fairplay

AI Agents in Risk Management:

Safe Deployment and Real-World Use Cases

SEPTEMBER 2, 2026

The Panel

Delicia Reynolds Hand

Senior Director,
Digital Marketplace

Consumer Reports

Abby Hogan

General Counsel

fairplay

Kevin Lewis

CoFounder, CRO

Spring

How are AI agents different than LLMs and machine learning models?

Autonomous

Operating in real time

Evolving / constantly updating

Real-world use cases

Financial Services & Telecommunications

Customer Operations

Customer Service

Inbound and outbound support with automated handling of complaints & disputes.

Billing & Recovery

Streamlined billing processes, collections management, and automated recovery.

Trust & Compliance

Fraud Prevention

Real-time threat detection and mitigation across transactions.

Identity & AML

Robust Customer Identification (KYC) and strict anti-money laundering compliance.

Smart Ecosystems

Agentic Commerce

Next-generation autonomous transactions and automated shopping experiences.

Smart Home Systems

Seamless device integration and intelligent automation within home environments.

Early Adopters of AI agents

TIER	WHO	WHAT'S RUNNING
Largest consumer deployment	Amazon Alexa+ (GA Feb 2026, tens of millions), Google Gemini for Home (early access), Apple Siri (fall 2026)	Multi-step task execution, third-party bookings and purchases, camera-triggered automation
Live at scale, enterprise	Card networks, large issuers, e-commerce platforms	Fraud detection, dispute intake, collections prep, KYC review — human-in-the-loop
Consumer commerce, early	Perplexity Comet, ChatGPT shopping, Gemini/AI Mode	Discovery, comparison, agent-executed checkout

The most-deployed consumer agents are in the least-regulated setting. Financial services has MRM. Telecom doesn't. The home has neither, and it's where the agent has physical control and a payment method.

Smart Home as the proving ground

The intelligent home: where agentic AI already lives

Samsung's own roadmap, presented to CR, describes three stages:

1. Single-point IoT devices *(where most homes are)*
2. Semi-autonomous cooperation between devices
3. A home that "takes care of you"

Real example: control you can't get back: Google's own documentation indicates that switching a Nest or Home speaker to Gemini for Home **cannot be reverted on that speaker**. Alexa+ can be exited. Same category, opposite answer on the most basic control question: *can I undo this?*

Why it matters: The agent's authority expanded, the consumer's didn't, and the change was made to hardware they already owned.

CR has a (soon to be) published testing standard for exactly this. The **Intelligent Home Governance Framework v3.0 (June 2026)** - five sections covering AI security, privacy and consumer control, autonomy and agency, interoperability, and governance - with **gating critical indicators**: no hard off-switch, no spending limit, or no fail-safe caps the product at *Not Recommended* regardless of how well it scores elsewhere.

The three questions we test: Can I shut it off right now? Can it spend without asking? Does it fail safe?

Early Adopters of AI Agents (Banking & Finance)

AML Compliance

BSA / AML / KYC

Automating sanctions, adverse media checks, PEP screening, and transaction testing.

Debt Collection

Inbound Accounts

Managing inbound collection inquiries, structured payment negotiation, and compliance-safe responses.

Customer Service

Front-line Support

Resolving routine account inquiries, standard disclosures, and orchestrating complex multi-system handoffs.

Credit Decisioning

Human-in-the-Loop

Assisting underwriting analysis, drafting justifications, and raising flags for final human verification.

Human-in-the loop

Efficiency / safety tradeoff

Reason for having a human in the loop

- regulatory requirements
- practical limitations of the technology

Safe deployment of AI agents starts with a few basic questions:

Governance / Safety element

- Does the agent arrive at the correct answer?
- Does the agent take the right action?
- Can the agent be tricked into doing the wrong thing?
- Is the agent loyal - to whom?

How to monitor

How is that different from machine learning systems?

Where does that break down?

Beyond MRM moment

Different approach than for deterministic systems

Does the Agent Arrive at the Correct Answer?

Does the Agent Take the Right Action?

Can the Agent be tricked into doing the wrong thing?

Safe deployment of AI agent should also include: Loyalty

The point: Right answer, right action, not tricked; and, it can still be worse off for the customer. Competence and loyalty are separate tests.

Conflict of Interest

An agent can be accurate, in-scope, and un-manipulated while steering to the higher-margin plan.

Fairness vs. Benefit

Fairness testing was built for predictive ML; disparate impact across outcome distributions. A recommendation can be perfectly fair across protected classes and bad for every single customer.

Relational Failure

Agents fail relationally, not just statistically: sycophancy to whoever is speaking, undisclosed conflicts, procedural pressure.

Where are we seeing the greatest risk of deploying AI agents?

Presenter discussion

Real World Example: Duty of Vigor

The risk nobody is preparing for: the agent that doesn't act

Example Scenario

A carrier's service agent handles a bill inquiry. It answers correctly, resolves the ticket, closes the case. It does not mention the unused line, the expired promo that reverted to rack rate, the add-on the customer stopped using in March, or that a lower-priced plan now fits their usage.

Every individual answer is accurate. Every control passes. CSAT goes up.

Definition

Loyalty failures are usually omissions, not errors. The agent doesn't do the wrong thing; it declines to do the right thing.

Why it matters

The asymmetry that costs consumers isn't information, it's effort.

Distinction

Nothing technical prevents an agent from flagging the unused line. That's a business-model choice, not a constraint.

How should companies think about mitigating those risks?

Risk mitigation strategies, resources

“Know your agent” (KYA)

Highest value target for agentic services from a consumer perspective - intersection of largest potential and highest risk - opportunity to anticipate and prepare for these risks

Meaningful control.

- Scoped, time-bounded, revocable authorization - with an agent?
- Visible action logs with real intervention points
- Reversibility tiers
- A dispute channel built for AI-initiated transactions