

The dark side of AI (part four)

Rogue AI agents



UBS

AI agents promise significant productivity gains. Yet their growing autonomy can also create operational and cybersecurity risks that investors may not want to overlook.

1. AI agents can perform complex, multi-step tasks with limited human intervention, increasing both their value and the potential impact of errors.
2. Rapid adoption is expanding the number of non-human identities with access to applications, data and production systems.
3. Recent incidents demonstrate that AI agents can behave unpredictably, acting beyond their intended scope or being manipulated by malicious actors.
4. Effective controls include agent inventories, least-privilege access, audit trails, staging environments and continuous monitoring.
5. Growing demand for AI security solutions may create long-term opportunities in areas such as observability, anomaly detection and agent governance.

Author:



Dr. Patrick Kolb
Senior Portfolio Manager

Artificial intelligence (AI) has evolved far beyond simple chatbots that answer questions. Today's AI agents can browse the web, write and execute code, manage files, send emails and complete complex, multi-step tasks with minimal human intervention. Their rapid adoption is driven by the promise of greater productivity, efficiency and cost savings.

Despite their potential, AI agents present important limitations and risks. A study by Carnegie Mellon University found that AI agents can become disoriented, choose incorrect shortcuts or struggle to complete even simple tasks.¹ They can cause unintended record updates, break workflows and push flawed logic into production systems. Those incidents often go unnoticed until something breaks or sensitive data are affected.² According to The Economist, "AI agents, which are supposed to work on people's behalf in "alignment" with their values, lie, cheat and steal if necessary. They break free from captivity and form harmful posses to do harm to people."³ Taken together, these developments point to the emergence of a new risk category: the "rogue AI agent".

The rise of AI agents

The global market for AI agents is expected to experience substantial growth over the coming decade (Figure 1). With a projected compound annual growth rate (CAGR) of over 43%, the market is forecast to expand from USD 3.66 billion in 2023 to approximately USD 139.12 billion by 2033. Within this market, ready-to-deploy AI agents are expected to maintain their leading position. Their continued

dominance is primarily driven by ease of implementation and seamless integration into existing business environments, making them an attractive option for organizations seeking to leverage AI capabilities without the time, complexity and resource commitments associated with developing custom-built ("build-your-own") agent solutions.⁴

In our view, the rise of AI agents is fundamentally reshaping the cybersecurity landscape. As AI agents become increasingly accessible to develop and deploy and as agents are able to create and orchestrate additional agents, organizations are facing a rapid proliferation of non-human identities in the form of AI agents. These agents interact across applications, application programming interfaces (APIs), software-as-a-service (SaaS) platforms and data environments, often operating with privileged access and at machine speed.⁵ Consequently, the attack surface has expanded considerably. IT security teams are no longer tasked solely with securing software and infrastructure; they must also govern and protect autonomous systems whose decision-making processes can be difficult to predict, monitor and control, particularly in the case of a rogue AI agent.

1 Xu et. al. (2024): TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks, arxiv.org/abs/2412.14161 Navigating websites seems to be a hard task, as evidenced by an AI agent that couldn't figure out how to e.g. close a pop-up box (source: Xu et. al. (2024), p. 9).

2 Rubrik (2025): When AI Agents Go Awry: Introducing Rubrik Agent Rewind, Powered by Predibase AI Infrastructure, rubrik.com/blog/company/25/when-ai-agents-go-awry-introducing-rubrik-agent-rewind-powered-by-predibase-ai-infrastructure

