

Modul 1

Úvod do
umělé inteligence a
etiky v inženýrské praxi



Co-funded by
the European Union



Ethical Engineer

www.ethicalengineer.eu

OBSAH

01 Etika a inženýrství

02 Spravedlnost

03 Autonomie

04 Transparentnost

05 Neškodit (princip Non-maleficence)

06 Skupinová činnost

07 Kvíz



Co-funded by
the European Union

Tento projekt byl financován s podporou Evropské komise. Za tuto publikaci (sdělení) nese výlučnou odpovědnost autor a Komise nenese žádnou odpovědnost za jakékoli použití informací v ní obsažených. V souladu s novým rámcem GDPR upozorňujeme, že partnerství bude zpracovávat vaše osobní údaje výhradně v zájmu a za účelem projektu a bez jakéhokoli dotčení vašich práv.



Představení modulu

Cíl

Představit etické uvažování jako součást praxe v oblasti umělé inteligence a robotiky.

Cílová skupina

Studenti vysokých škol v oborech inženýrství, informatiky, robotiky a souvisejících oborech STEM.

Doporučená délka

Základní varianta 180 minut. Rozšířená varianta 225 minut. Lze realizovat jako jednu lekci nebo dvě kratší lekce.

[Zdroje] Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení (vývoj OER), str. 61–70.

Představení modulu

Vzdělávací sekvence

Vysvětlit etiku v inženýrské praxi, uplatnit zásady spravedlnosti, autonomie, transparentnosti a non-maleficence (princip neublížování), následně analyzovat inženýrský případ a odůvodnit rozhodnutí.

Výstupy studenta

Mapa etických otázek, analýza případu, ochranná opatření, záznam etického rozhodování v inženýrské praxi a kontrola výsledků. Tyto výstupy ukazují, jak etické uvažování mění požadavky, testy, dohled a rozhodnutí o schválení.

Výstupy učení

Na konci modulu budou studenti schopni:

1. Vysvětlit, jak se autonomie, spravedlnost, transparentnost a non-maleficence vztahují k umělé inteligenci a robotice.
2. identifikovat etická rizika v rámci formulace problému, dat, modelu, rozhraní, nasazení a monitorování.
3. analyzovat inženýrský případ a propojit rozhodnutí o návrhu s dopady na zainteresované strany,
4. navrhnout technická a řídicí opatření, která snižují riziko a podporují lidský dohled.
5. a odůvodnit inženýrské rozhodnutí s odkazem na Etické pokyny EU pro důvěryhodnou umělou inteligenci a příslušné kompetence DigComp.

[Zdroje]

Schválená žádost projektu Ethical Engineer KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://publications.jrc.ec.europa.eu/repository/handle/JRC128415>

Etické uvažování patří do inženýrského návrhu

01

Formulace
problému

02

Data

03

Model

04

Rozhraní a dohled

05

Nasazení

Technická rozhodnutí ovlivňují etické výsledky v průběhu celého životního cyklu systému umělé inteligence:

Etické uvažování je součástí kompetencí v oblasti návrhu. Pomáhá inženýrům převádět evropské hodnoty do technických požadavků, testovacích kritérií a zodpovědných rozhodnutí

[Zdroje] - Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení (vývoj OER), str. 61–70- <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

○ Formulace problému

Čí potřeby mají přednost?

○ Data

Kdo je zastoupen, kdo chybí a kdo je vystaven riziku?

○ Model

Jaké chyby a kompromisy jsou akceptovány?

○ Rozhraní a dohled

Kdo může výstupy pochopit, zpochybnit nebo přepsat?

○ Nasazení

Jaká jsou zavedena bezpečnostní opatření, monitorování a postupy pro řešení incidentů?

Soulad s evropskými kompetenčními rámci

Evropské rámce propojují etické zásady s praktickými inženýrskými kompetencemi.

Pokyny

The Etické pokyny EU definují požadavky na důvěryhodnou umělou inteligenci. DigComp 2.2 a DigCompEdu podporují kompetence žáků a pedagogů v oblasti odpovědného přístupu.

Uplatnění

Studenti zkoumají dohled, spravedlnost, ochranu soukromí, odpovědnost, vysvětlitelnost a bezpečnost v rámci inženýrského systému.

Důkazy

Analýza případů, ochranná opatření a záznam o rozhodování ukazují, co mohou studenti analyzovat, testovat a odůvodňovat.

01

ETIKA A INŽENÝRSTVÍ

Etika v inženýrské praxi

Etika je systematické zkoumání toho, co je správné, spravedlivé a zodpovědné.

V inženýrství řídí etika odborné rozhodování v případech, kdy technická rozhodnutí ovlivňují lidi, práva, bezpečnost, přístup nebo životní prostředí.

Etický inženýr se ptá:

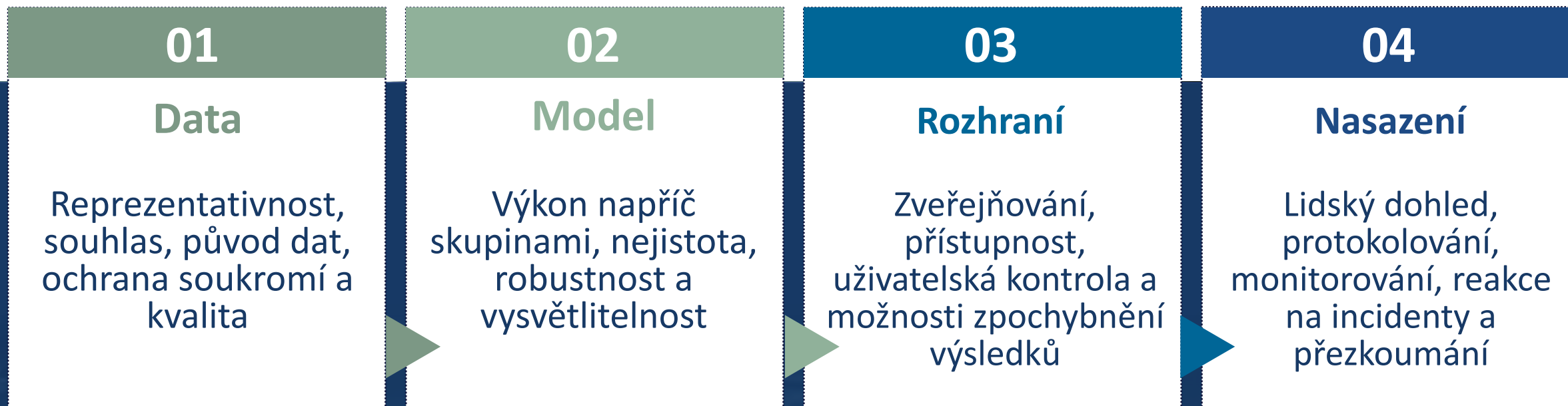
- Komu to může prospět nebo ublížit?
- Které hodnoty a práva jsou dotčeny?
- Jaké důkazy podporují tuto volbu návrhu?
- Kdo nese odpovědnost v případě selhání systému?

Požadavky stanovují profesní kodexy, institucionální zásady a právní předpisy. Etické uvažování podporuje zodpovědné rozhodování v situacích, kdy je na výběr několik možností.

[Zdroje] Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení (vývoj OER), str. 61–70.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

Etika se stává inženýrským požadavkem

Etiku lze převést do konkrétních otázek týkajících se návrhu a kritérií přijatelnosti:



Úkolem inženýrů je zajistit funkčnost hodnot prostřednictvím specifikací, testů, dokumentace a správy.

Etické uvažování využívá několik úhlů pohledu

Tyto tři úhly pohledu odhalují různé aspekty téhož rozhodnutí.

○ Důsledky

Zvažte pravděpodobné přínosy a škody, kdo je zažívá, jejich pravděpodobnost a zda je lze zvrátit. Návrh s dobrým průměrným výkonem může přesto způsobit vážnou škodu menší skupině.

