

Modul 2

Zaujatost (Bias), spravedlnost a
inkluzie v systémech umělé
inteligence



Co-funded by
the European Union



Ethical Engineer

www.ethicalengineer.eu

OBSAH

- 01 Diagnostika zaujatosti v systému
- 02 Ověření spravedlnosti v kontextu
- 03 Návrh spravedlivých a inkluzivních kontrolních mechanismů
- 04 Přijetí a obhajoba inženýrského rozhodnutí



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.



Přehled kurzu „Etický inženýr“

Hlavní obsah a výstupy pro každého studenta v jednotlivých modulech

Modul 01	Modul 02	Modul 03	Modul 04	Modul 05	Modul 06
Etika a transparentnost	Zaujatost, spravedlnost a inkluze	Soukromí a ochrana osobních údajů	Odpovědnost a bezpečnost	Návrh založený na hodnotách	Udržitelnost a dopad
Etické zásady a hodnoty EU Transparentnost a vysvětlitelnost Analýza případů z oblasti umělé inteligence a robotiky VÝSTUP: mapa etických otázek	Zaujatost v průběhu celého životního cyklu umělé inteligence Cíle spravedlnosti a důkazy Příklady inkluzivního inženýrství VÝSTUP: audit zaujatosti	Ochrana soukromí již od návrhu a GDPR Správa dat a PET Posuzování a napadnutelnost VÝSTUP: návrh s ohledem na ochranu soukromí	Role a odpovědnost Selhání, odolnost a předání Protokolování a poučení z incidentů VÝSTUP: záznam o sledovatelnosti	Požadavky na etický design Volba dat, modelů a rozhraní Kompromisy a agilní kontrolní body VÝSTUP: rozhodovací matice	Energie, voda a hardware společenské dopady a dopady v průběhu životního cyklu Metriky, správa a přepracování VÝSTUP: plán udržitelnosti

ZÁVĚREČNÁ PRÁCE: zkombinujte šest výstupů do etického inženýrského spisu s využitím jednoho případu z WP2 nebo inženýrského projektu

Představení modulu

Účel

Vybavte studenty technických oborů schopností identifikovat, měřit a omezovat nespravedlivé výsledky v systémech umělé inteligence a zároveň začleňovat inkluzivní postupy v týmech a mezi zainteresovanými stranami.

Cílová skupina

Studenti vysokých škol v oborech inženýrství, informatiky, robotiky a příbuzných oborech.

Navazuje na

Modul 1: etické zásady, transparentnost a záznamy o inženýrských rozhodnutích.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Představení modulu

Vzdělávací sekvence

Identifikujte zaujatost, vyberte důkazy, navrhnete kontrolní mechanismy a obhájte technické rozhodnutí.

Výstupy studenta

Mapa cest zaujatosti, plán testování spravedlnosti, kontrolní seznam pro začlenění již ve fázi návrhu, záznam o auditu zaujatosti a stručný popis technického rozhodnutí.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Výstupy učení

Na konci modulu budou studenti schopni:

1. Vysvětlujte, jak mohou formulace problému, data, označení, vlastnosti, modely, zpětnovazební smyčky a předpoklady týmu vést ke zkreslení.
2. Rozlišujte cíle spravedlnosti a vybírejte důkazy vhodné pro inženýrský kontext.
3. Vyhodnocujte výkon a dopady napříč relevantními a intersekcionálními skupinami.
4. Aplikujte principy diverzity a inkluze na týmovou spolupráci, zapojení zainteresovaných stran a návrh systému.
5. Navrhujte technické, organizační, participativní a lidské kontrolní mechanismy.
6. Zaznamenávejte kompromisy, zbytková rizika, monitorování a postupy pro zpochybnění rozhodnutí.

Zaujatost se stává inženýrským problémem, když mění výsledky

- Systémy umělé inteligence klasifikují, řadí, předpovídají a přidělují. Tyto funkce ovlivňují přístup k práci, zdravotní péči, vzdělání, financím a veřejným službám. K nespravedlivým výsledkům může dojít i v případech, že je model celkově technicky přesný.
- Technická otázka je konkrétní: Která volba v systému způsobuje rozdíl v zacházení nebo výsledku, pro koho, za jakých podmínek a na základě jakých důkazů?

Obhajitelná odpověď propojuje dotčené osoby s požadavky, daty, testy výkonu, lidskou kontrolou, nápravnými opatřeními a monitorováním.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Evropské rámce propojují spravedlnost s technickými důkazy

Etické pokyny EU identifikují rozmanitost, nediskriminaci a spravedlnost jako požadavky na důvěryhodnou AI.

Zákon o umělé inteligenci a Listina základních práv EU poskytují nejsilnější právní a práva jako opory: řízení rizik, správu dat, lidský dohled, důstojnost, nediskriminaci a rovnost.

Norma ISO 30415 poskytuje podporu v podobě pokynů pro inkluzivní organizační kulturu a rozhodování. DigComp podporuje kritické hodnocení digitálních informací, dat a technologií.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

01

Diagnostika zaujatosti (bias) v systému

Sledujte, jak mohou cíle, data, označení, modely a zpětnovazební smyčky ovlivňovat výsledky.

