



Human-in-the-loop AI

Západočeská Univerzita V Plzni
Katedra Kybernetiky
Semestrální práce - HKUI

Trong Quoc Huy Dinh
31. semestr
23. prosince 2025

1 Human-in-the-Loop Artificial Intelligence

1.1 Definice Human-in-the-Loop umělé inteligence

Human-in-the-Loop (HITL) Artificial Intelligence označuje třídu systémů strojového učení, ve kterých lidská expertiza tvoří nedílnou součást procesu učení, vyhodnocování nebo samotného provozu systému. Na rozdíl od plně autonomních pipeline, kde se model po natrénování chová nezávisle na lidském zásahu, HITL systémy cíleně zapojují člověka do strategických bodů rozhodovacího procesu s cílem zvýšit přesnost, bezpečnost a schopnost adaptace systému [2].

Tento přístup vychází z předpokladu, že lidé a stroje disponují odlišnými, avšak vzájemně se doplňujícími schopnostmi. Zatímco modely umělé inteligence vynikají ve zpracování velkých objemů dat a v odhalování statistických vzorů, lidé přinášejí kontextové porozumění, schopnost rozhodovat v místech nejistoty a dovednost interpretovat nejednoznačné situace. HITL paradigma tedy nesměřuje k nahrazení člověka, ale k vytvoření spolupracujícího systému, kde se lidský úsudek a automatizované zpracování vzájemně doplňují.

Ačkoliv je pojem Human-in-the-Loop používán v různých oblastech umělé inteligence, jeho základní význam zůstává konzistentní. Vždy se jedná o situaci, kdy lidský faktor aktivně ovlivňuje chování nebo vývoj systému. Tato interakce může mít podobu ruční anotace dat, poskytování korektivní zpětné vazby, kontroly výstupů, omezování chování modelu nebo finálního rozhodování v případech s vysokými důsledky. HITL tak zahrnuje široké spektrum metodických přístupů, které se liší zejména načasováním, strukturou a účelem lidské intervence.

1.2 Terminologie a rozsah

Navzdory rozšířenému používání pojmu HITL nejsou jeho hranice v odborné literatuře jednoznačně vymezeny. Někteří autoři označují jako HITL pouze interaktivní scénáře, kde je člověk přímo zapojen do trénování modelu, zatímco jiní tento termín používají pro jakýkoliv systém, který vyžaduje lidský dohled nebo schvalování výstupů. Vhodným způsobem interpretace HITL je proto vnímání tohoto přístupu jako kontinua lidské účasti, které sahá od minimální anotace nebo zpětné vazby až po hlubokou integraci lidského rozhodování během celého životního cyklu systému.

Je rovněž důležité odlišit HITL od příbuzných pojmů, jako je human-centered AI nebo hybrid intelligence. Human-centered AI se primárně zaměřuje na uživatelskou přívětivost a interakci se systémem, zatímco hybrid intelligence zdůrazňuje kombinaci kognitivních schopností člověka a stroje. HITL se naproti tomu soustředí především na mechanismy, kterými člověk přímo ovlivňuje chování a vývoj modelů strojového učení.

1.3 Role člověka v HITL systémech

Míra a forma zapojení člověka v HITL systémech se liší podle cíle a kontextu dané aplikace. Lidskou roli lze obecně rozdělit do několika funkčních kategorií.

První kategorií jsou přispěvatelé dat, kteří vytvářejí, anotují nebo kurátorují trénovací datasety. Tato role je klíčová zejména v oblastech, kde jsou kvalitní popisky vzácné, obtížně formalizovatelné nebo silně závislé na kontextu.

Druhou kategorií jsou interaktivní učitelé nebo spolupracovníci, kteří během učení modelu poskytují zpětnou vazbu, opravují chování systému a tím aktivně formují jeho učební trajektorii.

Další roli představují hodnotitelé a dohlížitelé, kteří posuzují chování modelu, ověřují jeho výstupy a rozhodují o připravenosti systému k nasazení.

V citlivých doménách pak člověk často zůstává finálním rozhodovatelem. Modely umělé inteligence zde slouží jako podpůrný nástroj, zatímco odpovědnost za konečné rozhodnutí nese lidský expert. Tyto role se mohou v praxi překrývat a jeden člověk může plnit více funkcí v různých fázích vývoje systému.