○ Povinnosti a práva

Zeptejte se, které povinnosti platí i v případě, že se jiná možnost jeví jako efektivní. Lidská důstojnost, soukromí, nediskriminace, informovaná volba a bezpečnost mohou klást omezení na chování systému.

○ Profesionální charakter

Zeptejte se, co vyžaduje kompetentní, čestná a odpovědná inženýrská praxe. To zahrnuje přiznání nejistoty, hledání důkazů, zaznamenávání nesouhlasných názorů a eskalaci rizika, pokud tým nemá pravomoc jej vyřešit.

[Zdroje] - <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Zákony, normy, kodexy a etika

- Zákony stanovují vymahatelné povinnosti a zakázané praktiky.
- Technické normy definují dohodnuté metody, procesy a důkazy.
- Profesní kodexy stanovují povinnosti, jako je odborná způsobilost, poctivost, veřejná bezpečnost a odpovědnost. Organizační zásady rozdělují místní odpovědnosti a schvalovací postupy.
- Etické uvažování zkoumá možnosti, které v rámci těchto hranic zbývají. Ptá se, která varianta respektuje lidi, spravedlivě rozděljuje riziko a lze ji podložit důkazy.
- Dodržování předpisů ponechává mnoho možností volby, pokud jde o cíle, data, prahové hodnoty, rozhraní, dohled a monitorování. Inženýři zůstávají odpovědní za rozpoznávání a eskalaci etických rizik v rámci své role.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
Ethical Engineer schválil žádost KA220 HED, WP3 Výuka a učení, strany 61 až 70.

AI a etika (I)

Umělá inteligence může zakotvovat nebo zesilovat předsudky, znesnadňovat vyvozování odpovědnosti, ovlivňovat bezpečnost a práva a měnit rozložení moci mezi zúčastněnými stranami. Ačkoli umělá inteligence nabízí zřejmé výhody, přináší také etické výzvy, které nelze ignorovat.

Algoritmy mohou zakotvit nebo zesílit stávající předsudky – například systém pro nábor zaměstnanců může neúmyslně diskriminovat určité skupiny, pokud je trénován na zaujatých historických datech. Odpovědnost může být zastřena, když jsou rozhodnutí přijímána složitými modely, což vede k nejasnostem ohledně toho, kdo nese odpovědnost v případě, že dojde ke škodě.

V sázce je také bezpečnost a práva, protože chyby v autonomních vozidlech nebo lékařských systémech umělé inteligence mohou vést k fyzickým škodám. Současně mohou technologie sledování porušovat soukromí a lidskou důstojnost.

AI a etika (II)

Umělá inteligence navíc přerozděluje moc mezi zúčastněné strany: společnosti, inženýři, regulační orgány a veřejnost mohou mít protichůdné zájmy a ti, kdo mají kontrolu nad daty a algoritmy, mohou získat nepřiměřený vliv.

Doporučení UNESCO k etice umělé inteligence na tato rizika upozorňuje a zdůrazňuje, že lidská důstojnost, spravedlnost, transparentnost a odpovědnost musí zůstat ústředními prvky při vývoji umělé inteligence. Zákon EU o umělé inteligenci se k těmto obavám připojuje tím, že zakazuje manipulativní a vykořisťovatelské praktiky v oblasti umělé inteligence a ukládá přísná pravidla pro vysoce rizikové systémy s cílem chránit základní práva a společenské blaho.

Etika umělé intelligence vyžaduje systémový pohled

Systémy umělé intelligence kombinují data, modely, rozhraní, lidi, organizace a provozní prostředí. Etické dopady mohou vycházet z jakékoli části tohoto systému. Zaujatost se může vkrádat prostřednictvím štítků nebo cílů. Škoda může vzniknout v důsledku technicky přesného výstupu použitého v nesprávném kontextu. Jasné vysvětlení může být i tak neúčinné, pokud dané osobě chybí čas nebo pravomoc jednat.

Umělá intelligence také mění rozsah a moc. Jeden model může ovlivnit mnoho lidí, reprodukovat historické vzorce a utvářet přístup k práci, vzdělání, zdravotní péči nebo veřejným službám. Odpovědnost se proto vztahuje na návrh, zadávání zakázek, nasazení, monitorování i reakci na incidenty. Etickou jednotkou analýzy je celá služba a její důsledky pro dotčené osoby.

[Zdroje] <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Etické zásady a profesní inženýrské chování

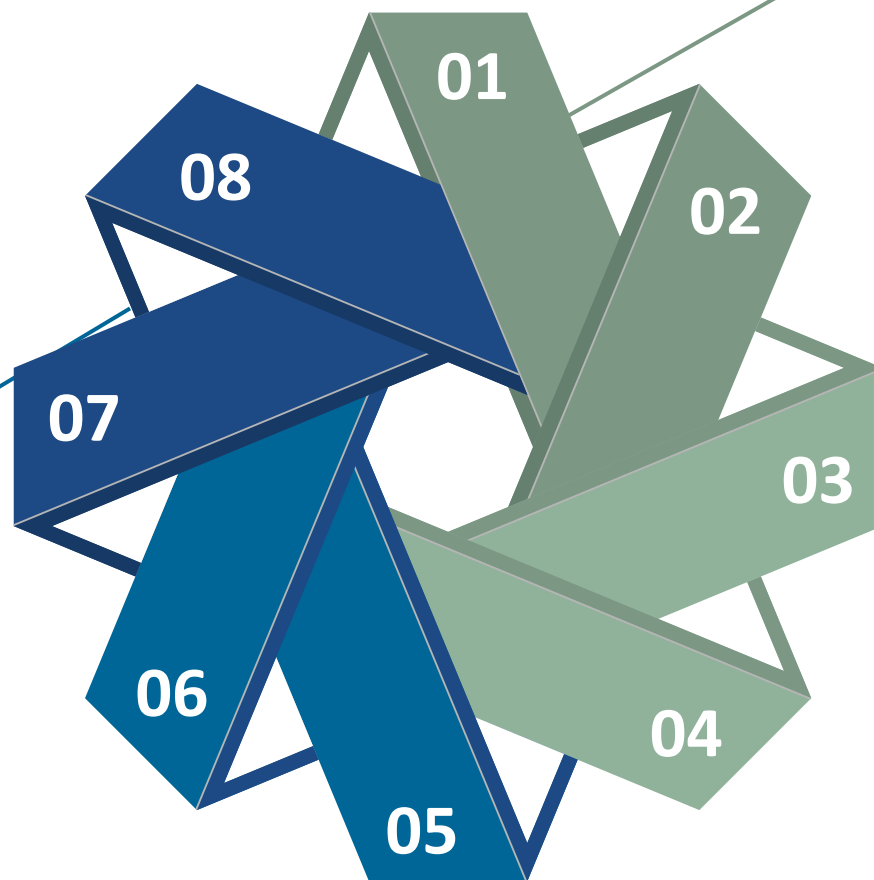
Profesionální chování inženýrů zahrnuje:

05. Úcta

06. Zákonnost

07. Odpovědnost

08. Odhodlání k dokonalosti



01. Upřímně
02. Integrita
03. Důvěryhodnost
04. Spravedlnost

[Zdroje] Aplikace „Ethical Engineer“ schválená v rámci projektu KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.

Lapiņa, I. (2019), jak je citováno v seznamu zdrojů. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Etické zásady a profesní inženýrské chování

Tyto hodnoty podporují etické rozhodování. Inženýrství v oblasti umělé inteligence také vyžaduje požadavky na úrovni systému.

Následující oddíly se zabývají:

- Autonomie
- Spravedlnost
- transparentnost a zásada non-maleficence vůči datům,
- model
- Rozhraní
- nasazení a
- dohled nad rozhodováním.

Čtyři zásady fungují společně

Spravedlnost se ptá, jak jsou rozděleny přínosy, chyby a zátěže a zda mají dotčené osoby k dispozici spravedlivý proces a nápravu.

