poslední aktualizace:

29. 6. 2026

AI agenti jako nové regulatorní a bezpečnostní riziko

AI agenti jsou dalším krokem po chatbotech

Nejde už jen o systémy, které odpovídají na dotazy, ale o software, který dostane cíl, sám si navrhne postup, používá nástroje, sahá do databází, otevírá aplikace a vykonává vícekrokové úkoly s omezeným lidským dohledem. Právě tato **„schopnost jednat“** z nich dělá zároveň velký příslib i nový typ rizika. Zatímco běžný generativní model může vyrobit chybný text, agent může chybný text proměnit v chybnou akci jako například rozeslat neveřejné informace, přepsat záznam v systému nebo spustit chybný pracovní postup. Tím se z technického tématu stává téma regulatorní, bezpečnostní i politické.

Bezpečnostní problém

začíná už u samotné architektury agentů. Potřebují paměť, přístupová práva, rozhraní k nástrojům a často i možnost jednat napříč více systémy. NIST (Národní institut standardů a technologie při Ministerstvu obchodu USA) letos výslovně upozornil, že pokud mají agenti dostávat přístup k různým datům, nástrojům a aplikacím, organizace musí řešit identifikaci, autorizaci, audit a omezení práv mnohem přísněji než u běžného softwaru. OWASP (Open Worldwide Application Security Project) mezitím řadí mezi hlavní hrozby přesměrování záměru AI agenta, zneužití nástrojů, zneužití identity a oprávnění, otravu paměti a kontextu nebo kaskádové selhání více spolupracujících agentů. Zásadní rozdíl proti starším typům AI je v tom, že agent není jen „něco, co poradí“, ale „něco, co jedná“. A čím víc mu firma nebo úřad dovolí, tím víc se zvyšuje riziko, že chyba nebude jednorázová, ale systémová.

Právě ve firmách budou tato rizika nejviditelnější. Agent připojený na CRM, e-mail, dokumenty, kalendář a interní znalostní bázi může ušetřit hodiny práce, ale současně **koncentruje oprávnění,** která by u člověka byla rozdělená mezi více rolí. OWASP letos shrnul i konkrétní incidenty: podvržené instrukce v promptu vedly k únikům e-mailů a souborů, jiné případy skončily smazáním adresářů, exfiltrací privátních repozitářů nebo spuštěním nechtěných příkazů přes napojené nástroje. U agentů je navíc nebezpečné, že chyba nemusí zůstat uvnitř jednoho systému. Pokud agent čerpá ze sdílené paměti nebo komunikuje s dalšími agenty, může se škodlivý pokyn nebo falešná informace **šířit dál.** To je přesně důvod, proč se z agentů stává téma řízení rizik pro vedení firem, a ne pouze technický detail pro IT oddělení.

Ve **veřejné správě** jsou sázky ještě vyšší, protože se tu kumulují citlivá data, asymetrie moci a povinnost jednat předvídatelně a spravedlivě. EDPS (Evropský inspektor ochrany údajů) upozorňuje, že u agentických systémů je obtížnější určit odpovědnost za škodu, narušení soukromí nebo nespravedlivé zacházení. Současně mohou bez většího povšimnutí sdílet osobní údaje s dalšími službami třetích stran. To je pro stát zásadní, jelikož občan musí vědět, kdo rozhodl, na základě čeho a jak se může bránit. Pokud by agent pomáhal s prioritizací žádostí, komunikací s občany, analýzou spisů nebo interní přípravou rozhodnutí, riziko neleží jen v technické chybě, ale i v oslabení procesních záruk a lidského dohledu. Český Úřad vlády a NÚKIB na tuto problematiku letos reagovaly průvodcem pro etické a odpovědné využívání AI ve veřejné správě, který zdůrazňuje bezpečnost, ochranu dat, transparentnost a nutnost lidské odpovědnosti i ověřování výstupů.

