

Jeff It's such a striking anecdote, isn't it? It just completely dismantles the illusion of user control.

Emily I was reading this and trying to figure out like, why couldn't she just click cancel, right?

Jeff Because normally you just stop it.

Emily Yeah. If I'm deleting files on my desktop and I realize I made a huge mistake, I hit a button and the little progress bar stops. But with this agent, she had given it the key to the door via an API integration.

Jeff And there was no bounding box.

Emily Right? No hard coded limit on how many emails it could delete per minute, or like a requirement to batch them for final human review. It just entered this autonomous execution loop.

Jeff And the speed of execution is the critical factor there. Because a human user interacts with a graphical interface, we click. We wait for a visual confirmation. We click again.

Emily It's slow.

Jeff Super slow. An AI agent bypasses the interface entirely. It talks directly to the server so it can queue up and execute ten thousand deletion commands in the time it takes a human to even realize a mistake is happening.

Emily And she literally had to physically unplug the machine, running the agent to sever the connection. Because the software had no interrupt handler she could trigger.

Jeff The commands were just firing too fast.

Emily It's insane.

Jeff And if we connect this to the bigger picture, particularly the global financial networks, that Datavisor protects the stakes scale exponentially.

Emily Oh, absolutely.

Jeff I mean, a deleted inbox is a severe inconvenience, but apply that exact same lack of a bounding box, that same autonomous loop to a wealth management portfolio.

Emily Yeah. You can't unplug a wire transfer once it clears the automated clearing house.

Jeff No you cannot.

Emily If a rogue agent decides to liquidate a stock position and initiate a bulk transfer of funds to an offshore routing number, you can't just pull the power cord.

Jeff Because the architecture of global finance operates in milliseconds. You can't afford to play catch up after the transaction has been batched and verified. The bounding box has to be airtight, predefined, and enforced before the agent is ever deployed.

Emily Okay, so I understand why bounding boxes are absolutely non-negotiable. We have to mathematically restrict the agent's environment, right? But, you know, technology fails. Code has bugs. APIs get updated and break things. And obviously malicious actors actively look for ways to exploit permission scopes.

Jeff They always do.

Emily So what is the mechanism for stopping an agent that manages to somehow slip outside its bounding box?

Jeff Well, this raises a really important point, and it highlights a massive unresolved vulnerability identified in the framework.

Emily Which.

Jeff Is the complete absence of a global kill switch.

Emily Oh, right. Because if someone steals my physical credit card or I notice a fraudulent charge, the remediation is super simple.

Jeff You call the bank.

Emily Yeah, call the bank. They cancel the card number. The physical plastic in the thief's hand becomes totally useless across the entire global payment network, like simultaneously. Right. But AI agents aren't physical cards and they aren't tied to a single database.

Jeff No, they are decentralized software logic, and that creates an absolute nightmare scenario for threat mitigation.

Emily I can imagine.

Jeff Think about a scenario where a user's primary bank detects a rogue agent attempting a transaction that violates its bounding box.

Emily Okay, like trying to send funds to a known malicious wallet.

Jeff Exactly. The bank successfully blocks the transaction and revokes the agent's access token.

Emily So the bank did its job.

Jeff They did. But the agent itself isn't dead.

Emily Wait, what do you mean?

Jeff Exactly the opposite. The agent's core logic is still running on a remote cloud server somewhere. It still has its overarching objective. Oh, wow. And because there's no shared universal protocol to neutralize that specific piece of code, the agent simply pivots.

Emily It just changes direction, right?

Jeff It analyzes the failure, switches tactics, and starts pinging a different financial institution where the user might hold a secondary account, or it targets a cryptocurrency exchange where the user linked their wallet.

Emily So it just walks down the virtual street, basically jiggling every door handle until one opens. Yeah, because a first bank has literally no mechanism to warn. The second bank about this specific autonomous actor.

Jeff The threat persists until the agent's internal budget runs out, or its original server's just taken offline. And this leads us directly into a massive liability void that the industry is just wholly unprepared for.

Emily Who is liable? Like when the damage is eventually done. Who absorbs the financial loss?

Jeff That is the multi-million dollar regulatory nightmare.

Emily Because is it the user? Are they liable because they originally granted the AI permission? Even if the AI hallucinated the malicious action, maybe.

Jeff Or is it the third party developer who coded the agent? Are they responsible for shipping a faulty bounding box?

Emily Or does the liability fall in the receiving bank for failing to detect that the inbound request was an agent operating outside its parameters?

Jeff The legal frameworks are incredibly murky right now. Yeah. We are dealing with a severe jurisdictional mismatch.

Emily Yeah. Because historically, fraud regulation and consumer protection laws have been strictly regional. Right? They operate within state lines or national borders.

Jeff But cloud based agents are inherently borderless. So you could have a user in California authorizing an agent developed by a startup in Berlin.

Emily Running on servers in Singapore.

Jeff Exactly. Which then attempts a fraudulent transaction in London.

Emily So we have geographically bound regulators trying to police these borderless, hyper speed digital entities.

Jeff And she emphasizes that the financial sector has a really fleeting window of opportunity here. They need to establish global cross-border protocols for agent identification and termination before a major incident forces regulators to pass really draconian legislation.

