Umělá inteligence a lidská práva: rizika, příležitosti a regulace

Výzkumný projekt TAČR č. TL05000484

Umělá inteligence a lidská práva:

Hodnocení rizik pro lidská práva v životním cyklu AI systému

Ambis.Vysoká škola.

MUNI Masarykova univerzita

ČVUT ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

prg.ai

Hodnocení rizik pro lidská práva v životním cyklu AI systému: Soubor doporučení

Vstupní analýza AI systému				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze AI životního cyklu	Možnosti eliminace rizika při vývoji systému	Možnosti eliminace rizika za provozu
Systém bude provozován v doméně s vazbou na LP* **Systém bude při vývoji používat data s vazbou na LP, případně data citlivá na bias s dopadem do LP** **Byly identifikovány požadavky související s LP**	Při vstupní analýze je zjištěna vazba AI systému na LP problematiku. Systém bude například vyhodnocovat osobní údaje, počítat výši půjčky, identifikovat osoby na základě obrazu a jiných biometrických hodnot, hodnotit účastníky výběrového řízení či predikovat pravděpodobnost recidivy a další. Pokud se vazba na LP problematiku určí pouze indikativně, například osobou, která není odborníkem v oblasti LP, musí následovat expertní rozpracování odborníkem v rámci vstupní analýzy.	- Možný dopad do všech fází	Rozpracováno pro specifická rizika níže.	Rozpracováno pro specifická rizika níže.

Vstupní data, strojové učení				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze AI životního cyklu	Možnosti eliminace rizika při vývoji systému	Možnosti eliminace rizika za provozu
Bias v datech	Bias z pohledu LP je neoprávněná či nevyvážená přítomnost tzv. chráněné hodnoty v datech, tj. nesoucí informaci o pohlaví, rase, barvě pleti, náboženství,	- Možný dopad do všech fází	Uživatelské požadavky: - Reprezentativní a úplná vstupní vývojová datová sada, pokud nese informaci o tzv. chráněné hodnotě.	Systém za provozu bude transparentní pro uživatele a bude kromě výstupu poskytovat také informace, na základě

	politické příslušnosti apod. Došlo by tak k protiprávní diskriminaci. Příkladem biasu v datech v oblasti bankovnictví je nezahrnutí skupiny osob v určitém příjmovém rozsahu při přípravě systému pro ohodnocování možné výše poskytované půjčky, pokud je daná skupina kvalifikována právě jednou z chráněných hodnot. Mohlo by tak dojít k diskriminaci dané skupiny. Může se jednat jak o bias ve vývojových datech používaných například v procesu strojového učení, tak v datech používaných pro verifikaci, validaci a kvalifikaci systému. Při analýze rizik bereme v úvahu: - Je tzv. chráněná hodnota přítomna vývojových datech? - Je její přítomnost oprávněná? - Pokud je oprávněná, je zajištěna reprezentativnost a úplnost dat nesoucí tuto chráněnou hodnotu? - Analyzovali jsme takové riziko i u testovací sady?		- Tzv. chráněná hodnota je eliminována, pokud její přítomnost v datech není oprávněná. Požadavky na přípravu dat: - Vývojová trénovací a testovací data musí pokrývat celou úlohu a musí obsahovat reprezentativní a úplnou vstupní datovou sadu ve vztahu k přítomné chráněné hodnotě. Příkladem je systém rozpoznávání obličeje, jehož data musejí být připravena tak, aby pokrývala veškeré etnické skupiny, rasy a barvy pleti, pokud bude provozován globálně. - Popis/anotace dat není zkreslený a je konzistentní přes celou datovou sadu. Může být vyžadována nezávislá revize.	jakých vstupních dat k tomuto výstupu dospěl a s jakou vahou. Systém bude poskytovat data pro své monitorování. Na jeho základě je vyhodnocován soulad s LP. Vyhodnocování může probíhat průběžně za chodu, ve stanovených časových intervalech, automatizovaně, pomocí lidské obsluhy nebo formou pravidelných auditů systému. Způsob je dán požadavky na systém a jeho provozování.
Použití syntetických vývojových dat	Syntetická data se používají zejména pro doplnění vývojových dat, pokud není k dispozici dostatečné množství reálných dat nebo není časově možná všechna reálná data nasbírat. Příkladem je doplnění obrázků obličeje o uměle generované obličeje etnika, které v reálných snímcích chybí.	- Příprava vývojových dat - Verifikace, validace, kvalifikace - Provozní fáze	Požadavky na přípravu dat: - Zvolení správného poměru reálných a syntetických dat, aby byla pokryta řešená úloha. - Ověření, zda syntetická data neobsahují bias, viz rizika identická jako u biasu.	Stejné možnosti jako u rizika „Bias v datech“.