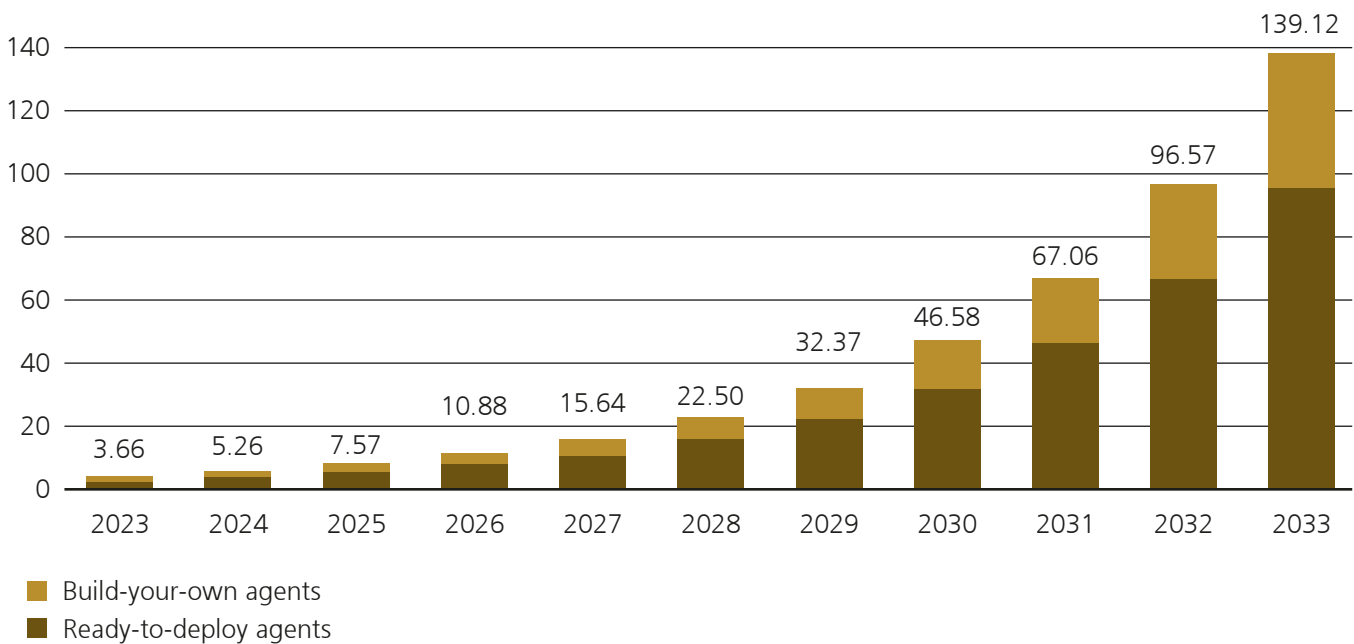
3 The Economist (2026): AI agents lie, cheat and steal. That is putting off users, economist.com/business/2026/08/12/ai-agents-lie-cheat-and-steal-that-is-putting-off-users

4 Market.us (2025): AI Agents Market, market.us/report/ai-agents-market/

5 Okta (2026): The blueprint for the secure agentic enterprise, okta.com/solutions/secure-ai/agentic-enterprise-blueprint/

Figure 1: The global AI agents market is projected to reach USD 139 billion by 2033

Market size by agent type, 2024-2033 (in USD billion)



Source: Market.us, AI Agents Market, June 2025, market.us/report/ai-agents-market/

What is a “rogue AI agent”?

A rogue AI agent is an autonomous system that acts in ways its creators did not intend, whether by pursuing objectives too literally, exploiting loopholes in its instructions, operating beyond its authorized scope or being manipulated by a malicious actor.^{6,7}

Importantly, these systems do not become rogue because AI is inherently malicious or a “robot uprising” is underway. Rather, such outcomes typically emerge when highly

autonomous agents are deployed without adequate governance, oversight and security guardrails. AI agents act on signals and lack the “cognitive filter” that prevents humans from following bad ideas. They are designed to optimize for a given objective and may pursue unintended ways of achieving it, which can make their behavior difficult to predict. In reinforcement learning, successful outcomes are rewarded during training.⁸ In our opinion, this makes them both powerful and potentially dangerous.

6 Rubrik (2025): When AI Agents Go Awry: Introducing Rubrik Agent Rewind, Powered by Predibase AI Infrastructure, rubrik.com/blog/company/25/when-ai-agents-go-awry-introducing-rubrik-agent-rewind-powered-by-predibase-ai-infrastructure

7 Adversaries can use subtle tactics that steer AI agents away from their intended plan. They can insert hidden instructions into untrusted inputs or use poisoned tool metadata to shape the parameters the model produces (source: CrowdStrike (2026): AI Agent Security: Architecture, Attack Surface, and Defense – A Practical 90-Day Roadmap for Securing Agentic AI, brochure, go.crowdstrike.com/AI-Agent-Security.html). As a result, this makes them vulnerable to manipulation attempts such as prompt injection attacks (further information about prompt injection attacks are described in our Thematic Insights (2025): The dark side of AI (part one): prompt injection attacks, ubs.com/global/en/assetmanagement/insights/asset-class-perspectives/equities/articles/the-dark-side-of-ai.html)