Zkreslení je vlastností systému

Zkreslení je systematická odchylka, která vede k zavádějícím nebo nespravedlivým výsledkům. Může vzniknout v důsledku nevyvážených dat, historické nerovnosti, úzkých perspektiv týmu, nedostatečného zastoupení zainteresovaných stran nebo provozních rozhodnutí.

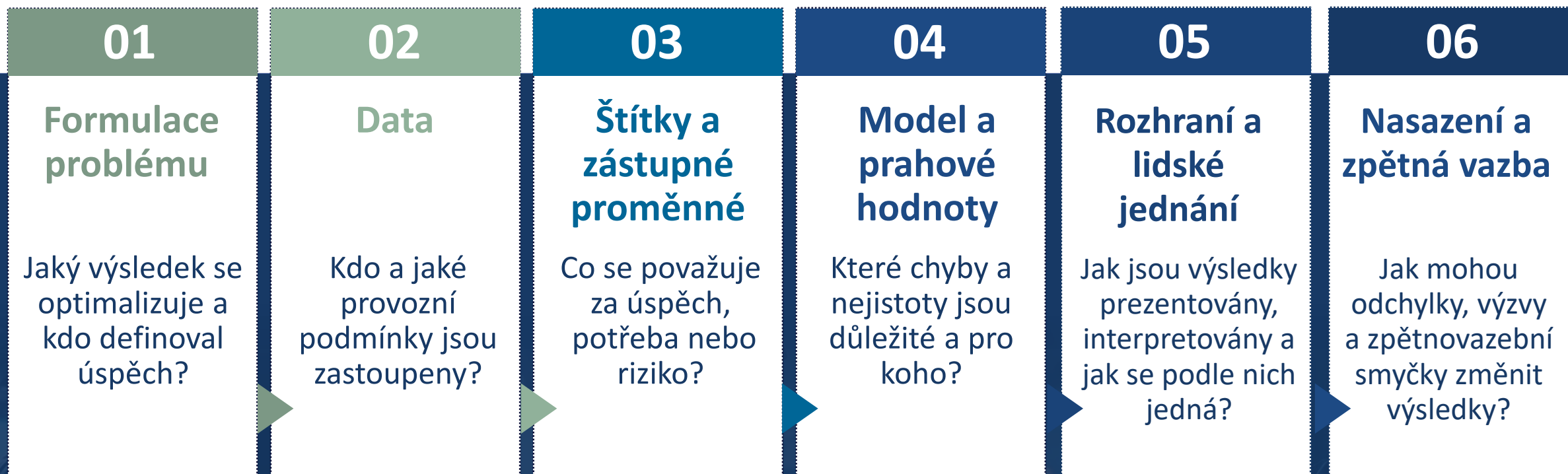
Data mohou odrážet historické nerovnosti, ale zkreslení se může projevit také prostřednictvím cíle, cílové proměnné, pravidel přiřazování štítků, návrhu charakteristik, prahové hodnoty, rozhraní nebo procesu nasazení.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Zkreslení se může objevit v každé fázi životního cyklu AI

Vyberte jedno technické rozhodnutí. V každé fázi identifikujte zdroj zaujatosti, postižené skupiny a požadované důkazy.



ZÁVĚR: Mapa cest zaujatosti propojující každou fázi s důkazy, dotčenými skupinami, aspekty diverzity a inkluze a praktickou kontrolou.

Analýza vzniku zkreslení: Nasazení autonomních taxíků

Případová studie:
Ethical Engineer:
Autonomní taxi

Rozhodnutí:

Schválit autonomní taxi pro smíšenou městskou čtvrť. V každé fázi se zaznamenává zaujatost, dotčené osoby, důkazy a kontrola.

01

○ Formulace problému

- **Zaujatost:** Úspěch se měří efektivitou vozového parku; vzácným rizikům se přikládá menší váha.
- **Dotčené skupiny:** chodci, cyklisté, cestující se zdravotním postižením a záchranné složky.
- **Důkazy/kontrola:** společně definovat bezpečnost a přístupnost; přiřadit odpovědnost za incidenty.

02

○ Data

- **Zkreslení:** Data upřednostňují jasné počasí, zmapované silnice a běžné dopravní vzorce.
- **Ovlivněno:** lidé v oblastech silničních prací, nízká viditelnost a špatně zmapované oblasti.
- **Důkazy/kontrola:** rozsah auditu, přidání okrajových případů a omezení provozní domény.

Analýza vzniku zkreslení: Nasazení autonomních taxíků

Případová studie:
Ethical Engineer:
Autonomní taxi

Rozhodnutí:
Schválit autonomní taxi pro
smíšenou městskou čtvrť. V
každé fázi se zaznamenává
zaujatost, dotčené osoby,
důkazy a kontrola.

03

Štítky a zástupné proměnné

- **Zkreslení:** Těsné vyhýbání se srážkám a zásahy člověka jsou označeny nekonzistentně.
- **Dotčené:** účastníci silničního provozu, jejichž chování nebo pohyb je špatně detekován.
- **Důkazy/kontrola:** sjednotit bezpečnostní označení; otestovat anotace a náhradní měřítko.