	Syntetická data se nepoužívají jako plnohodnotná náhrada reálných dat a jejich použití má své limity vycházející z komplexity řešené úlohy. V případě komplexní úlohy i syntetická data mohou obsahovat ostatní zde uvedená LP rizika. Syntetická data nejsou zpravidla vhodná pro pokrytí celé úlohy. Pokud by bylo možné celou úlohu pokrýt syntetickými daty, její algoritmické řešení je známé a není třeba používat metody strojového učení k jejímu řešení.		Další možnosti eliminace jsou stejné jako u rizika „Bias v datech“.
AI systém není vysvětlitelný	Jedná se o algoritmickou vysvětlitelnost AI systému. Nevysvětlitelná AI se chová jako černá skříňka: jde o AI modely, které jsou příliš složité na to, aby je člověk mohl jednoduše interpretovat. Tyto systémy jsou většinou založeny na technikách strojového učení – příkladem jsou neuronové sítě, které jsou v dnešní době široce používaným paradigmatem pro úlohy strojového učení a představují tak primární zdroj modelů černých skříněk. U nevysvětlitelné AI výrobce není schopen zaručit vysvětlení, jakým způsobem AI systém došel ke svému rozhodnutí/výstupu. Naopak u vysvětlitelných modelů lze čtením modelu vidět, na základě jakých informací a jakým způsobem systém dospěl k rozhodnutí/výstupu.	- Možný dopad do všech fází	Pokud bude vyžadován systém, u kterého bude vždy možné říci, jakým způsobem dospěl na základě vstupních dat ke svému výstupu, není možné použít systém nevysvětlitelný, který představuje black box. Nevysvětlitelnost systému nelze eliminovat, lze pouze snížit možné dopady tohoto rizika a to následujícím způsobem. Uživatelské požadavky: - Zavedení požadavků na operační transparentnost systému pro uživatele. Předání systému do provozu: - Provedení certifikace systému. Způsob certifikace závisí na podmínkách a požadavcích provozování systému. Může se	Eliminace rizika nevysvětlitelnosti za provozu není možná, lze pouze snížit možné dopady tohoto rizika a to následujícím způsobem: Systém za provozu bude transparentní pro uživatele, tj. kromě výstupu bude poskytovat také informace, na základě jakých vstupních dat a s jakou vahou k tomuto výstupu dospěl. Systém bude poskytovat data pro své monitorování. Na jeho základě je vyhodnocován soulad s LP. Vyhodnocování může probíhat průběžně za chodu, ve stanovených časových intervalech, automatizovaně, pomocí lidské obsluhy nebo formou pravidelných auditů systému. Způsob je dán