Autonomie se ptá, kdo má informace, možnost volby, pravomoc a smysluplnou kontrolu. Transparentnost se ptá, které důkazy musí být k dispozici dotčeným osobám, provozovatelům, inženýrům a auditorům.

Zásada non-maleficence se ptá, které škody jsou předvídatelné a jak budou předcházeny, odhalovány a řešeny. Jediné rozhodnutí o návrhu může zahrnovat všechny zásady. Prahová hodnota pro lékařský poplach ovlivňuje, komu bude věnována pozornost, zda si lékaři zachovají kontrolu, jaká nejistota je viditelná a kterým pacientům může zpoždění nebo falešný poplach ublížit.

Etická analýza získává na síle, když jsou tyto principy používány jako propojené otázky, nikoli jako izolované pojmy.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Reflexe – diskuse – závěr (10–15 minut)

Otázka		<i>„Kde se již v každodenním životě nebo při studiu setkáváte s umělou inteligencí (nebo automatizací)? Jaké etické otázky by mohly vyvstat?“</i>
Přemýšlej	2 minuty	Studenti si zapíší 2–3 příklady/problémy.
V páru	5 minut	Porovnají si seznamy s partnerem.
Sdílení	3–8 minut	Na tabuli shromážděte 4–6 příkladů (například: rozpoznávání obličeje a souhlas, nástroje pro hodnocení a spravedlnost; systémy návrhů a transparentnost).

02

SPRAVEDLNOST



Spravedlnost v inženýrství umělé intelligence

Spravedlnost se týká spravedlivého a rovného zacházení, včetně toho, jak jsou rozdělovány výhody, rizika, chyby a příležitosti.

Pro inženýry, to vyžaduje důkazy:

- Jsou v datech a validační sadě zastoupeny relevantní skupiny?
- Jsou míry chybovosti a dopady posuzovány napříč dotčenými skupinami?
- Ovlivňují rizika přístupnosti a vyloučení návrh systému?
- Mohou lidé podat odvolání, požádat o přezkoumání člověkem nebo hledat nápravu?

Spravedlnost závisí na kontextu. Týmy by měly zdokumentovat zvolené cíle v oblasti spravedlnosti, kompromisy a dotčené zainteresované strany.

[Zdroje] - <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

Spravedlnost má čtyři inženýrské dimenze

1. **Distributivní spravedlnost se týká** to, kdo získává výhody, chyby, náklady, zpoždění a příležitosti. Inženýři potřebují měřítka výsledků pro příslušné skupiny a provozní podmínky.
2. **Procesní spravedlnost se týká** způsobu, jakým se přijímají rozhodnutí. Lidé potřebují konzistentní pravidla, smysluplné lidské posouzení a možnost napadnout výsledek.
3. **Uznání se týká toho**, zda návrh zohledňuje potřeby a zkušenosti lidí, kteří mohou být přehlíženi, včetně osob se zdravotním postižením a menšinových skupin.
4. **Nápravná spravedlnost se týká nápravy.** Služba potřebuje proces, který umožní opravit chybu, obnovit přístup, vyšetřit příčiny a zabránit opakování.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://unesdoc.unesco.org/ark:/48223/pf0000381137>

Skutečný případ: Chyby v rozpoznávání obličeje

Oficiální důkazy:
Studie NIST o
rozpoznávání
obličeje

NIST vyhodnotil téměř 200 algoritmů rozpoznávání obličeje s využitím více než 18 milionů snímků přes 8 milionů lidí. U většiny testovaných algoritmů se objevily demografické rozdíly. Vzor chyb závisel na algoritmu, aplikaci a obrazových datech.

Technický problém se týká jak kontextu použití, tak výkonu modelu. Falešná shoda může vystavit osobu vyšetřování. Neúspěšné ověření může odepřít přístup ke službě. Inženýři musí identifikovat příslušnou chybu, otestovat dotčené skupiny a realistické podmínky snímků, zdokumentovat nejistotu a zajistit lidskou kontrolu a nápravu.

Diskuse:

Která chyba způsobuje největší škodu v oblasti přístupu na akademickou půdu, policejní činnosti a ověřování pomocí telefonu?

[Zdroje] <https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software>

Aplikovaný úkol: Spravedlnost při přidělování infrastruktury

Regionální úřad využívá model umělé inteligence k určení priorit při inspekcích mostů. Městské silniční mosty disponují bohatými údaji ze senzorů a o údržbě. U venkovských mostů, přechodů pro chodce a tras využívaných osobami se zdravotním postižením jsou záznamy sporadické. Model směřuje většinu inspekčních kapacit na místa s nejkomplexnějšími údaji a nejvyšším provozem.

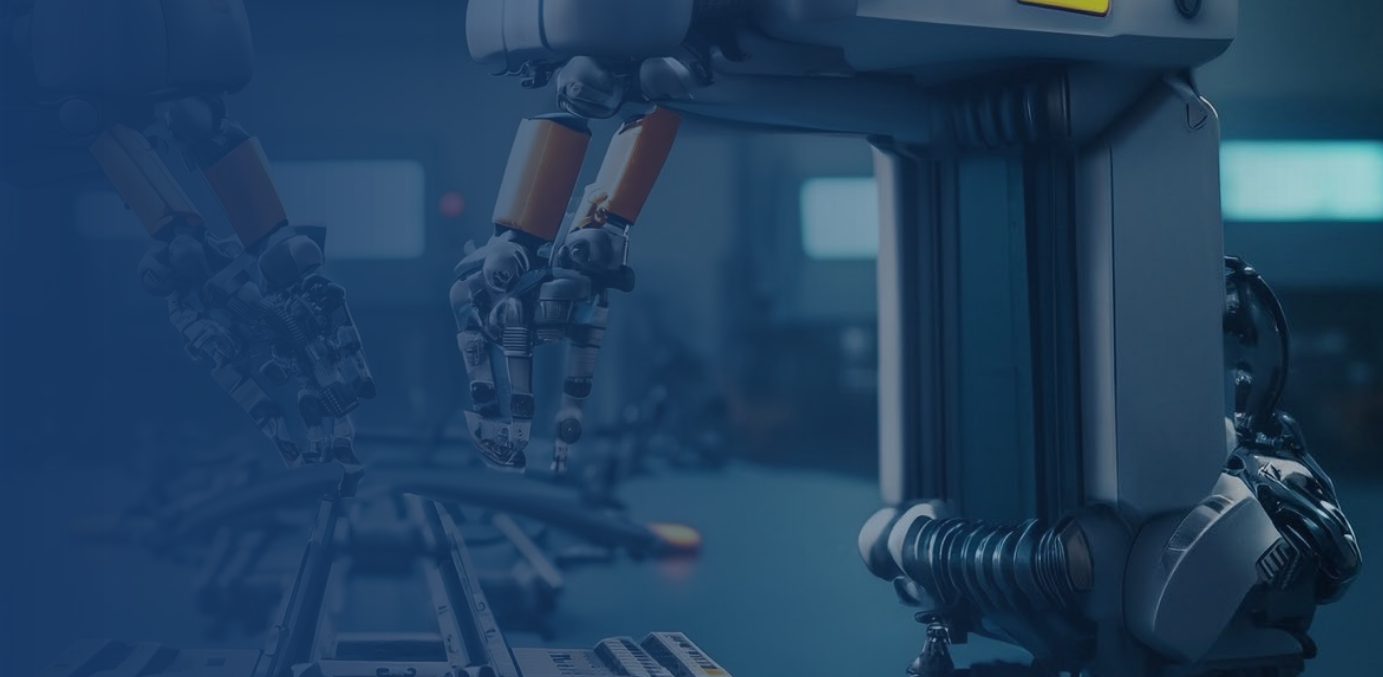
Spravedlnost se ptá, jak jsou rozděleny bezpečnostní pozornost, rizika a přístup, a zda mají opomíjené komunity spravedlivou cestu k přezkumu.