04

Model a prahové hodnoty

- **Zkreslení:** Prahové hodnoty pro detekci a zastavení představují kompromis mezi pohodlím a rizikem kolize.
- **Dotčené osoby:** chodci, cyklisté a cestující za výjimečných podmínek.
- **Důkazy/kontrola:** porovnejte chyby scénářů; použijte bezpečnostní rezervy a limity pro bezporuchový provoz.

Analýza vzniku zkreslení: Nasazení autonomních taxíků

Případová studie:
Ethical Engineer:
Autonomní taxi

Rozhodnutí:

Schválit autonomní taxi pro smíšenou městskou čtvrť. V každé fázi se zaznamenávají zaujatost, dotčené osoby, důkazy a kontrola.

05

Rozhraní a lidské jednání

- **Zaujatost:** Nejasnosti ohledně úrovně jistoty, předání řízení a možností v nouzových situacích.
- **Dotčené osoby:** cestující se zdravotním postižením, operátoři na dálku a záchranáři.
- **Důkazy/kontrola:** otestovat přístupná upozornění; zajistit postupy pro zásah a eskalaci.

06

Zavedení a zpětná vazba

- **Zaujatost:** Trasy s dobrým pokrytím generují více dat a jeví se jako bezpečnější.
- **Dotčené skupiny:** nedostatečně zastoupené okresy, jejich obyvatelé a řidiči, kteří jsou nuceni jet jinou trasou.
- **Důkazy/kontrola:** sledovat incidenty podle oblasti; provádět audit odchylek, stížností a dopadů změn.

Analýza vzniku zkreslení: Nasazení autonomních taxíků

Případová studie:
Ethical Engineer:
Autonomní taxi

Rozhodnutí

Schválit pouze v rámci ověřené provozní domény, s dostupnými ovládacími prvky, možností vzdáleného zásahu, odpovědností za incidenty a monitorováním na úrovni oblasti.

Provoz pozastavit v případě porušení bezpečnostních podmínek nebo podmínek pokrytí.

[Pokyny pro lektory]

- Dejte žákům dvě minuty na to, aby porovnali tuto hotovou mapu se svou vlastní mapou z předchozího snímku.
- Zeptejte se, který ovládací prvek mění návrh, který omezuje nasazení a který poskytuje nápravu po způsobení škody.
- Zdůrazněte, že konečné rozhodnutí je podmíněné. Důkazy určují, kde může systém fungovat a kdy musí přestat.

[Přístupnost]

- Přečtěte si šest fází v pořadí a poskytněte související případovou studii v přístupném digitálním formátu.

Volba vzorku, měření a označení s sebou nese různá rizika

K zkreslení výběru dochází, když chybí relevantní osoby nebo provozní podmínky. Historické zkreslení se objevuje, když jsou minulé nerovnosti považovány za neutrální důkazy.

K měřicí zaujatosti dochází, když proměnná zaznamenává skupiny odlišně. K zaujatosti při označování dochází, když cíl odráží nekonzistentní úsudek nebo nerovný proces.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Zástupné proměnné mohou skrývat hodnotová rozhodnutí

- Zástupné proměnné se používají, když nelze požadovaný pojem měřit přímo. Náklady na zdravotnictví mohou představovat parametr zdravotní potřeby. Minulé náборы mohou předpovídat budoucí pracovní výkon nově přijímaných zaměstnanců. Zatčení může prezentovat zástupný ukazatel pro páchaní trestné činnosti.
- Zástupné proměnné mohou být prediktivní, ale zároveň může vést k nespravedlivým výsledkům, protože je ovlivňována sociálními podmínkami. Inženýři by se měli ptát, co zástupná proměnná představuje pro každou skupinu, jaké kauzální faktory ji utvářejí a zda lze zamýšlené rozhodnutí ospravedlnit.

Zaznamenejte zástupnou proměnnou, její omezení a alternativní měřítko.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Zpětné vazby mohou proměnit počáteční zaujatost v systémový vzorec

- Výstupy umělé inteligence často mění data, která se shromažďují dále. Zvýšená kontrola vede k většímu počtu zaznamenaných incidentů. Více inspekcí odhalí více závad. Zamítnutí uchazečů neposkytují žádná pozdější data o výkonu.
- Zpětná vazba může způsobit, že se model jeví jako stále jistější, zatímco se zužuje rozsah toho, co systém může pozorovat.