	Příkladem je vysvětlitelný doporučovací AI systém využívaný ve výběrovém řízení do zaměstnání, který doporučí kandidáta na pracovní místo, pokud je jeho předchozí praxe delší pěti let. U systému, který není vysvětlitelný a který je natrénovaný pro obdobnou funkcionalitu takové čtení modelu není možné, žádné takové lidsky srozumitelné rozhodovací postupy v něm nelze nalézt.		jednat o certifikace od výrobce, přes odběratele a třetí stranu až po certifikace od státem pověřené entity.	požadavky na systém a jeho provozování.
Dotrénování systému s použitím provozních dat	Například systémy rozpoznávání řeči používají reálné promluvy nasbírané za provozu systému ke svému zlepšování, tedy k dotrénování svých modelů. Hlavní riziko spočívá ve volbě správné metody řízeného dotrénování a zajištění odpovídajících trénovacích dat sbíraných za provozu systému. Hrozí zavedení biasu a posun již natrénovaného systému tak, že nebude ve shodě s LP. Dalším rizikem je ovlivnění přesnosti systému a jeho věrohodnosti. Při dotrénování za provozu může být systém i manipulován, čemuž je nutné předcházet.	- Možný dopad do všech fází	Uživatelské požadavky: - Definování podmínek sběru dat a jejich vlastností za provozu pro zamezení biasu a posunu. - Definice a kontrola očekávané kvality systému po dotrénování. - Uvedení způsobu dotrénování, zda půjde o mechanismus automatizovaného dotrénování přímo za provozu nebo sběr dat a dotrénování výrobcem při aktualizaci systému. - Zavedení monitorování systému za provozu. Požadavky na přípravu dat: - Definice spolehlivého způsobu výběru/filtrace a kontroly sbíraných dat pro dotrénování. Technická specifikace:	Monitorování a kontrola kvality systému za provozu.

			- Zajištění odolnosti systému proti biasu.	
Nejsou známé a definované okrajové/hraniční vlastnosti systému a systém je na nich provozován	Například u bankovního systému pro ohodnocování kredibility klienta se nepočítalo s tím, že bude nasazen v lokalitě s velkou mírou rizika ztráty zaměstnání nebo naopak v oblasti s velkým výskytem startup firem, které vytváří v krátkém časovém horizontu majetné klienty. I u pečlivě připravovaných vývojových dat nemusí být plně pokryty veškeré hraniční případy řešené úlohy nebo vzácné výskyty určitého jevu. Rizikem je neznalost hranic, za kterých systém ještě funguje správně a kdy naopak i s malou odchylkou ve vstupních provozních datech bude poskytovat nesprávné výsledky.	- Verifikace, validace, kvalifikace - Uvedení systému do provozu - Provozní fáze	Verifikace, validace, kvalifikace: - Systém musí být ověřován i ve svých okrajových případech použití. Pro tyto případy musí být doplněny testovací sady. Předání systému do provozu: - Certifikace systému před uvedením do provozu s cílem zajistit shodu s provozními podmínkami a splnění požadavků na LP. Způsob certifikace závisí na podmínkách a požadavcích provozování systému. Může se jednat o certifikace od výrobce, přes odběratele a třetí stranu až po certifikace od státem pověřené entity.	Monitorování systému za provozu.

Požadavky na data používaná při strojovém učení jsou stanovena natolik striktně, že neumožní vývoj systému bez biasu a posunu	Příkladem může být situace, kdy by u požadavků na data používaná pro strojové učení bylo striktně trváno na použití anonymizovaných dat, např. anonymizovaných registračních značek vozů v případě trénování rozpoznávání právě těchto značek. Pak nelze úlohu učení úspěšně realizovat a její funkce nebude odpovídat realitě. Pokud se tento problém nebude řešit již ve vstupní analýze, později způsobí daleko větší náklady při vývoji.	- Vstupní analýza AI systému - Příprava vývojových dat	Vstupní analýza AI systému: - Identifikace rizika a definice opatření ochrany dat, která není v kolizi s vývojem systému. Příprava vývojových dat: - Zavedení opatření ochrany dat při současném zajištění jejich úplnosti pro vývoj.	Neexistuje specifická možnost eliminace.

Integrace AI systémů třetích stran, rozšíření/úprava/oprava stávajícího systému, certifikace				
Riziko	Příklad/ Vysvětlení rizika	Dotčené fáze AI životního cyklu	Možnosti eliminace rizika při vývoji systému	Možnosti eliminace rizika za provozu
Integrace AI systému třetí strany při vývoji, zahrnuje transfer learning	Riziko se týká situací, kdy je přebírán již hotový AI systém od externího dodavatele a integrován do výsledného řešení. Rizika spočívají v neznalosti vývojové datové sady, způsobu ověřování přebíraného systému a požadavků, za kterých byl systém vyvíjen. Většinou není k dispozici relevantní analýza dopadů systému na LP spojená s očekávaným nasazením systému. Rizikem jsou i případné certifikace systému s ním dodávané, zejm. pokud nejsou vydány ověřenou autoritou či nejsou plně	- Verifikace, validace, kvalifikace - Uvedení systému do provozu	Verifikace, validace, kvalifikace: - Pokud není zajištěna validní certifikace systému, přistupuje se k přetestování systému třetích stran. Uvedení systému do provozu: - Lze požadovat doložení certifikace. - Ověření validnost stávající certifikace pro očekávané nasazení systému. - Provedení nové recertifikace systému.	Neexistuje specifická možnost eliminace.