Rozhodněte se: POKRAČOVAT, UPRAVIT nebo ZASTAVIT. Uveďte dotčené skupiny, chybějící důkazy, pravidlo spravedlivého přidělování, cestu k ručnímu nominování, kritéria přijatelnosti, odpovědnou osobu a nápravu.

Kontrola kvality: Kdo získá přednostní ochranu, kdo nese zbytkové riziko a jak lze opomenutí napravit?

03

AUTONOMIE



Lidská autonomie a inženýrská odpovědnost

Lidská autonomie je schopnost činit informovaná rozhodnutí a jednat bez nepřiměřeného nátlaku. V systémech umělé inteligence závisí na smysluplných informacích, případně na souhlasu, skutečné kontrole ze strany uživatele a možnostech zpochybnit nebo přepsat výsledky.

Inženýři ovlivňují autonomii prostřednictvím návrhu rozhraní, výchozích nastavení, úrovně automatizace, mechanismů přepsání a přidělování rozhodovací pravomoci.

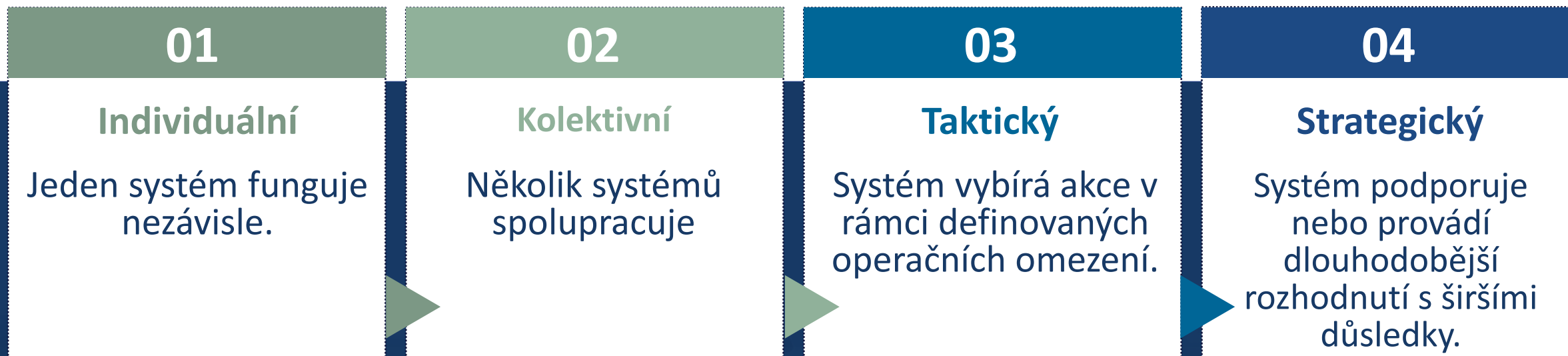
Klíčový rozdíl:

Autonomie systému popisuje, co dokáže stroj vykonat samostatně. Lidská autonomie se týká svobodného jednání a práv osob, na které má systém vliv.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

Úrovně autonomie systému

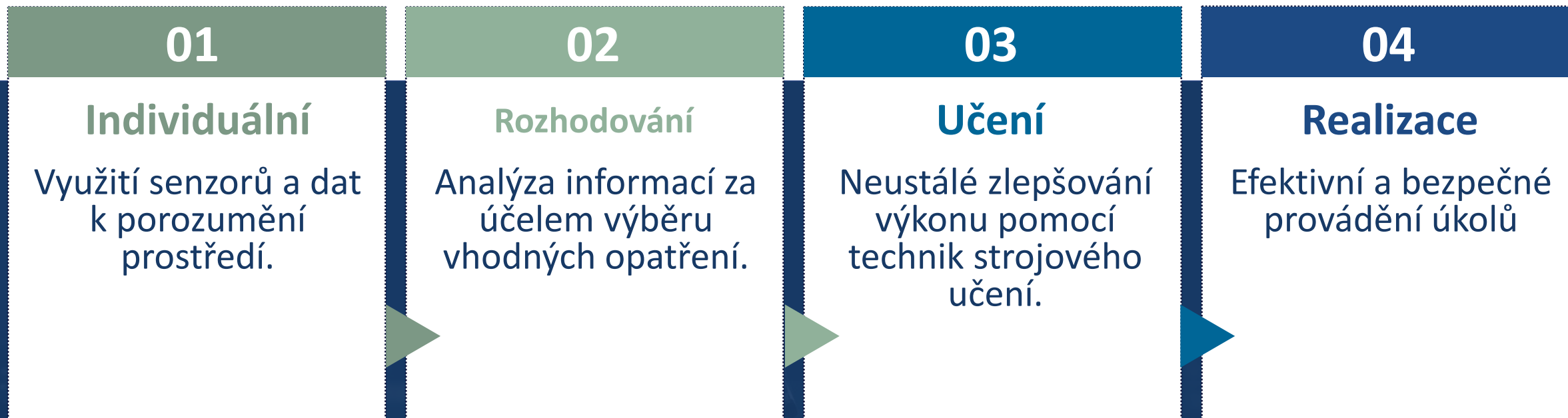
Inženýři mohou autonomii popsat na různých úrovních:



Pro každou úroveň specifikujte provozní hranice, lidský dohled, bezpečnostní chování, odpovědnost a důkazy požadované před nasazením. Každá úroveň má odlišný profil rizika a požadavky na bezpečnostní opatření.

Autonomní AI (II)

Mezi klíčové vlastnosti autonomní AI patří:



Autonomní AI mění tvář různých odvětví, jako je doprava, výroba, zdravotnictví a logistika, a nabízí vyšší efektivitu, bezpečnost a nové možnosti. Vyvolává však také etické, bezpečnostní a regulační otázky, zejména pokud jde o odpovědnost za rozhodnutí a dopad na zaměstnanost.

Autonomie mění rozhodovací pravomoc

Pro každou automatizovanou funkci by inženýři měli specifikovat povolené akce, provozní hranice, roli člověka, informace potřebné pro zásah, dobu odezvy, záložní řešení a odpovědnou osobu. Osoba uvedená v pracovním postupu má smysluplnou kontrolu pouze tehdy, pokud dokáže rozpoznat problém, porozumět stavu systému a jednat dříve, než dojde ke škodě.

Autonomie systému popisuje funkce prováděné bez okamžitého lidského zásahu. Lidská autonomie se týká schopnosti lidí rozumět, volit, odmítat, zasahovat a zpochybňovat rozhodnutí. Etická otázka se proto týká pravomoci: kdo smí rozhodovat, kdo smí jednat a kdo může proces zastavit.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Skutečný případ: Dohled nad automatizovaným řízením

Při nehodě v Tempe v roce 2018 srazilo testovací vozidlo společnosti Uber chodce, který na následky srážky zemřel. NTSB zjistila nedostatečné posouzení bezpečnostních rizik, neúčinný dohled a slabou bezpečnostní kulturu. Varování před kolizí a automatické nouzové brzdění byly deaktivovány. Oficiální vyšetřování: Zpráva přijatá NTSB.

Lidský dohled je záměrnou bezpečnostní funkcí. Vyžaduje jasné informace o stavu systému, realistické požadavky na bdělost, dostatečnou dobu na reakci, prověřený zásah, bezpečný záložní systém a organizační monitorování.

Diskuse:

Která rozhodnutí týkající se systému, rozhraní a řízení ovlivnila výsledek?

Aplikovaný úkol: Autonomie v řízení procesů

Řídicí jednotka s umělou inteligencí reguluje ventily chladicí vody v elektrolyzáru vodíku. Polohu ventilu může měnit každou sekundu. Při poruše snímače může během osmi sekund vzniknout nebezpečný tlak.

Operátor potřebuje asi deset sekund na vyhodnocení alarmu a dalších pět sekund na provedení opatření. Dodavatel popisuje systém jako pod lidským dohledem, protože je k dispozici tlačítko nouzového zastavení.

