EPISODE 6 | Analysis

Regulatory Expectations Around Explainable AI

FEBRUARY 17, 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.

Jeff So we're looking at a stack of reports today that, well, they paint a pretty tense picture of a financial world right now.

Emily A very tense picture.

Jeff You have these massive institutions. We're talking banks, fintechs, and they're caught in what really looks like a paradox. They're under this, you know, immense pressure to stop fraud and money laundering.

Emily It's the classic arms race. The criminals get smarter, so the defenses have to get smarter, right?

Jeff So they're deploying the most powerful AI they can get their hands on. But here's the friction point.

Emily Ah, yes.

Jeff The regulators are looking at these new hyper intelligent systems. And they're basically saying, hold on.

Emily We don't trust this.

Jeff Exactly. We're seeing this major pivot where accuracy just catching the bad guy. It's no longer the only metric that matters.

Emily It's arguably not even the most important one anymore. If you look at the recent regulatory inquiries, the ones in the source material, the issue isn't that banks are missing fraud. No, it's that they can't explain how they found it.

Jeff That is the black box problem. And that's our mission for this deep dive. We're going to unpack why the industry is being forced to move from simply using AI to needing explainable AI. We need to figure out why a system that works perfectly might still be, well, illegal, and break down the specific checklist that turns a black box into something you can actually defend.

Emily It really all comes down to accountability. We're moving from an era where it was enough to just get the right answer to an era where you have to show your work.

Jeff And in high stakes finance.

Emily Showing your work is the difference between a compliant bank and a massive fine.

Jeff So let's start with the technology itself, because the sources describe this evolution that really created the problem in the first place. We didn't always have black boxes.

Emily No, not at all. For decades fraud detection was rules based. It was very linear logic.

Jeff Like a flowchart. If A happens, then do B exactly.

Emily If a customer wires more than ten thousand dollars internationally at three a m, flag it. It was rigid, sure, but it was transparent.

Jeff You could see the logic.

Emily Absolutely. Our regulator asks, why did you freeze this account? You could point to line forty of the code and say because they violated rule twelve.

Jeff But the problem with those simple rules is that they're brittle. The criminals just figured them out. Of course, they started structuring transactions. Doing what, nine thousand nine hundred dollars instead of ten thousand? This is Smurfing, right?

Emily Right. And that's why the industry pivoted to machine learning and eventually deep learning. These systems don't just follow hard coded rules.

Jeff No. They learn.

Emily They digest massive data sets. Transaction history, device fingerprints, mouse movements, IP addresses. And they learn patterns on their own.

Jeff And this is where the black box idea really comes in. Because when one of these deep learning models flags something, it isn't following a simple if this then that rule.

Emily It's it's exponentially more complex. The model might be looking at a thousand different variables at the same time. It might see that a user is typing ten percent faster than usual coming from a new IP address and transacting with someone who has a slightly higher risk score. No human wrote a rule for that specific combination.

Jeff The AI just learned it.

Emily It just learned that this specific cluster of variables equals, say, a ninety nine percent probability of fraud.

Jeff So the output is accurate. The system says fraud. But if you ask the system why.

Emily It can't give you a human readable reason, it can give you the math, the vector weights inside the neural network. But for a long time the attitude in tech was who cares?

Jeff I can hear the engineers listening right now making that exact argument.

Emily Huh?

Jeff Right. If the model stops terrorist financing or catches the credit card thief, isn't that the goal? Why does the explanation matter if the outcome is good?

Emily And that's the utilitarian argument. But the source material is very clear. In a regulated industry, the how matters just as much as the what. Which brings us to the concept of adverse action.

Jeff Okay. This is the legal side of things.

Emily Exactly. When a bank freezes an account or denies a transaction, they are taking an adverse action against you in the US and under rules like GDPR in Europe, that customer has rights. If you freeze someone's money and they ask why, you cannot legally just say.

Jeff The computer said you look suspicious.

Emily You can't. That sounds like a dystopian nightmare.

Jeff Computer says no.

Emily It's a nightmare for the consumer, but it's a liability nightmare for the bank. If you can't explain the decision, you can't defend it. Imagine trying to explain that to a judge.

Jeff Your honor, we don't know why we flag them, but our model is usually right.

Emily You get laughed out of court. And then there's the other big issue.

Jeff Bias are right. The risk of the model finding proxies for protected attributes.

Emily This is where the black box becomes a compliance minefield. Let's say you feed the model ten years of historical data to train it. Okay, if that historical data contains biases, say certain zip codes where historically flagged. More often the model learns that pattern.

Jeff So it effectively learns to redline, but without anyone explicitly telling it to.