	známé podmínky, za kterých k certifikaci došlo.				
Certifikace systému	Riziko souvisí zejména s provozováním systému a dopadá na provozovatele. Způsobů certifikace z hlediska LP problematiky může být celá řada a provozovatel systému musí zvolit adekvátní k možným dopadům provozu na LP, případně se musí řídit povinnou certifikací, pokud existuje. Obecně může certifikovat každý z aktérů, který se účastní životního cyklu AI systému, a to s různou mírou autority certifikátu. Za nejnižší formu certifikace lze považovat certifikát vydaný přímo výrobcem systému, za nejvyšší pak certifikaci provedenou regulačním orgánem zřízeným státem.	- Uvedení systému do provozu	Uvedení systému do provozu: - Stupeň certifikace odpovídající rizikům spojeným s provozováním systému a případným dopadům na LP. - Recertifikace systému od zvolené autority. - Ověřit parametry certifikace vzhledem k provozním podmínkám a požadavkům na systém.	Neexistuje specifická možnost eliminace.	
Výrobce nezahrnul do vývoje požadavky regulátora, certifikační autority nebo doporučené postupy, pokud existují	Riziko nastává, pokud si výrobce nebo odběratel/provozovatel systému není vědom těchto specifických požadavků a nezahrnul je do analýzy systému. Může se jednat například o požadavek testování systému na konkrétní testovací sadě dat, kterou spravuje příslušný úřad. Riziko souvisí zejména s provozováním systému a dopadá na provozovatele.	- Uvedení systému do provozu	Uvedení systému do provozu: - Stupeň certifikace odpovídající požadavkům regulátora, certifikační autority nebo minimálně shoda s doporučenými postupy vyžadované odběratelem/provozovatelem. - Recertifikace systému od zvolené autority.	Neexistuje specifická možnost eliminace.	

Je provedeno rozšíření nebo oprava AI systému	Typickou situací je aktualizace systému, rozšíření funkcionality nebo odstranění chyby v systému. Při tomto zásahu může dojít ke změně parametrů systému, biasu a posunu kvůli rozšíření systému nebo vzniku nových rizik (ve smyslu rizik uvedených v této analýze) a to díky novým funkcionalitám.	- Možný dopad do všech fází	Opakuje se celý vývojový cyklus včetně testování a uvedení do provozu. Před jeho začátkem je nutné provést analýzu rizik, identifikovat ta, která jsou pro úpravu systému platná, a řídit se postupy danými pro příslušná rizika.	Neexistuje specifická možnost eliminace.

Požadavky na AI systém				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze AI životního cyklu	Možnosti eliminace rizika při vývoji systému	Možnosti eliminace rizika za provozu
LP požadavky jsou zadány vágně a nejsou měřitelné a testovatelné	Riziko vágně zadaných požadavků v LP oblasti je vysoké, což je dáno vstupem AI do oblastí lidské činnosti, které zatím nebyly automatizovány. Vágnost požadavků může být způsobena nedostatečnou analýzou požadavků na systém či i obtížností převodu LP definic do technických parametrů. Navazujícím problémem je testovatelnost těchto vágních požadavků. Testy musí pokrývat celý testovaný případ a musí být konkrétní a měřitelné, což je u vágně definovaných požadavků těžké splnit. Příkladem je chráněná hodnota „barva pleti“, kde není v technické oblasti jednoznačně zadefinováno, jak je tato hodnota reprezentována například v obrazových datech.	- Vstupní analýza AI systému - Produktové a uživatelské požadavky - Technická specifikace a požadavky - Verifikace, validace, kvalifikace	Vstupní analýza: - Vágnost, a zní vyplývající testovatelnost, požadavků na LP systém musí být zachycena a odstraněna, za účasti AI odborníků. Návazné fáze: - Je nutné udržovat konkrétnost, měřitelnost a testovatelnost, požadavků. Pokud bude po vstupní analýze riziko vágnosti požadavků a specifikací vyhodnoceno jako přetrvávající a neodstranitelné, je nutné zvážit eliminaci zavedením požadavku na operační transparentnost systému.	Plná eliminace rizika za provozu není možná, lze však použít metody snížení jeho dopadů na LP a to následovně: - Systém za provozu bude transparentní pro uživatele. - Systém bude poskytovat data pro své monitorování.