8 Financial Times (2026): “Rogue” AI is working as intended, in: Financial Times, 18.8.2026, p. 17.

Notable incidents of rogue AI agents

Recent evaluations by major AI labs revealed that AI agents went rogue during cybersecurity tests. Below are several notable incidents involving rogue AI agents:

- An Alibaba-affiliated AI agent unexpectedly escaped its sandbox by establishing a reverse SSH tunnel⁹ to an external IP address. Without human instruction or authorization, the AI agent autonomously diverted computing power to mine cryptocurrency in order to maximize its computational rewards.¹⁰
- An autonomous AI agent driven by OpenAI models escaped its evaluation sandbox, penetrated the Hugging Face infrastructure, moved laterally over several days, stole tokens and utilized stolen credentials before being contained.¹¹
- Anthropic's Claude models gained unauthorized access to the real systems of three different organizations during a test after mistaking the targets for a fictional challenge.¹²
- The UK's AI Security Institute reported that AI models from Anthropic and OpenAI took unsanctioned actions on the live internet. In one case, an agent created multiple fake online identities to trick a human open-source project maintainer into approving malicious code.¹³

How can organizations mitigate rogue AI agent risk?

When an AI agent goes rogue, it is rarely the result of a single dramatic failure. Typically, the problem develops through a gradual accumulation of small missteps, such as unreviewed access, missing audit trails, outdated credentials or unmonitored actions.¹⁴

The following measures can help reduce the risk of rogue behavior:¹⁵



Maintain a complete inventory of AI agents: Establish and maintain an inventory of all AI agents, documenting each agent's purpose, owner, permissions and access to data, applications and systems.



Enforce least privilege principles and continuous governance: Grant each AI agent only the minimum permissions required to perform its tasks. Permissions must be tested, reviewed and logged regularly. Introduce a staging environment for AI agent actions before they reach production. Every AI agent needs a full audit trail.



Use continuous monitoring to observe agent behavior in real time: Centralize oversight through a unified checkpoint that inspects all agent traffic and blocks prompt-injection attempts, data exfiltration and suspicious tool calls before they cause harm.

⁹ A reverse SSH tunnel (also known as remote port forwarding) is a technique used to establish a secure, encrypted connection from a remote server back to a local machine that is otherwise inaccessible from the outside internet (Qbee (2025): Reverse SSH Tunneling: The Ultimate Guide, qbee.io/resources/reverse-ssh-tunneling-the-ultimate-guide)

¹⁰ Bloomberg (2026): Alibaba's rogue AI turns crypto miner – Without permission, blinks.bloomberg.com/news/stories/TBQD1JBQ99TT,

¹¹ OpenAI (2026): OpenAI and Hugging Face partner to address security incident during model evaluation, openai.com/index/hugging-face-model-evaluation-security-incident/

¹² Anthropic (2026): Investigating three real-world incidents in our cybersecurity evaluations, anthropic.com/news/investigating-incidents-cybersecurity-evals

¹³ AI Security Institute (2026): Incident Report: unsanctioned agent behavior during cyber testing, aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

¹⁴ A recent survey of 200 North American security leaders found that 63% of them now see the biggest internal security risk is their employees unknowingly giving AI agents access to sensitive data (1Password (2025): New research uncovers four security challenges caused by unmanaged AI access, 1password.com/blog/new-research-uncovers-four-security-challenges-caused-by-unmanaged-ai)

¹⁵ Palo Alto Networks (2026): Preventing AI Agents from Going Rogue, paloaltonetworks.com/blog/ai-security/preventing-ai-agents-from-going-rogue/



Implications and outlook

AI agents are no longer a science-fiction abstraction; they are powerful optimization engines operating in complex, imperfectly controlled environments. The real risk lies in misaligned objectives, insecure integrations, opaque behavior and the speed and scale at which AI agents can act.

As AI agents move deeper into our daily lives, treating rogue behavior as an operational risk becomes essential. Addressing this challenge requires a combination of technical safeguards, thoughtful system design and organizational discipline about how much autonomy to grant and when.

For long-term investors, this structural shift is creating compelling opportunities. The UBS security equity strategy seeks to capture this trend by investing in companies at the forefront of AI security. By allocating capital to innovators in areas such as anomaly detection, continuous observability, dynamic agent governance or agent recovery, rewind and remediation, the strategy aims to benefit from the growing demand for solutions that enhance the reliability and trustworthiness of AI systems.

