watsonx.ai
watsonx.governance

Regulace umělé inteligence
EU AI ACT

Martin Ryšánek, watsonx Architect
martin_rysanek@cz.ibm.com

watsonx

IBM

Co reprezentuje EU AI ACT?

zákon EU, který upravuje vývoj a využívání umělé inteligence v Evropské unii (EU). Přístup EU AI ACT je založen **na riziku**, který mohou systémy umělé inteligence (AI) reprezentovat **pro lidské zdraví, bezpečnost, soukromí a právo**.

Umělá inteligence (AI) nahrazuje v mnoha běžných procesech rozhodování člověka. Člověk dokáže své rozhodování sledovat, vyhodnocovat a případně korigovat. EU AI ACT vyžaduje stejné kontrolní mechanismy pro procesy založené na AI, když zavádí přísné **požadavky na správu, řízení rizik a transparentnost** používání AI **podle rizika.**

Koho se zaváděná opatření týkají?

- Poskytovatelů a vývojářů
  vyvíjejí a poskytují modely a systém AI pro všeobecné použití (např. chatboty)
- Uživatelů
  zavedli a používají systémy umělé inteligence, tj. organizací, jednotlivců i statní správy
- Importérů
  provozují AI systémy na trhu a na území EU, přestože sami se nachází mimo EU

Časový plán zavádění EU AI ACT

- Schválení, zveřejnění a platnost
  13. 3. 2024 - byl schválen Evropským parlamentem
  12. 7. 2024 - byl zveřejněn v Úředním věstníku
  1. 8. 2024 - vstoupil oficiálně v platnost
- Účinnost zákona pro systémy se **zakázanými AI praktikami**
  1. 2. 2025 – zakázány všechny systémy s nepřijatelnými riziky
- Účinnost zákona pro **vysoce rizikové systémy**
  1. 8. 2026 – po 2 letech platnost pro vysoce rizikové systémy

EU AI ACT rozdělení AI systémů podle rizika

nepřijatelné riziko

Použití je považováno za nepřijatelné, protože ohrožují bezpečnost, základní práva a hodnoty EU.

- sociální bodovací systémy
- manipulativní AI
- neoprávněné biometrické identifikace
- vykořisťující dohled nad dětmi a zaměstnanci

vysoké riziko

Spojeno s významnými riziky pro práva a bezpečnost lidí.

- bezpečnostní komponenty nebo výrobky, podrobeny předpisům dle přílohy č. II
- oblasti v příloze č. III

1. biometrické údaje
2. kritická infrastruktura
3. vzdělání a odborná příprava
4. zaměstnání
5. přístup k základním soukromým a veřejným službám a výhodám
6. vymáhání práva
7. řízení migrace, azylu a ochrany hranic
8. výkon spravedlnosti a demokratické procesy

- profilování

omezené riziko

Vyžaduje specifické transparentnostní opatření, aby uživatelé rozuměli, jak systém funguje a jakým způsobem je používán.

- povinnost transparentnosti (informace pro koncové uživatele, deep fakes, informační povinnost pro navazující uživatele)

minimální riziko

Neregulováno.

- např. filtry nevyžádané pošty

Pokuty při nedodržení podmínek regulace EU AI ACT

Úroveň 1: Systémy využívající zakázané praktiky AI

- Pokuty: až do výše 35 milionů EUR nebo 7 % celosvětového ročního obratu společnosti, podle toho, která částka je vyšší.
- Týká se nejzávažnějších porušení, jako je nasazení systémů AI klasifikovaných jako systémy představující nepřijatelné riziko.

Úroveň 2: vysoce rizikové systémy AI

- Pokuty: až do výše 15 milionů eur nebo 3 % ročního obratu, podle toho, která částka je vyšší.
- Tato úroveň zahrnuje nedodržení požadavků na vysoce rizikové systémy AI, včetně nezavedení řízení rizik, nezajištění kvality dat či zajištění dostatečné transparentnosti.

