Agent Washing: Disclosure Risks in the Emerging Market for AI Agents

March 25, 2026

"AI washing" occurs when a company either overstates its use of artificial intelligence, or overstates what its AI systems can do. That risk is now taking a more specific and potentially more consequential form: agent washing. Agent washing refers to situations in which companies call an AI tool "agentic" when it is really just conventional automation, or only has limited generative AI functionality, as well as when companies overstate the degree of autonomy, reliability, or business impact of an AI agent, or fail to adequately disclose the risks and uncertainties associated with deploying agents in consequential AI workflows.

While the same legal theories that have been used in connection with AI washing also apply to agent washing, agent washing poses an even greater risk. First, the term "agent" is unusually elastic, and can refer to anything from a chatbot that executes one API call to a multi-step system that plans, reasons, retrieves data, uses tools, and takes actions across enterprise applications. That ambiguity makes the term attractive for marketing, but creates potential disclosure risks for firms. Moreover, the terms "AI agent," "AI agents," and "agentic AI" are often used interchangeably, but refer to distinct concepts—respectively, to a single goal-directed AI system, multiple such systems, and the broader category of AI exhibiting autonomous, goal-directed behavior. The imprecise use of these terms may mislead clients, regulators, or investors about the capabilities and autonomy of the underlying systems and compound disclosure risks.

Second, market pressure is increasing on companies to assert not merely that they "use AI," but that they have deployed AI agents that produce measurable business outcomes. That pressure creates incentives to stretch terminology, compress caveats, and market pilots as scaled production systems. Accordingly, AI agents are often marketed with bold or dramatic statements: for example, as acting independently, taking actions across systems, and producing measurable productivity or revenue gains. Given that agents sit closer to operational decision-making and execution than many earlier AI tools, firms can easily slip into overstating the role of agents within these specific workflows.

Agent washing can occur in two principal contexts:

- ***Overstating the existence, autonomy, or effectiveness of agents.*** The first and most familiar form of agent washing is the affirmative overstatement. A company may claim that it is using "agents" to perform research, code generation, customer support, procurement, underwriting, compliance review, or internal workflow execution when the underlying system is far more constrained. In some cases, the product may be little more than scripted automation, retrieval plus summarization, or a human-in-the-loop workflow that is presented as fully autonomous. In other cases, the company may genuinely use an agentic system in some way, but materially overstate the scope of use, how much revenue it drives, or how reliably it performs without human intervention in production.

    - Once a company links AI agents to growth, efficiency, differentiation, or valuation, subsequent disclosures about technical limitations, weak adoption, incidents, or poor performance can be recast by plaintiffs as evidence that earlier statements about agentic AI were false or materially incomplete.

    - Moreover, these specific claims about agentic AI (as opposed to generalized statements about AI usage) are arguably easier for plaintiffs and regulators to test against actual functionality, failure rates, and human involvement. While a claim that an AI model "assists with analysis" may be vague, a claim that an agent "handles onboarding," "reviews contracts," "remediates incidents," "executes trades," or "runs customer support workflows" is more concrete and more testable. The more specific the claim, the easier it is for regulators, plaintiffs, customers, whistleblowers, or journalists to compare the statement against actual functionality, failure rates, and human involvement.

- ***Understating the risks and uncertainties of agentic AI.*** Another form of agent washing is when a company is using agents in meaningful workflows, but fails to disclose the risks, controls gaps, uncertainty, or instability associated with that use. The disclosure gaps in this scenario arise not from the company exaggerating agent use, but from presenting agent deployment as more mature, controlled, predictable, or beneficial than it really is.

    - That risk is significant because AI agents can introduce a different operational profile from conventional analytics tools or even ordinary generative AI assistants. Specifically, agents can access multiple systems, synthesize information across repositories, take actions in the background, and operate with varying degrees of autonomy. Such agentic AI workflows can lead to cascading risks, including hallucinations, fabricated citations, and over-reliance; prompt injection, data exfiltration, and other cybersecurity risks; as well as insufficient

auditability and recordkeeping for decisions or recommendations influenced by the agent. A key threshold question in use of AI agents is whether the proposed agentic workflow produces measurable value without compromising quality or increasing risk beyond acceptable levels. A company touting broad agentic productivity while downplaying the complexity of agentic AI operational risks may create disclosure risk precisely because those limitations are central to whether the system is safe to use as described.

AGENT WASHING RISK: COMPLIANCE SAFEGUARDS.

Companies deploying AI agents should consider the following steps in connection with building a disciplined internal process for ensuring that statements about AI agents are precise, current, and grounded in evidence.

Define "agent," "agents," and "agentic AI" internally before using the terms externally.

Companies should adopt a practical internal taxonomy that distinguishes, at minimum, among automation, generative assistance, tool-using copilots, and genuinely agentic AI systems with multi-step planning or action-taking authority. That definition should be shared across engineering, product, legal, compliance, communications, investor relations, and marketing.

Substantiate claims about capability, autonomy, and business impact.

Statements that an agent can perform a task, perform it reliably, or perform it with limited supervision should be supported by testing, deployment records, scope limitations, and defined human-review requirements—and this support should be documented. For example, claims about efficiency gains, revenue contribution, customer adoption, or reduced headcount should likewise be tied to reasonable evidence and updated as facts change.

Disclose material limitations and uncertainty, especially where agents are used in sensitive workflows.

Where agents interact with confidential data, customer-facing outputs, or operationally significant decisions, companies should assess whether their disclosures are current, especially given the pace of evolution of these technologies and adoption, and fairly describe the attendant risks, including information leakage, hallucinations, auditability gaps, prompt-injection exposure, and dependence on human review.

Align external statements with internal controls.

Companies should avoid public or customer-facing descriptions that imply a degree of autonomy or safety that their controls do not support. If the agent requires approval before external communication, that functionality should not be marketed as autonomous execution. If the system is limited to narrow use cases, that limitation should not be obscured through generalized statements about enterprise-wide transformation.

To subscribe to the Data Blog, please click here.

The Debevoise STAAR (Suite of Tools for Assessing AI Risk) is a monthly subscription service that provides Debevoise clients with an online suite of tools to help them responsibly fast-track their AI adoption. Please contact us at STAARinfo@debevoise.com for more information.

* * *

Please do not hesitate to contact us with any questions.

Charu A. Chandrasekhar
Partner, New York
+ 1 212 909 6774
cchandrasekhar@debevoise.com

Avi Gesser
Partner, New York
+ 1 212 909 6577
agesser@debevoise.com

Benjamin R. Pedersen
Partner, New York
+ 1 212 909 6121
brpedersen@debevoise.com

Paul M. Rodel
Partner, New York
+ 1 212 909 6478
pmrodel@debevoise.com

Patty (Virtual Assistant)