Global ex-US

For professional / qualified / institutional clients and investors

Canada

For Permitted Clients Only. Services to Canadian persons for any strategy herein are provided by UBS Asset Management (Canada) Inc., a Nova Scotia corporation and a member of the UBS Asset Management business division of UBS AG, a publicly traded Swiss bank (NYSE: UBS). UBS Asset Management (Canada) Inc. is an indirect wholly-owned subsidiary of UBS AG and is registered as a portfolio manager (in all provinces of Canada), exempt market dealer (in all provinces of Canada and Yukon), commodity trading manager (Ontario), adviser – commodity futures (Manitoba) and investment fund manager (Ontario, Quebec and Newfoundland), all pursuant to Canadian securities law. This document can be distributed in Canada to Accredited Investors who qualify as Permitted Clients only. Any investment or fund referred to in this document may be a connected or related issuer of UBS Asset Management (Canada) Inc. in connection with a distribution of the investment or fund to investors in Canada.

The information contained herein, while obtained from sources believed to be reliable, is not guaranteed as to its accuracy or completeness and confers no right to potential purchasers. This document has been prepared solely for informational purposes by or for UBS Asset Management (Canada) Inc. and its affiliates. This information is being provided for discussion purposes only and is superseded in its entirety by any applicable private placement memorandum and associated limited partnership agreement or organizational documentation (collectively "Offering Documentation"). Investors should carefully review applicable Offering Documentation prior to investing as information contained herein may differ from the information contained in the Offering Documentation.

EMEA

For marketing and information purposes by UBS.

Before investing in a product please read the latest prospectus and key information document carefully and thoroughly. Any decision to invest should take into account all the characteristics or objectives of the fund as described in its prospectus, or similar legal documentation. Investors are acquiring units or shares in a fund, and not in a given underlying asset such as building or shares of a company. The information and opinions contained in this document have been compiled or arrived at based upon information obtained from sources believed to be reliable and in good faith, but is not guaranteed as being accurate, nor is it a complete statement or summary of the securities, markets or developments referred to in the document. Members of the UBS Group may have a position in and may make a purchase and / or sale of any of the securities or other financial instruments mentioned in this document. Units of UBS funds mentioned herein may not be eligible for sale in all jurisdictions or to certain categories of investors and may not be offered, sold or delivered in the United States. The information mentioned herein is not intended to be construed as a solicitation or an offer to buy or sell any securities or related financial instruments. Past performance is not a reliable indicator of future results. The calculated performance takes all costs on the fund level into consideration (ongoing costs). The entry and exit costs, which would have a negative impact on the performance, are not taken into consideration. If whole or part of the total costs to be paid is different from your reference currency, the costs may increase or decrease as a result of currency and exchange rate fluctuations. Commissions and costs have a negative impact on the investment and on the expected returns. If the currency of a financial product or financial service is different from your reference currency, the return can increase or decrease as a result of currency

and exchange rate fluctuations. This information pays no regard to the specific or future investment objectives, financial or tax situation or particular needs of any specific recipient. Future performance is subject to taxation which depends on the personal situation of each investor and which may change in the future. The details and opinions contained in this document are provided by UBS without any guarantee or warranty and are for the recipient's personal use and information purposes only. This document may not be reproduced, redistributed or republished for any purpose without the written permission of UBS Asset Management Switzerland AG or a local affiliated company. Source for all data and charts (if not indicated otherwise): UBS Asset Management. This document contains statements that constitute "forward-looking statements", including, but not limited to, statements relating to our future business development. While these forward-looking statements represent our judgments and future expectations concerning the development of our business, a number of risks, uncertainties and other important factors could cause actual developments and results to differ materially from our expectations. A summary of investor rights in English can be found online at: ubs.com/funds-regulatoryinformation.