Poskytnutí nesprávných, neúplných nebo zavádějících informací

- Pokuty: až do výše 7,5 milionu eur nebo 1 % ročního obratu, podle toho, která částka je vyšší.

Institucím, orgánům, úřadům a agenturám Unie

- Při nedodržení EU AI ACT zakázaných AI praktik, budou uloženy pokuty až do výše 1,5 milionu eur. Za jakékoli další nedodržení budou uděleny pokuty až do výše 750 000 EUR.

Co musí organizace podle EU AI ACT realizovat?

Klasifikace rizika:
AI systémy jsou rozděleny do kategorií podle úrovně rizika. Systémy s nepřijatelným rizikem jsou zakázány, zatímco vysoce rizikové systémy podléhají přísným regulacím. Organizace je zodpovědná určit riziko svých AI systémů.

Systém hodnocení a řízení rizik:
Organizace musí zavést systém řízení rizik, který pokrývá celý životní cyklus vysoce rizikového AI systému. To zahrnuje identifikaci, hodnocení a minimalizace dopadů rizik spojených s používáním AI.

Technická dokumentace:
Organizace musí vypracovat technickou dokumentaci prokazující shodu s regulačními požadavky, kterou budou moci poskytnout regulačním orgánům.

Monitoring, sledování a zaznamenávání (transparentnost):
Systémy musí být navrženy tak, aby umožnily automatické zaznamenávání událostí důležitých pro identifikaci rizik a změn v průběhu jejich životního cyklu.

Lidský dohled:
Společnosti musí zavést mechanismy kontroly člověkem, tedy nechat člověka dohlížet na sledované ukazatele tak, aby byla udržena kontrola nad systémy umělé inteligence a zabránilo se nezamýšleným důsledkům.

Registrace a hlášení incidentů:
Společnosti provozující vysoce rizikové AI systémy musí registrovat své systémy u příslušných regulačních orgánů, jsou povinny hlásit jakékoli incidenty nebo poruchy, které by mohly mít vliv na bezpečnost nebo fungování AI systémů. To zahrnuje i situace, kdy dojde k narušení soukromí uživatelů. Společnosti musí být připraveny poskytovat informace a audity provozování AI systémů.

watsonx.governance

Sledování životního cyklu a AI faktové listy

Faktografické listy

- faktografické listy dokumentují **data o použití modelů**
- obsahují konkrétní hodnoty používaných modelů (metadata) a hodnoty volání včetně použitých promptů během využívání modelu

Správa životního cyklu modelu

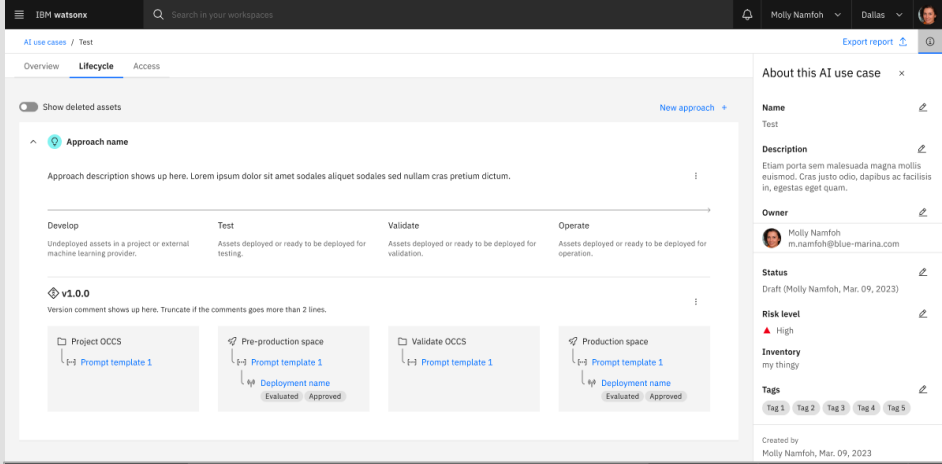
- Workflow procesy pokrývají celý životní cyklus modelu **od vývoje přes schvalování až po nasazení do produkce**
- Modely procházejí různými fázemi, jako je kandidátský proces, **schvalování, vývoj, validace a nasazení**