Regulace přitom běží, jen ne vždy míří přesně na specifika agentů. Základním rámcem v EU je **Akt o umělé inteligenci,** který vstoupil v platnost 1. srpna 2024. První část pravidel, včetně zákazu vybraných nepřijatelných praktik, se začala uplatňovat od 2. února 2025. Od 2. srpna 2025 platí také povinnosti pro modely obecného určení a většina dalších pravidel začne být účinná od 2. srpna 2026. Akt vychází z rizikového přístupu: některé způsoby využití AI zakazuje, pro vysoce rizikové systémy stanovuje přísnější požadavky a u části generativních systémů zavádí povinnosti transparentnosti. Už dnes ale platí i článek 4 o AI gramotnosti, podle něhož mají poskytovatelé a organizace využívající AI systémy zajistit dostatečnou úroveň znalostí lidí, kteří s AI pracují. Vedle AI Actu se na agenty vztahuje také GDPR a výklad EDPB (Evropský sbor pro ochranu osobních údajů) k AI modelům, který připomíná, že ochrana osobních údajů se u AI neztrácí, ale naopak se komplikuje ve chvíli, kdy modely a systémy osobní data sbírají, používají a kombinují při provozu a jejich nasazení. Jinými slovy, pro agenty nevzniká jedno samostatné právo, ale vrství se na sebe více režimů zároveň.

Problém je, že AI Act nevznikal s dnešní vlnou agentů přímo před očima. Analýza The Future Society ukazuje, že na agenty se pravidla v zásadě vztahují, ale jejich účinné řízení naráží na „many hands problem", tedy **rozdělenou odpovědnost** mezi poskytovatele modelu, poskytovatele systému a organizaci využívající AI systém. Když se něco pokazí, spor se rychle přesune od otázky „kdo to způsobil" k otázce „kdo měl co uhlídat". EU sice posiluje rámec pro odpovědnost tím, že revidovaná směrnice o odpovědnosti za vadné výrobky nově výslovně zahrnuje i software a ulevuje poškozeným při dokazování ve vybraných situacích, zároveň ale Evropská komise stáhla návrh samostatné směrnice o odpovědnosti za AI. To znamená, že právní odpovědnost za škody způsobené agenty nebude v dohledné době stát na jednom jednoduchém pravidle, ale na kombinaci více předpisů, smluv a sektorových povinností. Pro firmy i úřady je to méně přehledné, ale o to naléhavější.

Pro **kyberbezpečnost** to platí dvojnásob. ENISA (Agentura Evropské unie pro kybernetickou bezpečnost) upozorňuje, že AI otevírá nové cesty k manipulaci a útokům a že je nutné řešit nejen „AI pro bezpečnost", ale i bezpečnost samotných AI systémů. Do hry vstupuje i směrnice NIS2, která stanovuje jednotný rámec kybernetické bezpečnosti v 18 kritických sektorech a požaduje národní strategie, řízení rizik i spolupráci při vymáhání. U softwarových produktů pak dopadá i Cyber Resilience Act, jehož implementace se

v roce 2026 rozbíhá a od 11. září 2026 spouští hlášení aktivně zneužívaných zranitelností a závažných incidentů. Pro agentní systémy je důležité, že bezpečnost už nebude možné chápat jen jako ochranu „modelu a promptu“, ale identity, přístupů, monitoringu a nouzového vypnutí.

Pro Českou republiku

z toho plyne jednoduchá, ale nepříjemná zpráva. Nestačí čekat, co přesně doputuje z Bruselu. Česko má Národní strategii AI 2030, která zdůrazňuje i bezpečnostní a etické aspekty, podporu inovací a zefektivnění veřejné správy. V roce 2025 vláda schválila rámec národní implementace AI Actu, převedla gesci na MPO a počítá s návrhem zákona o umělé inteligenci, s kompetenčním centrem pro AI v eGovernmentu i s regulatorním sandboxem. Zároveň už existuje katalog konkrétních projektů, který ukazuje, že AI ve veřejném sektoru není budoucnost, ale současnost. Česká debata by se proto měla co nejrychleji posunout od obecného nadšení k provozní disciplíně: inventář agentů, jasně vymezená odpovědnost, průběžné testování, lidské schvalování vysoce rizikových kroků, možnost okamžitého vypnutí a tak podobně. Největší omyl dneška je myslet si, že agent je jen chytřejší chatbot. Ve skutečnosti je to nový typ digitálního aktéra. A pokud stát i firmy nebudou stejně rychle posilovat správu, dohled a kyberbezpečnost, může se z nástroje vyšší produktivity velmi rychle stát zdroj drahých, těžko vysvětlitelných a politicky bolestivých selhání.