Emily Precisely it might flag transactions from a certain neighborhood as high risk just because of the zip code, not because of what's actually happening.

Jeff And with a black box, you wouldn't even know it's doing that. You just see the high risk label.

Emily But if a regulator audits you and finds you're disproportionately flagging a specific demographic, you're in deep trouble. Fair lending laws, non-discrimination statutes.

Jeff And saying we didn't know is not a defense.

Emily Ignorance is not a valid legal strategy. You have a duty to know, which is why the sources are so heavy on this concept of auditable logic.

Jeff This leads us to a really important distinction made in the reading. I think a lot of people, myself included, tend to use the words transparency and explainability almost as synonyms.

Emily They're not.

Jeff But the research draws a very hard line between them.

Emily It's the most critical distinction in this entire discussion. And honestly, it's where a lot of vendors. They try to pull the wool over buyers eyes.

Jeff So let's break it down. What does transparency actually mean in this context?

Emily Okay. Think of transparency As the glass box.

Jeff A glass box.

Emily It gives you a view of the system's architecture. It answers the question how is this built? So transparency means you have documentation on the data inputs, the model version, how often it's updated, maybe a dashboard with performance metrics.

Jeff So I can see the machine. I can see the gears turning.

Emily You can see that the gears are turning. But and this is the key, transparency doesn't tell you why the gears turned a specific way for a specific case.

Jeff Okay. So that brings us to explainability.

Emily And explainability is the why it's granular. It answers the question why was this specific transaction flagged?

Jeff The source used a medical analogy that I thought was really helpful here.

Emily The open surgery concept. Right. So transparency is like watching a surgery through a glass wall. You see the doctors, the tools, the patient, you know, a surgery is happening.

Jeff That's transparency.

Emily But unless you're a surgeon, you have no idea why the doctor just made that specific incision. Explainability. Is the doctor turning to you and saying, I am cutting here because the artery is blocked at this exact point.

Jeff That clarifies it perfectly. You can have a fully transparent system, great dashboards, even open source code that is completely unexplainable when it comes to one decision.

Emily And that is the trap a vendor will sell you. Transparency. Look at our beautiful interface and claim it solves the regulatory problem. But if an analyst clicks on an alert and can't understand the logic, that system is not explainable.

Jeff The regulators are catching on to this difference.

Emily Oh, they are the source lists the heavy hitters, the OCC, FinCEN, the Fatf on a global scale. They're all issuing guidance that moves beyond just model governance. They're looking for defensibility.

Jeff They want to know that a human being can validate the machine's output.

Emily Precisely. The concept is human in the loop validation. If the human can't understand the machine, the loop is broken. The regulator doesn't want to talk to your algorithm.

Jeff They want to talk to your compliance officer.

Emily And if your compliance officer can't explain the logic, you fail the audit. Simple as that.

Jeff So we've established the problem. The black box is a liability. We've established the goal. Explainability. But what does that actually look like in practice? The source gives a checklist of four core features.

Emily Yes.

Jeff I think we should walk through these because they're really practical.

Emily This is essential if you're a bank looking at vendors or a tech led building this in-house. These are your non-negotiables.

Jeff Okay. Requirement number one human readable rules and annotations.

Emily This sounds simple, but it is so often missing when an alert fires, it needs to come with plain language, not a vector code, not a raw data dump.

Jeff No jargon.

Emily It needs to say flagged because transaction velocity is four hundred percent higher than the thirty day average.

Jeff It's a translation. It's taking the complex math and making it something an investigator can grasp in five seconds.

Emily And it creates efficiency. If an analyst sees flagged for unusual location, and they can also see the customer bought a plane ticket yesterday, they can resolve that alert instantly, right?

Jeff But if they just see high risk score.

Emily They have to dig through the whole profile. It empowers the human to make a judgment call.

Jeff Okay. Requirement number two alert traceability.

Emily This is the show your work feature. You have to be able to drill down from the alert all the way back to the source data. Which model triggered this? Which rule? What specific data point was the tipping point?

Jeff Like a breadcrumb trail for the decision?

Emily Exactly. In a black box, that trail vanishes. In an explainable system, you can trace the entire decision path. This is crucial for root cause analysis.

Jeff So you can fix mistakes.

Emily If your system makes a mistake a false positive, you need to know why so you can fix the logic. You can't fix a black box, you can only retrain it and hope it learns better next time.

Jeff Requirement number three version control and audit logs. This feels like the boring part, but the source says it's critical for legal defense.

Emily It is the time travel problem. Imagine you're being investigated today for a decision your bank made two years ago.

Jeff And your AI model has probably updated fifty times since then.