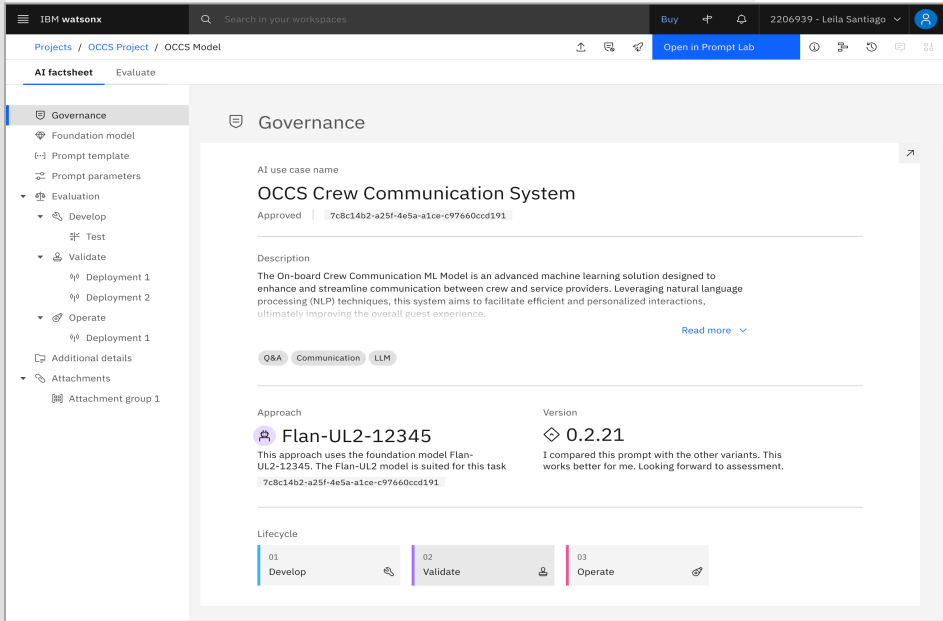
Správa verzí modelů

- Každý model může mít různé verze, které jsou sledovány a spravovány
- Změny a důvody volby konkrétní verze modelu jsou **zaznamenávány v auditní stopě**

Použití modelů v různých projektech a use cases

- Každý model může být použit v rámci **různých use cases**
- **Metriky a monitorování se definují pro každý use case zvlášť**