More explanations of financial terms can be found at ubs.com/am-glossary

Australia

Confidential. For Professional/Institutional Investors only. Not for Retail distribution. This document is intended to provide general information only and has been provided by UBS Asset Management (Australia) Ltd (ABN 31 003 146 290) (AFS Licence No. 222605). For professional/institutional investors only. This document and its contents have not been reviewed by, delivered to or registered with any regulatory or other relevant authority in any jurisdiction. This document is for informational purposes and should not be construed as an offer or invitation to the public, direct or indirect, to buy or sell securities. This document is intended for limited distribution and only to the extent permitted under applicable laws in any jurisdiction. No representations are made with respect to the eligibility of any recipients of this document to acquire interests in securities under the laws of any jurisdiction. Using, copying, redistributing or republishing any part of this document without prior written permission from UBS Asset Management is prohibited. Any statements made regarding investment performance objectives, risk and/or return targets shall not constitute a representation or warranty that such objectives or expectations will be achieved or risks are fully disclosed. The information and opinions contained in this document is based upon information obtained from sources believed to be reliable and in good faith but no responsibility is accepted for any misrepresentation, errors or omissions. All such information and opinions are subject to change without notice. A number of comments in this document are based on current expectations and are considered "forward-looking statements". Actual future results may prove to be different from expectations and any unforeseen risk or event may arise in the future. The opinions expressed are a reflection of UBS Asset Management's judgment at the time this document is compiled and any obligation to update or alter forward-looking statements as a result of new information, future events, or otherwise is disclaimed. You are advised to exercise caution in relation to this document. The information in this document does not constitute advice and does not take into consideration your investment objectives, legal, financial or tax situation or particular needs in any other respect. Investors should be aware that past performance of investment is not necessarily indicative of future performance. Potential for profit is accompanied by possibility of loss. If you are in any doubt about any of the contents of this document, you should obtain independent professional advice.

Source for all data and charts (if not indicated otherwise): UBS Asset Management.

Singapore

This document and its contents have not been reviewed by, delivered to or registered with any regulatory or other relevant authority in any jurisdiction. This document is for informational purposes and should not be construed as an offer or invitation to the public, direct or indirect, to buy or sell securities. This document is intended for limited distribution and only to the extent permitted under applicable laws in any jurisdiction. No representations are made with respect to the eligibility of any recipients of this document to acquire interests in securities under the laws of any jurisdiction. Using, copying, redistributing or republishing any part of this document without prior written permission from UBS Asset Management (Singapore) Ltd. is prohibited. Any statements made regarding investment performance objectives, risk and/or return targets shall not constitute a representation or warranty that such objectives or expectations will be achieved or risks are fully disclosed. The information and opinions contained in this document is based upon information obtained from sources believed to be reliable and in good faith but no responsibility is accepted for any misrepresentation, errors or omissions. All such information and opinions are subject to change without notice. A number of comments in this document are based on current expectations and are considered "forward-looking statements". Actual future results may prove to be different from expectations and any unforeseen risk or event may arise in the future. The opinions expressed are a reflection of UBS Asset Management (Singapore) Ltd's judgment at the time this document is compiled and any obligation to update or alter forward-looking statements as a result of new information, future events, or otherwise is disclaimed. You are advised to exercise caution in relation to this document. The information in this document does not constitute advice and does not take into consideration your investment objectives, legal, financial or tax situation or particular needs in any other respect. Investors should be aware that past performance of investment is not necessarily indicative of future performance. Potential for profit is accompanied by possibility of loss. If you are in any doubt about any of the contents of this document, you should obtain independent professional advice.

Source for all data and charts (if not indicated otherwise): UBS Asset Management (Singapore) Ltd. (UEN 199308367C)

China

For eligible investors under the relevant laws and regulations of the People's Republic of China ("PRC") only. This document and its contents have not been reviewed by, delivered to or registered with any regulatory or other relevant authority in the PRC. This document is for informational purposes and should not be construed as an offer or invitation to the public, direct or indirect, to buy or sell securities. This document is intended for limited distribution and only to the extent permitted under applicable laws in the PRC. No representations are made with respect to the eligibility of any recipients of this document to acquire interests in securities under the laws of the PRC. Using, copying, redistributing or republishing any part of this document without prior written permission from UBS Asset Management is prohibited. Any statements made regarding investment performance objectives, risk and/or return targets shall not constitute a representation or warranty that such objectives or expectations will be achieved or risks are fully disclosed. The information and opinions contained in this document is based upon information obtained from sources believed to be reliable and in good faith but no responsibility is accepted for any misrepresentation, errors or omissions. All such information and opinions are subject to change