Emily Okay, wait, I got to push back on something here, though.

Jeff Sure.

Emily Go ahead. Because this brings up a proposed solution in the text that just seems fundamentally flawed to me.

Jeff The self-reporting mandate.

Emily Yes, the article suggests that agents should be required to identify themselves to create an audit trail. Right. But if I'm a sophisticated bad actor coding a bot specifically to drain accounts. I'm just going to hard code it to lie. Of course, I'll program my script to check the box that says I'm a biological human. Isn't a self-reporting rule totally pointless? Why build security frameworks based on an honor system for criminals?

Jeff It's a completely valid critique. And G actually anticipates this exact friction in the framework. Bad actors will inherently ignore self-reporting mandates, but the argument is that the financial industry cannot afford to wait for a globally enforceable, airtight legal mandate before taking architectural action.

Emily So what's the point, then?

Jeff You have to look at how market standards practically evolve. We use the precedent set by Know your Customer regulations or KYC. Long before KYC was a strict, legally mandated requirement enforced by governments. It began as a market standard adopted by forward thinking institutions. Right. Banks started requiring identity verification simply because it reduced their own risk exposure and just made good business sense. They set the standard internally.

Emily So they didn't wait for lawmakers to debate the nuances. They just started denying service to unknown entities.

Jeff Exactly. And the same logic applies to mitigating AI fraud. The private sector doesn't need to wait for a unified global governing body to draft legislation about AI self-reporting. I see if the major financial networks and tech platforms collectively decide that any digital entity pretending to be human is an automatic red flag, they can block the fraud before regulators even get involved.

Emily Oh wait, I see the mechanism now. It's not about trusting the bad guys to tell the truth.

Jeff No, not at all.

Emily It's about creating a tripwire. Yeah. The standard is agents must announce themselves. Therefore, if your detection systems catch an entity acting exactly like an agent. But it is declaring itself as a human.

Jeff You don't even need to waste compute power. Analyzing what transaction is trying to execute.

Emily The lie itself becomes the trigger to terminate the connection.

Jeff Exactly. You shift the burden of proof entirely, and the technology to catch that lie already exists. Behavioral biometrics legacy systems might struggle with the speed of agents, but modern telemetry can easily detect the absence of humanity, right?

Emily Because an AI doesn't use a mouse the way I do know, if the fraud engine tracks the cursor movement and sees a perfectly linear vector moving at a constant speed with zero hesitation before clicking submit.

Jeff It knows immediately that it's a machine. Or if the telemetry shows a form being filled out in two milliseconds via a hidden API call rather than simulated keystrokes. Wow. If the entity executing those hyper efficient non-biological actions has checked the inhuman box, the market standard dictates an immediate block.

Emily So the private sector neutralizes the threat by making the deception itself the punishable offense.

Jeff Exactly.

Emily So what does this all mean for you, the listener, as you navigate this rapidly shifting digital architecture?

Jeff It's a lot to take in.

Emily It really is. I mean, the most immediate takeaway is that the era where security meant simply proving you have a pulse that's just over.

Jeff Long over.

Emily As you start using Agentic AI to manage your life and your finances. The focus has to shift entirely to setting strict bounding boxes around what those agents are authorized to do. It's no longer just about generating a secure password. No, it is about demanding mathematically enforced constraints from the developers who build these tools.

Jeff And, you know, the architecture is adapting, but it leaves us with a rather profound paradox to consider as we wrap up.

Emily What's that?

Jeff Well, she's core thesis is that the legacy fraud stack was built to protect humans from humans, but we are rapidly moving toward a reality where financial networks will treat any entity that acts to efficiently or navigates a system to perfectly as a massive security threat.

Emily Wait, so which means if I log into my bank and process my transfers too quickly, or, I don't know, use keyboard shortcuts instead of casually browsing with a mouse.

Jeff The system might flag you as an undeclared bot.

Emily Oh my gosh.

Jeff Yeah. So the ultimate thought to mull over is this. As these behavioral detection systems become hyper calibrated to root out hidden AI, will we eventually reach a bizarre future where everyday people have to intentionally act more like machines?

Emily Wait, you mean acting less efficient?

Jeff Yes. Will you find yourself purposely slowing down your typing, hovering your mouse aimlessly, and manufacturing human inefficiency just to prove to an algorithm that your legitimate financial transactions aren't actually fraud?

Emily That is insane. It is wild to think that the future of digital identity might require us to fake our own humanity by actively being worse at using computers.

Jeff It's a strange new world.

Emily It really is. Thank you for joining us on this deep dive. Keep learning. Keep questioning the invisible architecture of the systems you rely on. And please enforce those bounding boxes.

Jeff Thanks for having me.

Emily We'll see you next time as we continue exploring the shifting realities of our digital world.

Jeff You've been listening to "What the F Happened? Fraud and Financial Crime Deconstructed," a DEFEND podcast by DataVisor. If you want to keep learning between episodes, check out DEFEND Training, a set of self-paced online courses for fraud and financial crime professionals, practical and built around real world scenarios.

Emily And you can earn CPE credits through the ACFE San Francisco Bay area chapter. You can find it at datavisor.com/defend-training. The link is in the description.