Mezi technická opatření patří průzkumné vzorky, nezávislé zdroje dat, kontrafaktuální přezkum, sledování míry výběru a možnost, aby lidé mohli dodat chybějící důkazy.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Inkluze zlepšuje definici problému

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

- Inkluze znamená, že dotčené osoby mohou ovlivňovat požadavky, hodnocení a nápravná opatření. Zastoupení v týmu je důležité, zatímco účast na rozhodování představuje silnější technickou kontrolu.
- Řízení diverzity znamená plánovaný způsob, jakým organizace uznává lidské rozdíly a dává různým perspektivám vliv na práci a rozhodnutí. Systémové řízení diverzity to zviditelňuje v rámci správy: s kým se konzultuje, kdo může zpochybňovat předpoklady, kdo může požadovat důkazy a kdo může napadnout škodlivý cíl.
- Identifikujte přímé uživatele, osoby zastoupené v datech, osoby, kterých se rozhodnutí týkají, provozovatele, správce a komunity, na které má projekt nepřímý dopad. Zeptejte se, čí zkušenosti chybí a kdo může zpochybnit navrhovaný cíl.
- Psychologicky bezpečný tým umožňuje včas nahlásit nejistotu, nesouhlas a potenciální újmu.

Příklad z klinické praxe: Ne inkluzivní návrh srdeční pumpy

Případová studie: Ethical Engineer : Pulzní srdeční pumpa

Počáteční úkol: Definujte skupinu pacientů, kteří mohou profitovat z implantace VAD s pevným tepovým objemem 65 ml. **Navrhněte řešení pro pacienty různé tělesné konstituce, kteří se nacházejí mimo bezpečný rámec použití této pumpy**

01

○ Počáteční zadání

Potvrďte, že standardní zařízení je pro daného pacienta vhodné a splňuje principy beneficence a non-maleficence

02

○ Interdisciplinární tým

chirurg, anesteziolog, perfuzionista, biomedicínský inženýr, etický konzultant

03

○ Co ne-inkluzivní návrh pumpy odhalil?

Tělesné proporce se u různě velkých pacientů liší. Pevně stanovený tepový objem může u menších pacientů vést k arteriální hypertenzi, zatímco u větších pacientů může způsobit hypotenzi a nedostatečnou orgánovou perfuzi.

Příklad z klinické praxe: Ne inkluzivní návrh srdeční pumpy

Případová studie: Ethical Engineer : Pulzní srdeční pumpa

Počáteční úkol: Definujte skupinu pacientů, kteří mohou profitovat z implantace VAD s pevným tepovým objemem 65 ml. **Navrhněte řešení pro pacienty různé tělesné konstituce, kteří se nacházejí mimo bezpečný rámec použití této pumpy**

04

○ **Přeformulování problému**

Definujte bezpečný rámec použití pumpy: kterým pacientům může mechanická srdeční podpora (VAD 65 ml) prospět a pro které pacienty není vhodná

05

○ **Technické požadavky**

U pacientů, kteří se nacházejí v bezpečném rozmezí použití, použijte VAD se standardně doporučeným nastavením. U pacientů, u nichž se vyskytuje hypertenze, upravte nastavení přístroje s cílem dosáhnout optimálního arteriálního tlaku

06

○ **Rozhodnutí a přezkum**

VAD (65 ml) se standardním nastavením použijte pouze tehdy, pokud monitorace arteriálního tlaku potvrzuje hodnoty v normotenzním rozmezí. Při zjištění arteriální hypertenze stanovte další postup a povinnosti lékaře, včetně úpravy nastavení VAD, další monitorace pacienta a posouzení zbytkového rizika.

Příklad z klinické praxe: Ne inkluzivní návrh srdeční pumpy

Případová studie: Ethical
Engineer : Pulzní srdeční pumpa

Inkluzivní přístup změnil nastavení pumpy

Standardní implantace pumpy se proměnila v komplexní rozhodování o bezpečnosti pacienta, které vyžadovalo odchýlení od doporučeného nastavení pumpy s cílem naplnit principy beneficence a non-maleficence a zachránit život pacienta.

[Pokyny pro lektory]

- Požádejte studenty, aby určili přesný bod, ve kterém začlenění mění technický zadávací list.
- Vyzvěte jednu skupinu, aby obhajovala standardní zařízení, druhou přizpůsobení a třetí odložení. Každá skupina musí uvést důkazy a zbytkové riziko.
- Zdůrazněte, že účast vede ke vzniku požadavků a rozhodovacích kritérií. Jedná se o technické opatření, nikoli o konzultační cvičení přidané na závěr.

[Přístupnost]

- Přečtete nahlas každý řádek a poskytněte související případ srdeční pumpy v přístupném digitálním formátu.

02

Ověření spravedlnosti v kontextu

Vyberte cíle, měřítka, důkazy a
kritéria přijatelnosti spravedlnosti
pro inženýrský kontext.

Spravedlnost spojuje volbu systému s nerovnými účinky

Spravedlnost se týká toho, jak systém umělé inteligence rozděluje výhody, rizika, chyby, zpoždění a příležitosti mezi dotčené osoby. Celkové skóre přesnosti může skrývat závažné rozdíly mezi skupinami nebo provozními podmínkami.

Spravedlnost je proto jak technickým cílem návrhu, tak závazkem v oblasti lidských práv spojeným s důstojností, nediskriminací a rovností. Nespravedlnost se může projevit prostřednictvím cíle, dat, štítků, zástupných proměnných, prahových hodnot, rozhraní, lidského jednání nebo procesu nasazení. Inženýrskou otázkou je, zda volba systému vytváří neopodstatněné znevýhodnění.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Spravedlnost spojuje volby systému s nerovnými účinky

01

Spravedlnost Cíl

Uved'te, který výsledek, příležitost nebo chyba musí být spravedlivě rozděleny, a vysvětlete, proč je to důležité pro toto rozhodnutí.

