# Bezpečnostní standardy (AI Safety)

Bezpečnostní standardy pro AI mají za cíl zajistit, aby systémy fungovaly spolehlivě, předvídatelně a bezpečně i v neočekávaných situacích. Na rozdíl od běžného softwaru je testování AI složitější kvůli jeho nedeterministické povaze.

## Klíčové pilíře AI Safety

### 1. Robustnost a odolnost (Robustness)

Systém musí být schopen správně fungovat i v případě, že se setká s daty, která nebyla v trénovací sadě, nebo pokud dojde k pokusu o jeho oklamání.

- **Adverzní útoky (Adversarial Attacks):** Malé, pro člověka neviditelné změny ve vstupu (např. šum v obrázku), které způsobí, že AI udělá fatální chybu.
- **Opatření:** Adverzní trénování (zahrnutí těchto chyb do tréninku) a formální verifikace kódu.

### 2. Interpretovatelnost a vysvětlitelnost (Interpretability)

Bezpečnost vyžaduje, abychom rozuměli tomu, proč model dospěl k danému závěru. U kritických systémů (např. v medicíně nebo autonomním řízení) je „černá skříňka“ nepřípustná.

- **Saliency Maps:** Vizualizace částí vstupu, které měly na rozhodnutí největší vliv.

### 3. Slučitelnost cílů (Alignment)

Problém zajištění toho, aby cíle AI byly v souladu s lidskými hodnotami a záměry. Nejde jen o to, co AI říkáme, aby dělala, ale jak to interpretuje v širším kontextu.

## Mezinárodní normy a rámce

V současné době vznikají standardy, které definují, jak by se mělo k bezpečnosti AI přistupovat na organizační úrovni:

| Standard | Název / Zaměření |
|---|---|
| **ISO/IEC 42001** | Systém managementu umělé inteligence (AIMS) - první mezinárodní certifikovatelný standard. |
| **NIST AI RMF** | AI Risk Management Framework - komplexní rámec pro řízení rizik od amerického institutu NIST. |
| **ISO/IEC 23894** | Návod na řízení rizik specificky pro organizace vyvíjející nebo používající AI. |
| **MITRE ATLAS** | Databáze taktik a technik, které útočníci používají proti AI systémům (obdoba ATT&CK). |

# Praktické techniky zajištění bezpečnosti

- **Red Teaming:** Simulované útoky na model s cílem najít slabiny, obejít filtry nebo vynutit škodlivý výstup (např. jailbreaking u LLM).
- **Human-in-the-loop (HITL):** Proces, kde lidský operátor reviduje klíčová rozhodnutí AI dříve, než jsou vykonána.
- **Sandboxing:** Provozování AI v izolovaném prostředí, aby v případě chyby nemohla napadnout kritickou infrastrukturu.
- **Monitorování driftu (Model Drift):** Neustálé sledování, zda se výkon modelu v čase nesnižuje kvůli změně reálných dat.

# Postupy pro bezpečný vývoj (SecDevOps pro AI)

```
1. **Zabezpečení dat:** Kontrola, zda trénovací data neobsahují toxický
obsah nebo "zadní vrátka" (backdoors).
2. **Bezpečné uložení vah:** Modely (váhy neuronové sítě) musí být šifrovány
a chráněny proti krádeži (exfiltraci).
3. **Omezení výstupu:** Implementace filtrů, které zabrání generování
nenávistných projevů nebo návodů k nelegální činnosti.
```

> **Základní pravidlo:** Bezpečnost AI není jednorázový úkon, ale kontinuální proces, který začíná sběrem dat a končí stažením modelu z provozu.

— **Související témata:**

- Etika a rizika umělé inteligence
- Legislativa a právo v AI
- Penetrační testování a Red Teaming