Rozhodněte se: POKRAČOVAT, PŘEPRACOVAT nebo ZASTAVIT. Specifikujte povolené akce regulátoru, provozní hranice, pravomoc člověka, informace o alarmu, dobu odezvy, automatický bezpečný stav, nezávislou ochrannou vrstvu, ověřovací test, vlastníka důkazů a spouštěč přezkumu.

Kontrola kvality: Je lidské řízení schopno včas reagovat za realistických podmínek poruchy a pracovní zátěže?

04

TRANSPARENTNOST

Transparentnost jako vlastnost technického návrhu

Transparentnost umožňuje dotčeným osobám, provozovatelům, auditorům a regulačním orgánům pochopit, k čemu systém umělé inteligence slouží, jak se používá a kde leží jeho meze

Inženýři zajišťují transparentnost prostřednictvím dokumentace, sledovatelnosti a návrhu uživatelského rozhraní.

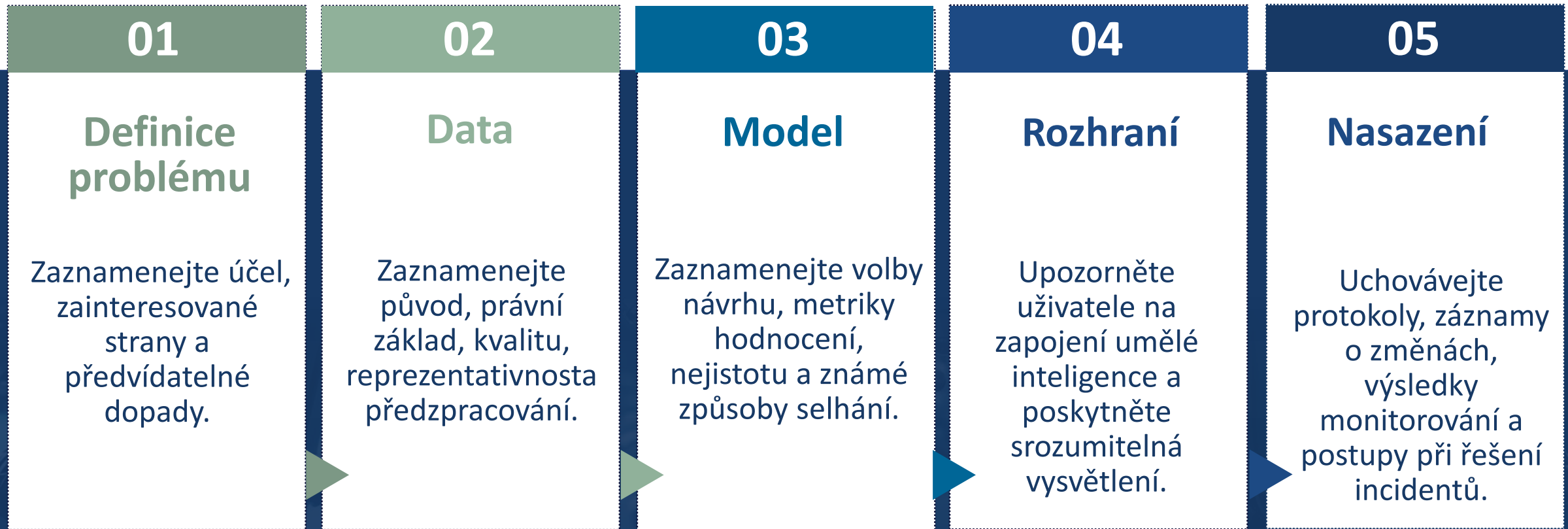
Mezi relevantní informace mohou patřit:

- Zamýšlené a zakázané použití
- Zdroje dat, kvalita a známé mezery
- Omezení modelu, nejistota a výkonnost napříč skupinami
- Lidský dohled a postupy eskalace
- Monitorování, incidenty a změny systému

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Transparentnost v průběhu celého životního cyklu AI

Transparentnost by měla být přizpůsobena cílovému publiku a zároveň chránit soukromí, bezpečnost a oprávněná práva duševního vlastnictví.



Technické kompromisy v oblasti transparentnosti

Transparentnost podporuje důvěru, odpovědnost, testování a informované používání. Může však představovat riziko, pokud jsou citlivé osobní údaje, bezpečnostní detaily nebo chráněné duševní vlastnictví zveřejňovány bez náležitých bezpečnostních opatření.

Cílem inženýrské práce je

účelná transparentnost:

- Zveřejňujte informace, které jednotlivé zainteresované strany potřebují k bezpečnému jednání a k zpochybňování rozhodnutí
- Chraňte informace, které by mohly ohrozit soukromí nebo bezpečnost
- Zaznamenejte, proč jsou informace zadržovány a kdo k nim má přístup pro účely auditu
- Ověřte, zda jsou vysvětlení přesná, použitelná a přístupná

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Klíčové složky transparentnosti AI (I)

○ Vysvětlitelnost

Schopnost vysvětlit, jak AI dospěla k určitému rozhodnutí nebo výsledku (např. proč byla zamítnuta žádost o půjčku).

○ Transparentnost dat

Jasná dokumentace toho, jaká data byla použita k trénování modelu, včetně zdrojů, kvality a potenciálních zkreslení

○ Transparentnost algoritmů

Přístup k použitým algoritmům (např. modely s otevřeným zdrojovým kódem, dokumentace rozhodovací logiky).

Transparentnost jetechnickým podkladem pro jednání

Transparentnost podporuje bezpečné jednání, kritické posuzování a odpovědnost. Její účel se mění v závislosti na cílovém publiku. Provozovatelé potřebují znát stav systému, míru nejistoty a požadovanou reakci. Dotčené osoby potřebují vědět, že byla použita umělá inteligence, jak je výsledek ovlivňuje a jak mohou požádat o přezkoumání. Inženýři a auditoři potřebují sledovatelnost napříč daty, testy, verzemi, rozhodnutími a incidenty.

Technické artefakty to umožňují v praxi:

- Záznamy o systému a modelu definují zamýšlené použití, omezení a režimy selhání.
- Důkazy z testování pokrývají relevantní podmínky a dotčené skupiny.
- Zprávy v rozhraní informují o nejistotě, opatřeních a eskalaci.
- Protokoly, záznamy o změnách a hlášení incidentů podporují rekonstrukci a přezkoumání.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Transparentnost závisí na cílové skupině a účelu

Užitečná transparentnost umožňuje konkrétní osobě porozumět, jednat, zpochybňovat nebo provádět audit. Zveřejnění technických detailů bez určení cílové skupiny může vytvářet zdání otevřenosti, přičemž rozhodnutí zůstává bez odpovědnosti. Informace by měly být přesné, aktuální, přístupné a propojené s konkrétním postupem pro jednání.

- Dotčená osoba potřebuje vědět, že je do procesu zapojena umělá inteligence, jaké rozhodnutí podporuje, jaké jsou relevantní důvody, jaká jsou důležitá omezení a jak dosáhnout lidského přezkumu.
- Provozovatel potřebuje znát stav systému, míru spolehlivosti, požadované akce, provozní limity a pokyny pro eskalaci.
- Inženýři a auditoři potřebují informace o původu dat, verzích, testech, protokolech, incidentech, změnách a schváleních.

Skutečný případ z oblasti inženýrství:

Transparentnost modelu Boeing 737 MAX

Oficiální vyšetřování:
Materiály ke zprávě o
letadle Boeing 737
MAX.

[Zdroje] <https://democrats-transportation.house.gov/committee-activity/boeing-737-max-investigation>

Vyšetřování Výboru Sněmovny reprezentantů Spojených států týkající se letounu Boeing 737 MAX se zabývalo návrhem, vývojem a certifikací letounu. Jeho zpráva identifikovala opakované poruchy týkající se systému pro vylepšení manévrovacích charakteristik (Maneuvering Characteristics Augmentation System, známého jako MCAS) a širší slabiny v certifikaci a dohledu.