watsonx.governance

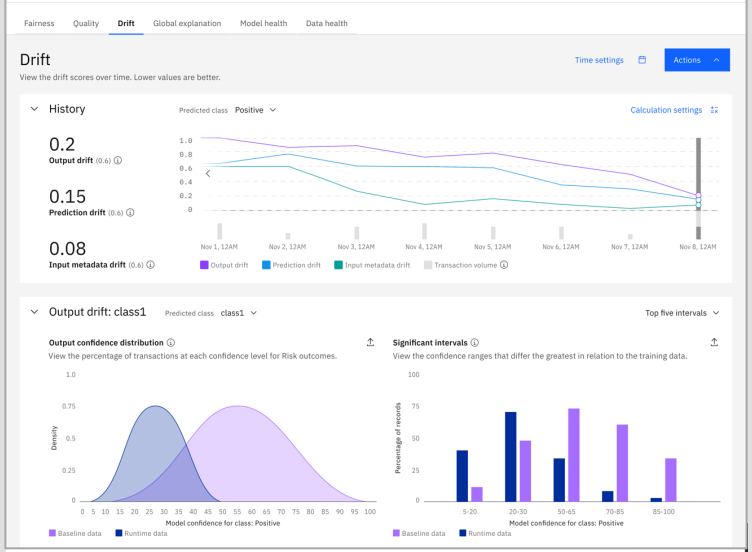
Monitorování a vyhodnocování

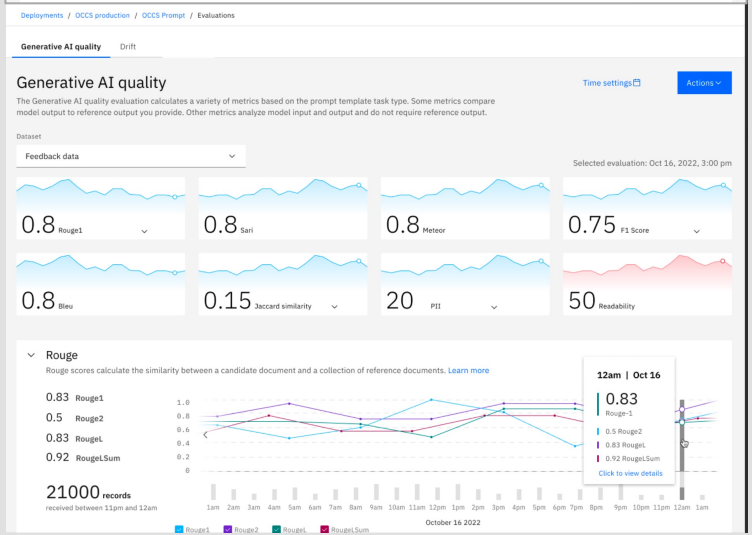
Monitoring

- sledování **kvality, čitelnosti a přesnosti LLM** pro běžné úlohy, jako je sumarizace, generování obsahu, otázky a odpovědi
- sledování vstupních promptů a výstupů vzhledem k personálním informací (PII) a použití toxického jazyka
- snadné sledování, zda model funguje v návaznosti na trénovací data, či se v průběhu času odchyluje. Při zjištění odchylky je možné vyvolat událost / upozornění.

Vyhodnocení

- **vyhodnocování a monitorování kvality, přesnosti a spravedlnosti modelu v průběhu celého životního cyklu - od vývoje až po nasazení**
- vyhodnocuje spravedlnost/objektivitu pro tradiční ML s automatickým odstraňováním zkreslení.
- upozorňuje uživatele při **překročení prahové hodnoty metriky**
- možnost pochopit zdroje / kontext, který se podílí na sestavení odpovědi pro RAG dotazovací systém.





watsonx.governance

Dodržování předpisů a regulace EU AI ACT

Watsonx.governance pro soulad s EU AI ACT:

- automatizovaná transformace regulačních požadavků do vymahatelných standardů a politik
- **dokumentace modelů, podpora auditovatelnosti modelů prostřednictvím AI faktů** (klíčová metadata modelů)
- proaktivní **identifikace a zmírnění rizik**, jako jsou předsudky a drift
- komplexní **správa životního cyklu AI** od předběžného nasazení až monitorování produkčního nasazení
- usnadňuje dodržování předpisů a podporuje auditovatelnost modelů prostřednictvím údajů o životním cyklů AI modelů

Řízení rizik

- platforma proaktivně identifikuje a zmírňuje potenciální rizika spojená s AI modely, včetně sledování předsudků (bias) a driftu.
- generuje **výstrahy, když metriky výkonu vybočují z předem definovaných prahových hodnot**, což organizacím umožňuje řešit problémy dříve, než eskalují

Předpřipravené nástroje a schopnosti

- předpřipravené panely zobrazující veličiny, pracovní postupy snadno přizpůsobitelné pro **implementaci regulačních standardů**
- interaktivní uživatelské rozhraní usnadňující práci i netechnickým uživatelům, **snadno přizpůsobitelné obchodním potřebám a procesům a potřebám regulátora**