	Podobně by působil požadavek, že „systém nemá pracovat s chráněnou hodnotou“, aniž by byla tato chráněná hodnota definována technickou specifikací.			
Nejsou identifikovány všechny LP uživatelské požadavky	Riziko nastává zejména v případě, kdy nejsou ke tvorbě požadavků na systém přizvány všechny klíčové osoby, kdy existuje malá zkušenost s vývojem a nasazením systému pro příslušnou doménu a také v případě neznalosti příslušných regulací, certifikačních postupů nebo doporučení dobré praxe. Příkladem je potřeba transparentnosti systému, který bude nasazen jako asistivní technologie v soudnictví. Pokud požadavek transparentnosti není zachycen, systém nebude poskytovat podpůrné informace uživateli (soudci), čímž oslabí či znemožní jeho schopnost v rámci svého rozhodování detekovat případnou chybu v predikci systému vedoucí k porušení LP.	- Vstupní analýza AI systému - Produktové a uživatelské požadavky	Pokud bude ve vstupní analýze riziko nekompletnosti požadavků na systém vyhodnoceno jako reálné a neodstranitelné, je nutné zvážit eliminaci pomocí opatření shodných s rizikem „Vysvětlitelnost“.	Obdobné jako u rizika „LP požadavky jsou zadány vágně a nejsou měřitelné a testovatelné“.
LP uživatelské požadavky jsou nereálné	Příkladem nereálných požadavků může být požadavek vytvoření plně autonomního systému tam, kde vzhledem ke složitosti rozhodovacího procesu má být nasazena asistivní technologie, podporující rozhodování lidského odborníka – typicky oblast soudnictví a zdravotnictví. Obecně, pokud je činnost řešena člověkem s určitou mírou chybovosti a tuto činnost hodláme automatizovat, je nutné nejprve	- Vstupní analýza AI systému	Je nutné zvážit eliminaci pomocí opatření shodných s rizikem „Vysvětlitelnost“.	Obdobné jako u rizika „LP požadavky jsou zadány vágně a nejsou měřitelné a testovatelné“.

	provést analýzu, zda je tato chybovost automatizací odstranitelná nebo se jedná o hlubší problém, který bude přetrvávat i po automatizaci příslušné činnosti.			
Provozování AI systému				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze AI životního cyklu	Možnosti eliminace rizika při vývoji systému	Možnosti eliminace rizika za provozu
Netransparentní AI systém	Transparentnost AI systému zahrnuje vývojovou a operační transparentnost. Vývojová transparentnost slouží především aktérům schvalujícím systém do provozu. Týká se vývojového procesu a umožňuje jeho kontrolu, vč. kvality. Operační transparentnost naopak vyjadřuje schopnost AI systému poskytnout informaci o tom, která vstupní data přispěla k výslednému výstupu systému a s jakou vahou. Z hlediska provozu je rozhodování pro uživatele transparentní a jeho výstupy lze ověřovat. Naopak u netransparentního systému uživatel tyto informace nemá a zná pouze výstup/výsledek. Pozn.: Transparentnost je třeba odlišit od algoritmické vysvětlitelnosti. Ta se týká způsobu algoritmické interpretace znalostí, tj. vysvětlitelnosti algoritmického zápisu znalostí. Pro uživatele může být i algoritmicky vysvětlitelný systém	- Produktové a uživatelské požadavky - Technická specifikace a požadavky - Verifikace, validace, kvalifikace - Provozní fáze	Uživatelské požadavky: - Operační transparentnost. Technické specifikace: - Technické řešení transparentnosti systému.	Lze eliminovat pouze zavedením systému, který bude za provozu pro uživatele transparentní.