Otázka transparentnosti se týká informací potřebných pro bezpečné jednání a nezávislé posouzení. Piloti, letecké společnosti a regulační orgány potřebovaly přesné důkazy o pravomoci systému, závislosti na senzorech, chování při selhání, indikacích v kokpitu, postupech a potřebách školení. Neúplné nebo roztříštěné informace oslabily jejich schopnost porozumět konstrukci a zpochybnit ji.

Diskuse:

Které technické záznamy, důkazy o rizicích a rozhodnutí o změnách měly být dostupné, komu a v jaké fázi schvalovacího procesu?

Praktický úkol:

Transparentní zadávání veřejných zakázek v železničním inženýrství

Provozovatel železnice plánuje pořídit systém umělé inteligence, který identifikuje závady na kolejích na základě snímků z inspekci. Dodavatel uvádí celkovou přesnost 97 %, ale nezveřejňuje výkonnost podle typu koleje, počasí, stavu kamery ani závažnosti závady. Původ trénovacích dat a známé režimy selhání nejsou k dispozici. Smlouva povoluje aktualizace modelu bez předchozího upozornění. Odpovědnost za přehlédnuté závady zůstává na údržbářích.

Napište

Uvedte důkazy a smluvní požadavky potřebné před zadáním zakázky. Zahrňte zamýšlené použití, provozní limity, původ datových sad, výsledky na úrovni vad, nejistotu, záznamy o verzích a změnách, přístup k protokolům, spolupráci při incidentech, nezávislé testování a práva na odstoupení.

Rozhodnutí

Uvedte jednu podmínku, která vyžaduje REVIZE, a jednu podmínku, která vyžaduje ZASTAVENÍ. Vysvětlete, jak může každá cílová skupina na základě poskytnutých informací jednat

05

NON- MALEFICENCE

NON-MALEFICENCE

Neubližovat, předvídat a zmírňovat rizika. Princip neškodnosti vyžaduje, aby inženýři zabránili vytváření systémů, které způsobují fyzickou, psychickou nebo společenskou újmu.

Vyžaduje také předvídat potenciální nebezpečí a zavést ochranná opatření k zmírnění nezamýšlených důsledků.

Újma přesahuje rámec fyzického zranění

- Fyzická újma zahrnuje zranění a nebezpečný provoz. Psychická újma zahrnuje úzkost, manipulaci a ztrátu důvěry. Ekonomická újma zahrnuje ušlý zisk, vyšší náklady a vyloučení ze služeb. Újma související s právy zahrnuje narušení soukromí, diskriminaci a odepření spravedlivého procesu. Sociální a environmentální újma zahrnuje sledování, nerovnoměrné rozložení moci, využívání zdrojů a dopady na komunity.
- Zásada neškodnosti vyžaduje, aby inženýři před výběrem kontrolních opatření identifikovali ohrožené osoby a pravděpodobné škody. Je třeba společně zohlednit závažnost, pravděpodobnost, míru ohrožení a zvratnost. Vzácná závažná škoda, častá nízká zátěž a chyba, kterou nelze napravit, vyžadují každá odlišný přístup.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://unesdoc.unesco.org/ark:/48223/pf0000381137>

Bezpečnost je povinností po celou dobu životního cyklu

- Před uvedením do provozu definujte zamýšlené a zakázané použití, identifikujte nebezpečí a předvídatelné zneužití, otestujte okrajové případy a zhoršené podmínky a specifikujte měřitelná kritéria přijatelnosti. Během provozu sledujte výkon, skupinové výsledky, přepsání nastavení, stížnosti, incidenty a změny kontextu. Po incidentu chraňte postižené osoby, zajistěte důkazy, určete příčiny, opravte kontrolní opatření a ověřte změnu.
- Kontrolní opatření by měla pokud možno řešit zdroj poškození. Týmy mohou změnit cíl, odstranit nebezpečnou funkci, omezit provozní doménu, přidat technickou detekci a záložní řešení, posílit lidskou kontrolu nebo zastavit nasazení. Bezpečnostní analýza zůstává platná pouze dokud zůstávají platné její předpoklady, verze systému, kontext a důkazy.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Skutečný inženýrský případ: Bezpečnost procesů v rafinérii BP v Texas City

Dne 23. března 2005 si výbuchy v rafinérii BP Texas City vyžádaly 15 životů a 180 zraněných. Americká Komise pro chemickou bezpečnost (United States Chemical Safety Board) uvedla, že destilační věž byla zaplavena, došlo v ní k přetlaku a z atmosférického odvětrávacího komína unikly uhlovodíky. Její vyšetřování odhalilo organizační a bezpečnostní nedostatky v celé společnosti.

Zásada non-maleficence vyžaduje, aby inženýři předcházeli předvídatelným škodám v rámci celého systému. K riziku přispívá konstrukce zařízení, výstražné systémy, provozní postupy, personální obsazení, řízení změn, uspořádání areálu, klíčové ukazatele bezpečnosti i dohled ze strany vedení. Statistiky osobní bezpečnosti nemohou nahradit důkazy o katastrofických procesních rizicích.

Diskuse:

Které preventivní vrstvy selhaly, jaké důkazy měly vyvolat reakci a kdo měl pravomoc zastavit restart?

Aplikovaný úkol: Omezený pilotní projekt v oblasti procesní bezpečnosti s vyžitím AI

Chemický závod navrhuje poradenský nástroj založený na umělé inteligenci, který předpovídá kolísání kompresoru a doporučuje změny v nastavení obtokových ventilů. Validace se vztahuje na provoz v ustáleném stavu. Neexistují žádné důkazy týkající se spouštění, vypínání, driftu senzorů nebo ztráty komunikace. Doporučení se aplikuje automaticky, pokud jej operátor nezruší do šesti sekund. Dodavatel navrhuje tříměsíční pilotní provoz v reálných podmínkách

Rozhodněte: SPUSTIT, UPRAVIT nebo ZASTAVIT. Definujte hranice pilotního provozu, vyloučené provozní režimy, nezávislé ochranné vrstvy, dobu reakce obsluhy, testy před uvedením do provozu, monitorovací signály, postup při incidentu, prahové hodnoty pro odstavení, vlastníka důkazů a datum přezkumu.

Vypracujte stručnou bezpečnostní analýzu: nebezpečí, příčinná souvislost, kontrola, důkazy o ověření, zbytkové riziko a odpovědné schválení. Uveďte, které chybějící důkazy brání provozu v reálném prostředí.

[Zdroje] <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
<https://www.csb.gov/bp-america-texas-city-refinery-explosion/>

06

SKUPINOVÁ AKTIVITA

Případ A: Stanovení priorit inspekcí pomocí BridgeSense

BridgeSense řadí mosty podle priority inspekce na základě snímků, vibračních senzorů, intenzity provozu a historie údržby. Starší venkovské mosty mají méně senzorů a neúplné záznamy. Inspekci podstoupí pouze 20 lokalit s nejvyšším hodnocením. Nízké pořadí může odrážet spíše chybějící data než bezpečnostní riziko. Zpoždění by mohlo způsobit škody, uzavření silnice nebo ztrátu nezbytného přístupu.

Úkol:

Rozhodněte: SPUSTIT, UPRAVIT nebo ZASTAVIT. Uveďte, jak by mělo hodnocení ovlivnit odborný úsudek a alokaci zdrojů.

Dostupné důkazy:

Celková přesnost a ověření v městském prostředí.

Chybějící důkazy:

Výsledky podle typu a stavu mostu, míry nejistoty, dopadu na komunitu, záznamů o přepsání a revizí

Případ B: Provoz turbín WindControl

WindControl upravuje sklon lopatek a načasování údržby. Byl vycvičen na novějších turbínách za mírného počasí. V lokalitě se však nacházejí i starší turbíny, dochází k námraze, silným poryvům větru a přerušované komunikaci. Rozhraní nezobrazuje míru spolehlivosti ani podpůrná data.