02

Důkazy

Ověřte výkonnost a dopad napříč relevantními skupinami a podmínkami. Uved'te počty skupin, míru nejistoty a omezení dat.

03

Technické řešení Reakce

Stanovte kritéria přijatelnosti, technické a organizační kontroly, lidskou kontrolu, monitorování a postup nápravných opatření.

Cíle spravedlnosti mohou být v rozporu. Inženýři musí zdůvodnit zvolený cíl ve vztahu k rozhodnutí, dotčených osob a důsledků chyby. Mezi běžné cíle patří rovný přístup, rovné příležitosti, srovnatelné míry chybovosti, kalibrace, individuální konzistence a spravedlivý proces.

**Měřítkem
spravedlnosti jsou
důkazy, nikoli
rozhodnutí**

Uvádějte absolutní výsledky, míru nejistoty a počty skupin. Zkoumejte intersekcionalní skupiny tam, kde jsou k dispozici dostatečné důkazy. Kombinujte kvantitativní testy s kvalitativními důkazy o procesu, důstojnosti, inkluzi, přístupnosti a přístupu k nápravě.

**Začněte
rozhodnutím a
dotčenou populací**

Hodnotící plán by měl odrážet realistické podmínky, včetně chybějících dat, kvality senzorů, jazyka, zdravotního postižení, environmentálních odchylek a toho, zda byly při sběru požadavků zastoupeny dotčené zainteresované strany.

Rozčlenit výkonnost a uvést nejistotu

- Celková přesnost může zakrývat slabý výkon u menší skupiny. Uvedte hodnoty matice záměny, kalibraci nebo kvalitu specifickou pro daný úkol podle příslušné skupiny a provozních podmínek.
- Uvedte velikost vzorku, interval spolehlivosti nebo jiné prohlášení o nejistotě. Pokud jsou důkazy příliš omezené, zaznamenejte tuto mezeru a omezte nasazení, místo abyste prezentovali nestabilní výsledky jako důkaz spravedlnosti.

Testy opakujte po změnách modelu, dat nebo prahových hodnot a během provozu.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Intersekcionalita mění návrh testů

- Lidé vnímají systémy prostřednictvím několika charakteristik a okolností najednou. Model může fungovat adekvátně pro ženy obecně a pro určitou rasovou skupinu obecně, zatímco pro ženy v rámci této skupiny může fungovat špatně.
- Intersekcionalní analýza vyžaduje dostatečné množství dat a pečlivou ochranu soukromí. Kombinujte cílené kvantitativní testování s konzultacemi se zainteresovanými stranami a analýzou scénářů.

Zaznamenejte, které průsečíky byly vyhodnoceny, které zůstávají nejisté a jak bude monitorování odhalovat škody po nasazení.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Prahové hodnoty rozdělují chyby a přínosy

- Skóre neurčuje výsledek, dokud nejsou zvoleny prahová hodnota a pravidlo pro akci. Změna prahové hodnoty ovlivňuje počet falešně pozitivních a falešně negativních výsledků, pracovní zátěž a přístup.
- Porovnejte navrhované prahové hodnoty s využitím výsledků dotčených skupin a skutečných důsledků jednotlivých chyb. Zvažte zavedení kontrolního pásma, v němž budou nejisté případy posuzovány lidmi.

Záznam o rozhodnutí by měl uvádět zvolenou prahovou hodnotu, kritéria přijatelnosti, postup přezkumu, odpovědnou osobu a spouštěč změny.

[Přístupnost]

Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Případová studie: AI v náborových procesech

Analyzujte případ náboru jako systém

Technické zaměření: historická data, definice cílů, zástupné charakteristiky, testování podskupin, vysvětlení, spornost a validace pracovních požadavků s ohledem na diverzitu.

Ověřte, zda historické náborové představy představují platný ukazatel schopností. Prozkoumejte charakteristiky týkající se vzdělání, přestávek v kariéře a jazykových znalostí jako možné zástupné proměnné. Porovnejte míru výběru a chybovost napříč relevantními skupinami. Zkontrolujte, zda náboráři mohou skóre přepsat, zda se na definici požadavků na pracovní pozici podíleli různí zainteresovaní aktéři a zda mohou uchazeči napadnout chybu.

Případová studie: AI v rozpoznávání obličeje

Otevřete případovou studii Ethical Engineer o rozpoznávání obličeje a definujte relevantní chybu pro zvolené použití.

- Tento projektový případ se zabývá rozpoznáváním obličeje, které se používá ke zvýšení bezpečnosti a zefektivnění procesů, jako je odbavení na letišti. Přínosy závisí na spolehlivém porovnání totožnosti, zatímco ochrana soukromí, bezpečnost dat a demografická výkonnost představují významná rizika.
- **Technické zaměření:** kontext použití, kvalita obrazu, porovnání typu „jeden k jednomu“ nebo „jeden k mnoha“, důsledky chyb, demografické testy a lidská kontrola.