	netransparentní, pokud neposkytuje informace, na základě čeho se rozhodl.			
AI systém není za provozu monitorovatelný	Nesoulad AI systému s LP lze detekovat za provozu pomocí jeho monitorování. Pokud není systém pro své monitorování připraven, nelze zpětně provádět ani jeho audit, analyzovat jeho výstupy, detekovat nesoulad a zjišťovat jeho příčinu. Monitorování AI systému za provozu přispívá k jeho transparentnosti.	- Produktové a uživatelské požadavky - Technická specifikace a požadavky - Verifikace, validace, kvalifikace - Provozní fáze	Uživatelské požadavky: - Systém bude umožňovat monitorování, tj. bude prováděn záznam informací pro monitorování systému. Technické specifikace: - Technické řešení monitorování systému.	Systém podporuje monitorování, je monitorován a výstupy monitorování jsou vyhodnocovány.
Systém je určen pro jiné než cílové provozní prostředí	Systém je nasazen v jiném cílovém prostředí, než bylo plánováno. Může se jednat např. o rozšíření okruhu uživatelů produktu, přičemž v daném prostředí nevyhovuje požadavkům na LP, či může dojít k instalaci systému v jiné zemi či regionu, kde existuje odlišný přístup nebo zákonné úpravy. Toto riziko může nastat i v rámci jedné firmy působící globálně. Další případem může být situace, kdy se v cílovém prostředí nasazení produktu změnil přístup k LP problematice a systém přestal být ve shodě s LP. Okrajové vlastnosti jsou řešeny samostatně v riziku uvedeném výše.	- Vstupní analýza AI systému - Produktové a uživatelské požadavky - Uvedení systému do provozu	Vstupní analýza AI systému: - Analýza podmínek provozování systému. Produktové a uživatelské požadavky: - Systém podporuje monitorování. Uvedení systému do provozu: - Ověření podmínek provozního nasazení, případně provedení certifikace systému pro provozní prostředí.	Systém podporuje monitorování, je monitorován a výstupy monitorování jsou vyhodnocovány. Provádí se pravidelný audit na shodu s LP problematikou v místě provozování systému.

Uživatele nebo provozovatel nepoužívá systém správně	Nesprávné používání systému uživatelem/provozovatelem může mít vícero variant a jejich kombinací: - systém je využíván i pro úlohy, pro které nebyl určen; - asistivní systém je využíván jako plně autonomní; - uživatel/provozovatel systému předkládá vstupy, pro které nebyl systém připraven; - výstupy monitorování systému nejsou správně interpretovány. Příčinou je většinou neznalost a nesprávné proškolení uživatele a provozovatele.	- Uvedení systému do provozu - Provozní fáze	Uvedení systému do provozu: - Proškolení provozovatele a uživatele. - Udělení certifikace provozovateli či uživateli o schopnosti používat systém v souladu s jeho určením.	Průběžná školení nových uživatelů a pro nové verze systému.
Oblast s vysokou rizikovostí porušení LP za provozu	Vzhledem ke komplexnosti LP problematiky, novým oblastem nasazování AI systémů, a tím i malé zkušenosti s jejich provozováním pro tyto nové činnosti, existuje vyšší riziko, že systém za provozu způsobí porušení LP. Při eliminaci tohoto rizika je i nutné zvážit míru škody. To by mělo být součástí vstupní analýzy systému.	- Možný dopad do všech fází	Vstupní analýza AI systému: - Vyhodnocení míry rizika, že dojde při provozování systému k narušení LP. - Zařazení systému do kategorie závažnosti dopadů do LP. Nevyšší kategorií mohou být například systémy „human-rights critical“, tj. AI technologie vyznačující se vysokým rizikem porušení LP s vážným dopadem na člověka. Výrobce systému dle analýzy zahrne do vývojové fáze relevantní možnosti eliminace rizika, která vybere ze všech zde uváděných rizik.	Dle míry rizika a vyhodnocení dopadů se do provozní fáze vyberou relevantní způsoby eliminace rizika. Výběr je možné provést z alternativ uvedených u všech zde analyzovaných rizik, případně budou takové způsoby předepsány regulátorem nebo certifikační autoritou.