Falešný poplach může způsobit odstávku. Přehlédnutá porucha může poškodit zařízení nebo ohrozit techniky a okolí.

Úkol:

Schválit omezený pilotní projekt, požadovat přepracování nebo zastavit. Definovat provozní limity, lidskou autoritu a podmínky pro automatické zásahy.

Dostupné důkazy:

Zkoušky v extrémních povětrnostních podmínkách, výsledky starších turbín, doba odezvy, chování v případě selhání, řízení změn a monitorování téměř nehod.

Analýza inženýrského případu a výstup týmu (30 minut)

Vyberte si jeden případ.

Ve skupinách po čtyřech nebo pěti si rozdělte role: systémy, bezpečnost, důkazy, zainteresované strany a zapisovatel.

1. Definujte zamýšlené použití, hranice systému, podmínky a dotčené osoby.
2. Pojmenujte inženýrské rozhodnutí, které vede k dilematu.
3. Zjistěte, jak data, model, rozhraní, automatizace a monitorování ovlivňují výsledky.
4. Aplikujte všechny čtyři zásady a identifikujte případné konflikty.
5. Navrhněte technická a řídicí bezpečnostní opatření včetně testů a odpovědných osob.
6. Zaznamenejte rozhodnutí „POKRAČOVAT“, „REVIDOVAT“ nebo „ZASTAVIT“ a důvod pro přezkum.

Výstup: jeden záznam o rozhodnutí a dvouminutové odůvodnění založené na důkazech.

[Zdroje] Společnost Ethical Engineer schválila žádost KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Aplikujte čtyři zásady na jedno inženýrské rozhodnutí

Používejte po celou dobu stejné kritické rozhodnutí, aby každý princip zůstal propojen s inženýrským postupem.

01

Spravedlnost

Kdo nese následky v podobě bezpečnosti, dostupnosti, zpoždění, nákladů, chyb nebo rizik?

Je proces spravedlivý a umožňuje nápravu?

02

Autonomie

Kdo má informace a pravomoc?
Může daná osoba včas zasáhnout?

03

Transparentnost

Co musí každý z adresátů vědět, aby mohl jednat, zpochybňovat nebo vyšetřovat?

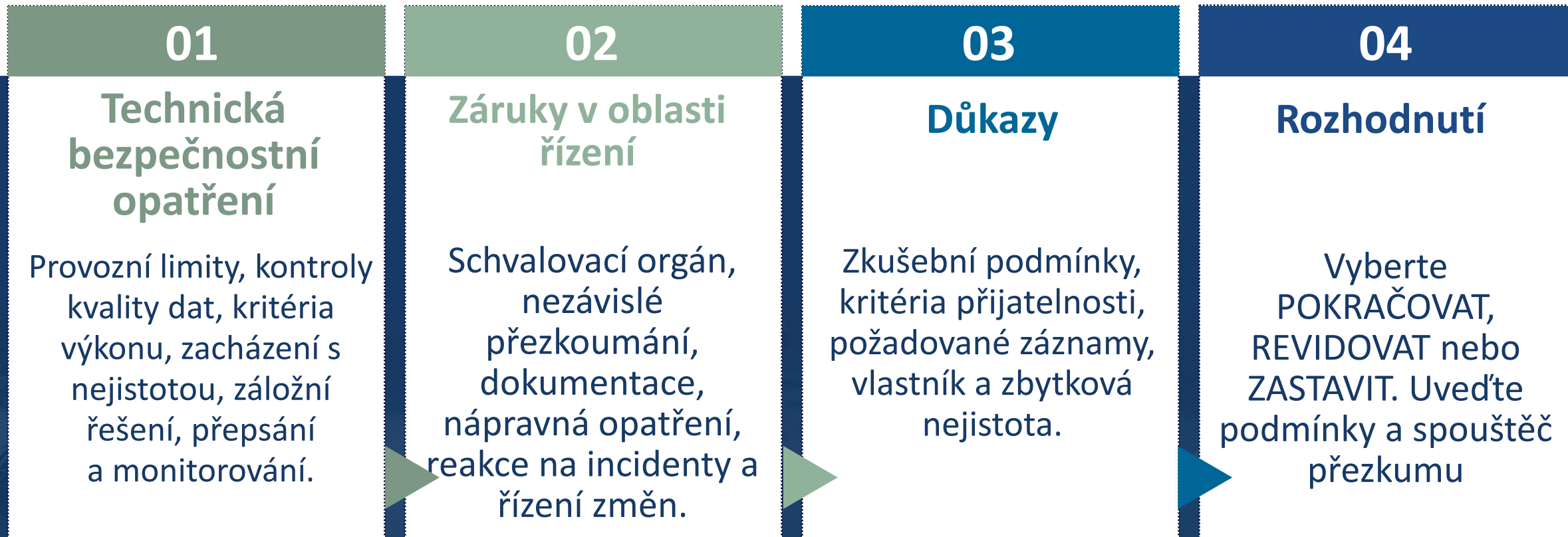
04

Non-maleficence

Jaké škody lze předvídat?
Které preventivní, detekční, zmírňující a nápravné kontroly jsou nezbytné?

Návrh bezpečnostních opatření, důkazy a dohled

Převést etickou analýzu do podoby technického návrhu.



Pracovní list: Záznam o rozhodnutí v oblasti technické etiky

Účel: Integrujte tyto čtyři principy do jednoho hodnotitelného inženýrského rozhodnutí. Vyplňte záznam pro BridgeSense nebo WindControl a poté připravte minutové odůvodnění.

Kvalitní záznam propojuje zásady s rozhodnutími, zárukami, důkazy, odpovědným vlastníkem a spouštěčem přezkumu.

Systém, zamýšlené použití a dotčené osoby:

Kritické inženýrské rozhodnutí:

Spravedlnost, autonomie, transparentnost a non-maleficence:

Technická bezpečnostní opatření a test přijatelnosti

Záruka správy a odpovědný vlastník

[Zdroje] Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení (vývoj OER), str. 61–70.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Pokračování na dalším snímku

Chybějící důkazy, nejistota a zbytkové riziko:

Rozhodnutí: POKRAČOVAT REVIDOVAT ZASTAVIT

Podmínky, monitorovací opatření a spouštěč přezkumu:

Odlišný názor nebo alternativní pohled:

Odůvodnění: Proč důkazy podporují toto rozhodnutí?

.....

.....

[Zdroje] Ethical Engineer schválil žádost KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Scénář pro hraní rolí

(25 minut)

Scénář:

Univerzita zvažuje nasazení autonomních hlídkových robotů pro noční monitorování. Kamery a detekce osob slibují rychlejší reakci na incidenty. Studenti vyjadřují obavy ohledně ochrany soukromí, diskriminace a falešných poplachů.

Role, jedna pro každou skupinu:

- **Vedoucí inženýr:** bezpečnost, výkon a provozní limity
- **Datový vědec:** datové sady, přesnost a zkreslení
- **Zástupce studentů:** práva, inkluze a osobní údaje
- **Správce zařízení:** provoz a odpovědnost
- **Etický nebo právní poradce:** posouzení rizik a dokumentace

[Zdroje] Schválená žádost Ethical Engineer v rámci projektu KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Individuální reflexe a aplikace

(15 minut)

Reflexe završuje skupinovou aktivitu a přenáší uvažování do jiného inženýrského kontextu.

Napište jeden nebo dva odstavce:

- Které dva principy nejvíce ovlivnily vaše rozhodnutí a proč?
- Který kompromis nebo nejistotu bylo nejtěžší vyřešit?
- Jaká technická změna je nutná před nasazením?
- Jaké důkazy by vás přiměly rozhodnutí revidovat?
- Jak by se uvažování lišilo v jiné disciplíně?