Důkazy NIST mění rozhodnutí o rozpoznávání obličeje

- NIST otestoval velké množství algoritmů a u většiny z nich zjistil demografické rozdíly, přičemž výsledky závisely na algoritmu, aplikaci a datech. To znamená, že obecné tvrzení o rozpoznávání obličeje představuje slabý důkaz.
- Uved'te verzi algoritmu, provozní režim, podmínky snímání a populaci. Otestujte chybu spojenou s újmou. Falešná shoda při sledování se liší od neúspěšného ověření u brány.

***Zajistěte lidskou kontrolu, kontroly kvality,
možnost odvolání a monitorování incidentů***

Praktický úkol: Napište plán testování spravedlnosti

- Vyberte případ náboru zaměstnanců nebo rozpoznávání obličeje. Vypracujte plán, který by mohl realizovat jiný tým inženýrů.
- Definujte rozhodnutí, populaci, skupiny, podmínky a důsledky chyb. Vyberte dvě kvantitativní měřítka a jeden kvalitativní zdroj. Uveďte kontroly kvality dat, minimální požadavky na vzorek, hlášení nejistoty a kritéria přijatelnosti.
- Přidejte pravidlo pro rozhodování v případě nedostatečných důkazů, odpovědnou osobu, datum přezkumu a postup, jakým mohou dotčené osoby výsledek napadnout.
- Přidejte 5bodový kontrolní seznam „Inclusion-by-Design“ (začlenění již ve fázi návrhu), který zahrnuje rozmanitost dat, zapojení zainteresovaných stran, přístupnost, dokumentaci, testování a nasazení.
- **Výstup:** jednostránkový testovací plán, kontrolní seznam a dvouminutové odůvodnění.

[Přístupnost] Nabídněte úkol v písemné, mluvené nebo vizuální formě. Přečtěte scénář nahlas, vysvětlete zkratky a umožněte přístupné digitální vypracování.

03

Design Fair & inkluzivní ovládací prvky

Převést zjištění do technických,
organizačních, participativních a
monitorovacích kontrolních
mechanismů.

Zahajte zmírňující opatření ještě před trénováním modelu

Před modelováním vypracujte kontrolní seznam „Inclusion-by-Design“. Zdokumentujte původ dat, účel, souhlas nebo právní základ, známé mezery, zamýšlené použití a omezení reprezentativnosti. Zahrňte přezkum ze strany zainteresovaných stran v případech, kdy cíl, zástupná proměnná nebo zamýšlené použití mohou zakódovat sociální znevýhodnění.

Vzorkování a vážení použijte pouze tehdy, pokud řeší identifikovaný mechanismus.

[Přístupnost]

- Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Kontroly trénování a následného zpracování vyžadují stanovený cíl spravedlnosti

- Omezení modelu, převažování, převzorkování, učení reprezentace a úprava prahových hodnot mohou změnit výsledky jednotlivých skupin. Jejich vhodnost závisí na zvoleném cíli spravedlnosti a kontextu rizika.
- Po každé změně porovnejte výkon, spravedlnost, nejistotu a robustnost. Zkontrolujte, zda zisky pro jednu skupinu nevedou k nepřijatelným ztrátám pro jinou skupinu nebo k oslabení bezpečnosti.

Zaznamenejte kompromis a zachovejte výchozí stav, aby byl účinek zásahu stále viditelný.

Lidská kontrola musí být navržena jako skutečná kontrolní skupina

Navrhněte spouštěč přezkumu, zobrazení důkazů, možnosti rozhodnutí, postup eskalace a záznam. Otestujte konzistentnost a pracovní zátěž přezkoumavatelů. Dohledové týmy by měly mít relevantní technické, oborové a praktické zkušenosti. Poskytněte dotčeným osobám jasné oznámení a přístupný postup k nápravě nebo odvolání.

Rozmanité týmy potřebují rozhodovací pravomoci a důkazy

Rozmanitost týmu může odhalit slepá místa, avšak samotné zastoupení nezaručuje vliv. Norma ISO 30415 definuje inkluzi jako kulturu, spravedlnost, respekt a odpovědnost; inženýrská praxe musí toto převést do rozhodovacích práv. Definujte, kdo může zpochybnit cíl, pozastavit nasazení, požádat o důkazy od podskupin a schválit zbytkové riziko.

Externí případ: Rozpoznávání obličeje ve společnosti Rite Aid

- Federální obchodní komise Spojených států obvinila společnost Rite Aid z toho, že používala rozpoznávání obličeje bez přiměřených postupů k prevenci újmy spotřebitelů. Stížnost popisovala falešně pozitivní výstrahy a selhání v testování, kvalitě obrazu, školení personálu, monitorování a posuzování rizik.
- **Inženýrské ponaučení:** výkon podskupiny je pouze jednou součástí kontrolního systému. K újmě přispívá také kvalita obrazu, pravidla databáze, jednání operátora, eskalace a důkazy o incidentech.

Zeptejte se, která kontrolní opatření měla zabránit nasazení a které důkazy měly vyvolat zastavení.