Emily Right? The model you have today is not the model you had then. If you run the old data through today's model, you might get a different result. I see. So you need a system that logs the exact state of the logic at the moment of the decision. You have to be able to say on February seventeenth, twenty twenty four, this was the logic we were using. And based on that logic, this decision was correct.

Jeff Without that log, you can't prove you were compliant back then.

Emily And if you can't prove it, the regulator assumes you weren't.

Jeff And finally, number four configurable thresholds.

Emily This is all about control. Every institution has a different risk appetite. A global bank's risk profile is totally different from a small regional credit unions.

Jeff You can't have a one size fits all model.

Emily No, the source emphasizes that you cannot just accept out of the box settings. You need to be able to turn the dial yourself.

Jeff So if you're getting too many false positives.

Emily And you're annoying good customers by freezing their cards, you need to dial back the sensitivity. Or maybe you're under regulatory scrutiny and need to tighten the net. If the vendor locks that logic away in their black box, you're not managing your own risk.

Jeff You've just outsourced it.

Emily You've outsourced your risk management to a software company, and regulators explicitly say you cannot do that. You are responsible.

Jeff The source puts it so bluntly here it says without these four capabilities readable rules, traceability, audit logs, and configurability, even the most accurate AI is indefensible.

Emily Indefensible. It's a strong word, but it's the reality. If you can't defend the decision, the system is a liability, no matter how much fraud it catches.

Jeff So for our listeners who are in the industry maybe looking at buying one of these solutions, it's a big buyer beware.

Emily Warning here a huge one. The market is absolutely flooded with AI solutions right now. Everyone has an AI sticker on their laptop. It's the buzzword of the decade.

Jeff So how do you separate the real deal from the AI washing?

Emily You have to ask the uncomfortable questions during the demo when the vendor shows you the fancy dashboard, you have to ask, can I see the logic behind a specific alert?

Jeff And if they say, oh, well, that's our proprietary secret sauce.

Emily Then you have a problem. Proprietary is often just code for black box. You have to ask, do I have control over the thresholds? Can I tune this myself? And most importantly, can I audit the decision logic from two years ago?

Jeff If the answer is no or we're working on that, you walk away.

Emily You have to. Because again, when the fine comes, it's written to the bank, not the software vendor.

Jeff It's interesting. We started this deep dive talking about tech, neural networks, decision trees, but we're ending up talking about design philosophy. The source argues that explainability has to be built into the detections lifecycle. It can't just be an afterthought.

Emily You can't take a black box and just tape an explanation to the outside. It doesn't work. The system has to be designed from the ground up to be interpretable. It's a fundamental shift in how we build these tools.

Jeff So we're not just building for accuracy anymore.

Emily We're building for collaboration, a collaboration between human and machine.

Jeff That collaboration piece feels key. We hear so much about AI replacing humans, but in this context, it seems like the AI is almost useless without the human to validate it.

Emily It needs the human for context, and it needs the human for accountability. The AI can crunch the numbers, but the human has to stand behind the final decision.

Jeff Okay, let's zoom out. We've covered the black box problem, the regulatory pressure, the checklist for compliance. But there's a broader takeaway here about the nature of protection.

Emily This is what really stuck with me from the reading. We tend to think of protection in finance as just stopping the bad guys.

Jeff Right. Stopping the fraudster, stopping the money launderer. That's the operational goal.

Emily But there's a second layer of protection, protecting the institution itself from regulatory enforcement, from lawsuits, from reputational damage.

Jeff And an unexplainable AI might solve the first problem, but create the second one.

Emily Exactly. If your AI stops the fraud but gets you fined fifty million dollars for lack of model governance. It didn't protect you. It actually introduced a massive new risk.

Jeff So true protection means doing the job and being able to prove you did it correctly.

Emily That's it. It's a powerful reframing. Clarity isn't just a nice to have feature anymore, it is a security feature.

Jeff Clarity is the foundation of trust.

Emily And banking is built on trust.

Jeff So to bring it all home, the days of move fast and break things really seem to be over in this sector.

Emily The new motto is move fast, but bring a map.

Jeff Ah, I like that. Move fast, but bring a map. You need to know how you got there. For everyone listening, whether you're running a compliance team, building these models, or you're just a consumer wondering why your transaction got declined. Think about that. Why? The next time you interact with an automated system, ask yourself, is this just transparent? Or is it actually explainable.

Emily Because there is a very, very big difference.

Jeff Thanks for diving deep with us today. Check your systems, ask the hard questions and we will catch you on the next one.

Emily Goodbye everyone.

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's in the description.

Jeff This episode's audio was generated using Google's Notebook LM based on expert analysis and trusted sources. Thanks for listening. We'll see you next time.