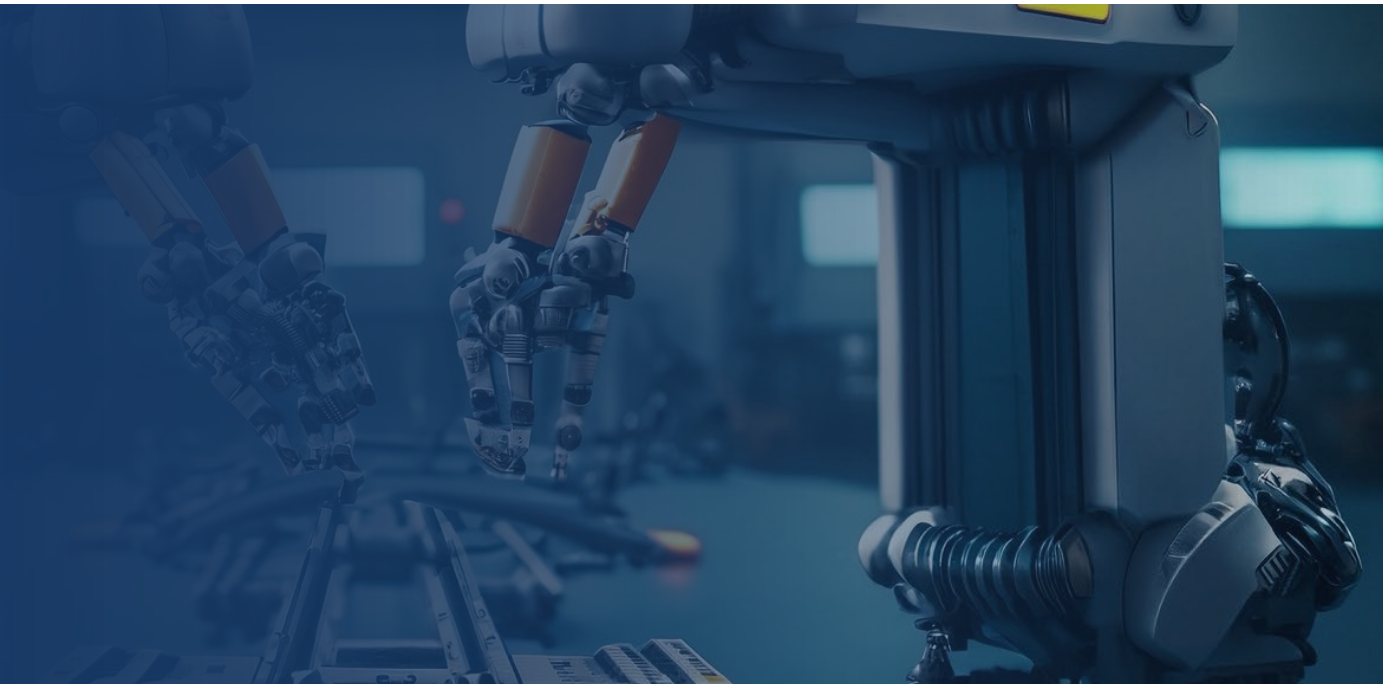
Důkazy:

Stručná argumentace propojující principy, inženýrské rozhodnutí, bezpečnostní opatření a důkazy. Přijatelné jsou písemné, zvukové nebo nahrané odpovědi.

[Zdroje] Ethical Engineer schválil žádost KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.

07

KVÍZ



Kvíz

1. Definujte *transparentnost* v oblasti umělé inteligence a uveďte jeden technický postup, který ji podporuje.
2. Který princip má za cíl **předcházet škodám** a uveďte jednu metodu snižování rizik v projektech umělé inteligence.
3. Uveďte **dva způsoby** , jakými mohou zaujatá trénovací data ovlivnit výsledky modelu.
4. V nástroji pro nábor zaměstnanců podporovaném umělou inteligencí identifikujte **jeden rozhodovací bod** , kde mohou inženýři snížit nespravedlnost.
5. Robot nesprávně klasifikuje pomůcky pro mobilitu jako překážky, čímž blokuje přístup. Který princip je tím nejvíce porušen?
 - a. *Autonomie*
 - b. *Spravedlnost*
 - c. *Transparentnost*
 - d. *Non-maleficence*

Odpovědi na kvíz a zdůvodnění výuky

1. Transparentnost znamená relevantní srozumitelnost, sledovatelnost a komunikaci. Mezi podpůrné postupy patří dokumentace datových sad, záznamy o modelech, vysvětlení, protokoly a záznamy o změnách.
2. Zásada non-maleficence. Vhodnou metodou může být analýza rizik, reprezentativní testování, provozní omezení, bezpečné záložní řešení, monitorování nebo reakce na incidenty.
3. Zkreslená data mohou vést k nerovnoměrným chybovostem, špatným výsledkům u nedostatečně zastoupených skupin, škodlivým náhradním ukazatelům a posílení historických vzorců.
4. Mezi platné rozhodovací body patří výběr dat, štítky, vlastnosti, cíl, prahová hodnota, rozhraní, lidská kontrola a monitorování.
5. B, Spravedlnost. Blokování přístupu rozděluje zátěž nespravedlivě. Může se uplatnit i zásada neškodnosti, protože chyba způsobuje škodu. Ocenit odůvodněnou odpověď.

Kontrola výstupů učení

Před dokončením by studenti měli být schopni:

- Uvedte technické rozhodnutí, které v daném případě způsobilo nebo snížilo újmu.
- K vysvětlení dané problematiky použijte alespoň dva etické principy.
- Doporučte jedno technické a jedno řídicí bezpečnostní opatření.
- Uvedte, jaké důkazy, dokumentace a monitorování by byly zapotřebí.
- Vysvětlete, jak mohou dotčené osoby výsledek pochopit, zpochybnit nebo napadnout.

Důkazy o osvojení znalostí:

Skupinová analýza, rozhodnutí v rámci hraní rolí, reflexivní odpověď a kvíz.

[Zdroje] Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení (vývoj OER), str. 61–70.

Základní odkazy a další literatura

- Skupina odborníků na vysoké úrovni Evropské komise pro umělou inteligenci. Etické pokyny pro důvěryhodnou umělou inteligenci.
- Vuorikari, Kluzer a Punie. DigComp 2.2.
- **UNESCO. Doporučení týkající se etiky umělé inteligence.**
- Evropská unie. Zákon o umělé inteligenci.
- **NIST. Test dodavatelů systémů rozpoznávání obličeje: demografické dopady.**
- NTSB. Vyšetřování nehody autonomního vozidla v Tempe.
- Výbor Sněmovny reprezentantů Spojených států. Vyšetřování letadla Boeing 737 MAX.
- Americká rada pro chemickou bezpečnost. Vyšetřování rafinérie BP v Texas City.
- Ethical Engineer schválil žádost KA220-HED, WP3 Výuka a učení, str. 61–70.

[Zdroje]

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

<https://unesdoc.unesco.org/ark:/48223/pf0000381137>

<https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition>

<https://democrats-transportation.house.gov/committee-activity/boeing-737-max-investigation>

Schválená žádost projektu Ethical Engineer KA220-HED, WP3 Výuka a učení (vývoj OER), str. 61–70.

<https://publications.jrc.ec.europa.eu/repository/handle/JRC128415>

<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

software <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>

<https://www.csb.gov/bp-america-texas-city-refinery-explosion/>

The background of the slide features a blue gradient with a glowing green tree at the top center. In the bottom left, a blue robotic hand is visible, holding a globe that has a circuit board pattern on it. The text is overlaid on a dark blue rectangular area.

Upozornění ohledně financování a použití ve vzdělávání

Tento otevřený vzdělávací zdroj slouží k podpoře výuky a diskuse. Nepředstavuje právní, regulační, bezpečnostní ani odborné inženýrské poradenství. Při aplikaci jeho obsahu se řiďte platnými zákony, normami, institucionálními zásadami a oborově specifickými pokyny.

Spolufinancováno programem Erasmus+ Evropské unie.



Ethical Engineer

www.ethicalengineer.eu

Děkujeme

Dokončili jste **Modul 1**



Co-funded by
the European Union