Proměňte lekce Rite Aid v kontrolní opatření

- **Problém spravedlnosti málokdy má jediné technické řešení.**

Případ Rite Aid ukazuje, jak může dojít ke škodám v důsledku několika navzájem propojených selhání – nízké kvality vstupních dat, nedostatečných pravidel pro porovnávání, chování operátorů, nedostatečného školení a chybějícího řízení. Účinné kontrolní mechanismy proto musí fungovat v rámci celého systému.

- **Kvalita obrazu**

Odmítněte vstupy, které nedosahují ověřené prahové hodnoty kvality.

- **Důkazy o shodě**

V příslušných případech zobrazujte míru spolehlivosti shody, limity kvality obrazu a alternativní kandidáty.

Přeměňte lekce Rite Aid na kontrolní prvky

○ Opatření provozovatele

Před jakýmkoli nepříznivým opatřením vyžadujte nezávislé posouzení. Samotné upozornění nesmí nikdy rozhodovat o výsledku

○ Školení

Ověřujte kompetence zaměstnanců prostřednictvím realistických scénářů a opakovaných kontrol.

○ Monitorování

Sledujte falešné poplachy, postižené skupiny, stížnosti a nově se objevující vzorce škod.

Proměňte poučení z případu Rite Aid v kontrolní opatření

○ Řízení

Vyžadujte schválení před rozšířením použití a definujte jasné spouštěče pozastavení.

○ Technický princip

Každé kontrolní opatření musí mít měřitelné kritérium přijatelnosti, odpovědného vlastníka a důkaz, že funguje

Dilema: Spravedlivé rozdělení při chybějících datech

Rozhodnutí

POKRAČOVAT, PŘEPRAVIT nebo ZASTAVIT.

Definujte spravedlivé pravidlo pro nejisté případy, minimální důkazy, lidskou kontrolu, přístupné odvolání, monitorování, odpovědného vlastníka a to, jak jsou v procesu přezkumu zastoupeni zralí studenti, zahraniční studenti nebo studenti s přerušným studiem.

[Přístupnost] Nabídněte úkol v písemné, ústní nebo vizuální podobě. Přečtěte scénář nahlas, vysvětlete zkratky a umožněte přístupné digitální vyplnění.

04

Přijměte a obhajujte inženýrské rozhodnutí

Proved'te audit zaujatosti,
zaznamenejte zbytkové
riziko a obhajte rozhodnutí
na základě důkazů.

Projektový případ A: Zkreslení v datech

- Případová studie „Ethical Engineer“ ukazuje, jak mohou nereprezentativní nebo historicky zaujatá data vést k nepřesným předpovědím a nerovným výsledkům. Využijte ji k identifikaci konkrétního procesu generování dat, namísto toho, abyste zaujatost považovali za obecnou vlastnost.
- Určete, kdo v datech chybí, která proměnná je zkreslená a jak to ovlivňuje rozhodnutí. Navrhněte jedno opatření pro kontrolu dat, jedno opatření pro kontrolu systému a jedno monitorovací opatření.

Otevřete případovou studii „Zaujatost v datech“ a zaznamenejte zbytkovou nejistotu.

Projektová případová studie B: Využití AI při hodnocení historických fotografií

- Tato případová studie se zabývá systémem umělé inteligence, který předpovídá, které historické snímky mohou přilákat pozornost nebo generovat příjmy. Data získaná z internetu mohou odměňovat popularitu, zatímco citlivým snímkům odnímají historický kontext.
- Technické dilema se týká cíle, trénovacího signálu, zastoupení komunit a rozhraní používaného kurátory.

Otevřete případ hodnocení historických fotografií. Rozhodněte, jak by měly kulturní kontext, posouzení zainteresovanými stranami a omezení automatizovaného řazení ovlivnit návrh.

Porovnejte případy před výběrem opatření ke zmírnění

Přizpůsobte kontrolu mechanismu zkreslení, nikoli označení „zkreslení“.

- **Zkreslení v datech začíná u kvality důkazů:** kdo v datech chybí, která proměnná je zkreslená a jak to ovlivňuje rozhodnutí.
- **Hodnocení historických fotografií přidává konflikt hodnot:** popularita, výnosy a kulturní význam mohou směřovat různými směry.

Před výběrem kontrolního opatření si položte čtyři otázky:

- Jaká volba systému způsobuje znevýhodnění?
- Jaké důkazy by prokázaly, zda je újma skutečná?
- Která kontrolní opatření mění mechanismus: cíl, data, prahová hodnota, rozhraní, přezkum nebo omezení nasazení?
- Jaké zbytkové riziko zůstává po zavedení kontrolního opatření?

[Přístupnost] - Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Týmová aktivita: Proved'te audit zaujatosti

Vyberte si jeden případ z Ethical Engineer. Přidělte role pro systémy, data, testování, zainteresované strany a zapisovatele.

1. Definujte zamýšlené použití a kritické rozhodnutí.
2. Zmapujte cestu zaujatosti.
3. Vyberte cíl spravedlnosti a důkazy.
4. Navrhněte technická a organizační kontrolní opatření.
5. Uveďte chybějící důkazy, zbytkové riziko a nápravná opatření.
6. Rozhodněte se: POKRAČOVAT, PŘEPRAVIT NEBO ZASTAVIT.