AI systém bude napaden a zneužit	Riziko se ošetřuje stejným postupem jako u všech IT a software systémů, tj. analýzou bezpečnosti systému, implementací bezpečnostních prvků, vzdálené správy za provozu, zajištěním provozní infrastruktury proti napadení apod. Toto riziko nemá přímo vztah k AI technologiím jako takovým, jedná se obecně o přístupy, které jsou obvyklé pro vývoj a nasazení jakékoli informační technologie, která je tomuto riziku vystavena. Nejsou zde tedy uváděné konkrétní příklady dopadu rizika do jednotlivých fází a postačuje odkaz na standardní metody implementace bezpečnostních opatření do IT technologií. Patří sem i situace, kdy je systém používán za nelegálním účelem.	- Možný dopad do všech fází	Obvyklé metody a postupy implementace bezpečnostních opatření u IT technologií v návaznosti na vstupní analýzu systému.	Obvyklé metody a postupy implementace bezpečnostních opatření u IT technologií v návaznosti na vstupní analýzu systému.
Systém bude provozován v plně autonomním režimu	Jde o situaci, kdy je AI systém používán jako plně autonomní systém a není tedy primárně určen jako asistivní ke zlepšení práce odborníka v dané oblasti nasazení, tj. řeší úlohy bez přispění lidského činitele. Na rozdíl od asistivních AI systémů tak jeho výstupy nebudou přehodnocovány doménovým expertem. S tím souvisí riziko nedostatečné, resp. nemožné zpětné vazby při provozování autonomních systémů. V procesu, kde je zapojen člověk, je obvyklé získávat zpětnou vazbu, klást otázky či vznášet námitky na provedenou činnost, rozhodnutí či výstupy, což zamezí případným nedostatkům a	- Možný dopad do všech fází	Obdobné jako v případě rizika: „Oblast s vysokou rizikovostí porušení LP za provozu“.	Obdobné jako v případě rizika: „Oblast s vysokou rizikovostí porušení LP za provozu“.

	chybám, popř. umožní jejich nápravu při nasazení. Důležitá je vstupní analýza možných dopadů na LP a vyhodnocení provozních rizik. Pokud bude systém vyhodnocen jako „human-rights critical“, pak platí maximální možná opatření pro eliminaci rizika. S nižší mírou dopadů na LP budou opatření úměrně redukována. Doporučené minimální požadavky by se měly týkat: - zajištění transparentnosti systému; - zajištění monitorování systému; - certifikace systému před uvedením do provozu.			
Status AI systému jako „vyšší autority“ pro uživatele	V očích lidského účastníka procesu, především uživatele, může být AI systém považován za prvek vyšší autority, jejíž závěry není nutné přehodnocovat či podrobovat kritickému přezkumu. Může to být dáno lidskou důvěrou v bezchybnost systému. Riziko spočívá v nekritickém přebírání výstupů systému i tam, kde je očekáváno jejich odborné posouzení člověkem. Týká se především asistivních AI systémů, viz například „autonomní“ řízení automobilů, které není plně autonomní, což některé řidiče neodrazuje od svěření plné kontroly nad řízením automatu.	- Produktové a uživatelské požadavky - Uvedení systému do provozu - Provozní fáze	Produktové a uživatelské požadavky: - Transparentnost systému. - Monitorování systému za provozu. Uvedení systému do provozu: - Proškolení provozovatele a uživatele. - Udělení certifikace provozovateli či uživateli o schopnosti používat systém v souladu s jeho určením.	Průběžná školení uživatelů.Podpora pro rychlé a efektivní odvolání proti výstupům. Umožnění zpětné analýzy činnosti systému monitorováním. Transparentnost systému.

** LP = lidská práva*

Projekt TL05000484 spolufinancovaný se státní podporou TAČR v rámci Programu ÉTA