1.4 Klíčové charakteristiky HITL

Navzdory rozdílům mezi jednotlivými aplikacemi lze u HITL systémů identifikovat několik společných charakteristik. Jednou z nich je iterativní zlepšování, kdy je lidský vstup využíván k postupnému zpřesňování modelu prostřednictvím opakovaných cyklů učení a zpětné vazby.

Dalším znakem je selektivní lidská intervence. Člověk není zapojen neustále, ale zejména v situacích, kdy systém čelí nejistotě, nejednoznačnosti nebo vysokému riziku chyby.

HITL rovněž staví na principu komplementarity, kdy kombinace lidských a strojových schopností vede k výsledkům, kterých by ani jedna strana samostatně nedosáhla.

Významnou vlastností HITL je také podpora vysvětlitelnosti a transparentnosti. Lidská interakce se systémem přirozeně vytváří tlak na srozumitelnost chování modelu, což usnadňuje jeho pochopení a přijetí uživateli.

1.5 Srovnání s plně autonomními systémy

Plně autonomní systémy umělé inteligence předpokládají, že po natrénování mohou fungovat spolehlivě bez další lidské účasti. Tento přístup je vhodný především pro stabilní a dobře definovaná prostředí s dostatkem reprezentativních dat.

V reálných podmínkách však často dochází k situacím, které nebyly při trénování zohledněny, například k výskytu vzácných událostí nebo nejednoznačných vstupů. HITL systémy tyto problémy řeší zapojením lidského úsudku do klíčových rozhodovacích bodů. Tím zvyšují robustnost systému a snižují rizika spojená s čistě automatickým rozhodováním.

2 Aplikace Human-in-the-Loop Artificial Intelligence

Human-in-the-Loop přístupy jsou stále častěji využívány v reálných systémech, kde autonomní modely čelí nejistotě, proměnlivým podmínkám nebo vysokým nárokům na odpovědnost. Konkrétní podoba lidského zapojení se liší podle domény, společným jmenovatelem však zůstává snaha zvýšit robustnost, správnost a důvěryhodnost systému.

Jednou z hlavních aplikačních oblastí je interaktivní anotace a kurace dat. Aktivní učení umožňuje modelům vybírat vzorky, které jsou pro člověka nejvíce informativní, čímž se zvyšuje kvalita datasetu při nižších nákladech na manuální práci [2], [3].

Významné uplatnění nachází HITL také v posilovaném učení. Lidská zpětná vazba zde pomáhá stabilizovat trénink a usměrňovat chování agentů směrem k bezpečným a kontextově vhodným strategiím.

Další důležitou oblastí jsou systémy pro podporu rozhodování v citlivých doménách, jako je zdravotnictví, právo nebo finance. Modely umělé inteligence zde slouží jako nástroj pro analýzu a doporučení, zatímco člověk interpretuje výsledky a nese odpovědnost za finální rozhodnutí.

HITL přístupy se uplatňují také při detekci a zmírňování zkreslení v datech a modelech. Nástroje jako D-BIAS ukazují, že lidský dohled je nezbytný

pro identifikaci neférových kauzálních vztahů [1].

V dynamických prostředích, například při moderaci obsahu nebo detekci podvodů, umožňuje HITL reagovat na nové a nečekané situace. Lidské zásahy v těchto případech slouží nejen k okamžité korekci, ale také jako nové trénovací signály, díky nimž se systém může postupně adaptovat na měnící se podmínky.

Odkazy

[1] B. Ghai, N. Alverio et al., "D-BIAS: A Human-in-the-Loop Tool for Detecting and Mitigating Bias in Tabular Data," *arXiv preprint arXiv:2208.05126*, 2022.

[2] E. Mosqueira-Rey, D. Alonso-Bartolomé et al., "Human-in-the-loop artificial intelligence: A review of key approaches," in *2022 International Conference on Data Science*, 2022.

[3] R. Pandey, H. Purohit, C. Castillo a V. Shalin, "Modeling and Mitigating Human Annotation Errors in Hybrid Stream Processing Systems with Human-in-the-Loop Machine Learning," *International Journal of Human–Computer Studies*, roč. 165, s. 102 772, 2022.