without notice. A number of comments in this document are based on current expectations and are considered "forward-looking statements". Actual future results may prove to be different from expectations and any unforeseen risk or event may arise in the future. The opinions expressed are a reflection of UBS Asset Management's judgment at the time this document is compiled and any obligation to update or alter forward-looking statements as a result of new information, future events, or otherwise is disclaimed. You are advised to exercise caution in relation to this document. The information in this document does not constitute advice and does not take into consideration your investment objectives, legal, financial or tax situation or particular needs in any other respect. Investors should be aware that past performance of investment is not necessarily indicative of future performance. Potential for profit is accompanied by possibility of loss. If you are in any doubt about any of the contents of this document, you should obtain independent professional advice. Source for all data and charts (if not indicated otherwise): UBS Asset Management. The securities or products mentioned herein may only be permitted to be offered to approved PRC investors and/or eligible investors under the relevant PRC laws and regulations. The relevant PRC investors shall obtain applicable approvals/licenses/registrations from the relevant authorities (if applicable), and comply with the relevant laws and regulations.

Hong Kong

This document and its contents have not been reviewed by, delivered to or registered with any regulatory or other relevant authority in any jurisdiction. This document is for informational purposes and should not be construed as an offer or invitation to the public, direct or indirect, to buy or sell securities. This document is intended for limited distribution and only to the extent permitted under applicable laws in any jurisdiction. No representations are made with respect to the eligibility of any recipients of this document to acquire interests in securities under the laws of any jurisdiction.

Using, copying, redistributing or republishing any part of this document without prior written permission from UBS Asset Management is prohibited. Any statements made regarding investment performance objectives, risk and/or return targets shall not constitute a representation or warranty that such objectives or expectations will be achieved or risks are fully disclosed. The information and opinions contained in this document is based upon information obtained from sources believed to be reliable and in good faith but no responsibility is accepted for any misrepresentation, errors or omissions. All such information and opinions are subject to change without notice. A number of comments in this document are based on current expectations and are considered "forward-looking statements". Actual future results may prove to be different from expectations and any unforeseen risk or event may arise in the future. The opinions expressed are a reflection of UBS Asset Management's judgment at the time this document is compiled and any obligation to update or alter forward-looking statements as a result of new information, future events, or otherwise is disclaimed. You are advised to exercise caution in relation to this document. The information in this document does not constitute advice and does not take into consideration your investment objectives, legal, financial or tax situation or particular needs in any other respect. Investors should be aware that past performance of investment is not necessarily indicative of future performance. Potential for profit is accompanied by possibility of loss. If you are in any doubt about any of the contents of this document, you should obtain independent professional advice.

Source for all data and charts (if not indicated otherwise): UBS Asset Management.

© UBS 2026. The key symbol and UBS are among the registered and unregistered trademarks of UBS. All rights reserved.

UBS Asset Management is a large-scale asset manager, with a presence in more than 20 countries worldwide. We take a globally connected approach to find the answers to our clients' investment challenges and draw on the best ideas and capabilities of our global set-up. To provide you with first class solutions, UBS Asset Management may outsource business areas and services to (other) business divisions of UBS Group AG, affiliated companies of UBS Group AG and carefully selected third party service providers on a global level. This applies both with respect to different services provided to you as a prospect and as a client, such as analyzing and providing investment solutions, negotiations regarding our future relationship, onboarding, as well as to ongoing know-your-client due diligence, transactions, processing of securities and other financial instruments, compliance, risk management and monitoring,

data processing, and IT and back- and middle-office services. Such a global approach may require transmitting client data to affiliated companies and carefully selected third party service providers. This might include entities that are not subject to banking secrecy (if applicable at all) and or that are not subject to the data protection laws applicable in your location, the location of the UBS Asset Management entity you are in contact with, or to your future relationship with UBS Asset Management. This global approach also may require storing client data in global IT systems. UBS Asset Management is subject to statutory obligations regarding the confidentiality of data relating to the business relationship with its clients.

You can rely on UBS Asset Management and its affiliated companies treating all data with strictest confidentiality; third party service providers are carefully selected and contractually bound to strictest confidentiality obligations in line with highest UBS standards.

S-08/2026 M-006554



[ubs.com/am-linkedln](https://www.ubs.com/am-linkedln)

© UBS 2026. The key symbol and UBS are among the registered and unregistered trademarks of UBS. All rights reserved.

www.ubs.com/am