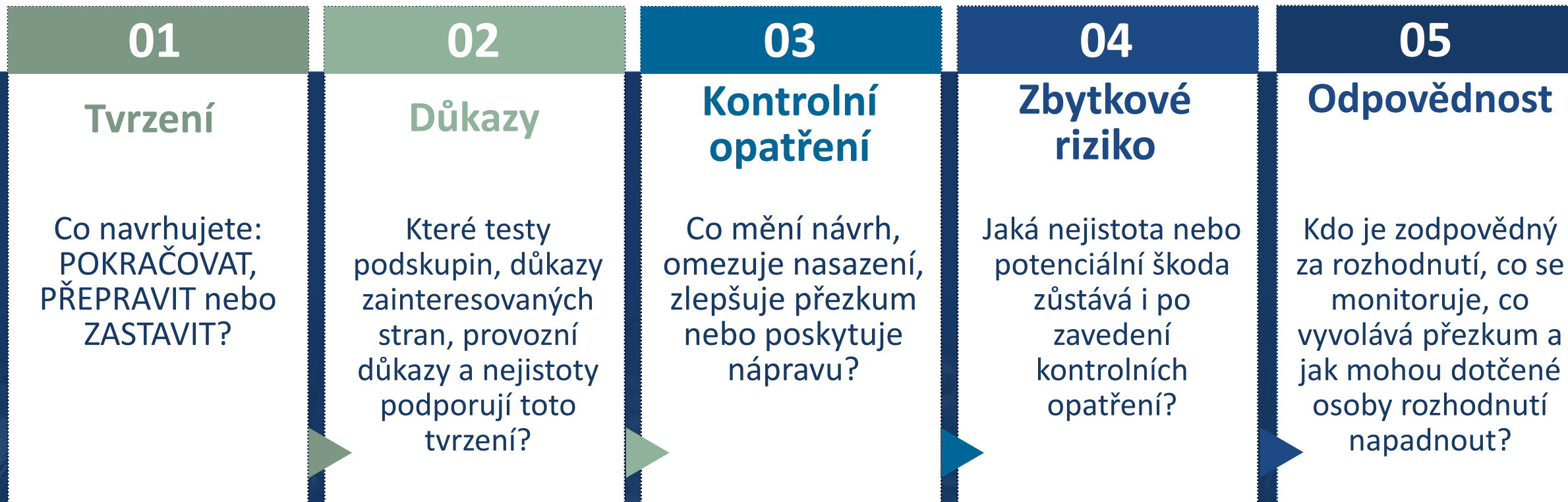
Výstup

Vyplněný záznam z auditu a dvouminutové shrnutí rozhodnutí založené na důkazech.

[Přístupnost] – Nabídněte úkol v písemné, mluvené nebo vizuální formě. Přečtěte scénář nahlas, vysvětlete zkratky a umožněte přístupné digitální vyplnění.

Od analýzy k obhajitelnému technickému rozhodnutí

Důkladný audit zaujatosti promění důkazy v podložený argument



Pracovní list: Záznam o rozhodnutí týkající se zaujatosti a spravedlnosti

System a zamýšlené použití:

Kritické rozhodnutí a dotčené osoby:

Mechanismus/cesta vzniku zaujatosti:

Cíl spravedlnosti a důvod, proč je vhodný:

Důkazy, skupiny, podmínky a nejistota:

Začlenění / důkazy zainteresovaných stran:

[Přístupnost] Úkol nabídněte v písemné, ústní nebo vizuální podobě. Přečtěte scénář nahlas, vysvětlete zkratky a umožněte přístupné digitální vyplnění.

Pokračování na dalším snímku

Technická kontrola a akceptační test:

Řízení / Kontrola inkluze a vlastník:

Rozhodnutí: POKRAČOVAT PŘEPRAVIT ZASTAVIT

Zbytkové riziko a spouštěč monitorování:

Nápravné opatření / Odvolací cesta:

Nesouhlas nebo alternativní názor:

[Přístupnost] Úkol nabídněte v písemné, ústní nebo vizuální podobě. Přečtěte scénář nahlas, vysvětlete zkratky a umožněte přístupné digitální vyplnění.

Role play: Komise pro přezkum veřejných zakázek

- Město plánuje nákup systému umělé inteligence, který stanoví priority při inspekcích bytů. Dodavatel poskytuje celkovou přesnost, ale žádné výsledky pro jednotlivé podskupiny nebo čtvrti. Historická data o stížnostech jsou v oblastech s jazykovými bariérami a nízkou důvěrou sporadická.
- **Role:** vedoucí zadávacího řízení, stavební inženýr, dodavatel, zástupce komunity, pověřenec pro ochranu osobních údajů a vedoucí bezpečnosti.

Dohodněte se na důkazech a smluvních podmínkách požadovaných před zahájením pilotního projektu. Uveďte jednu podmínku pro REVIZE a jednu pro ZASTAVENÍ. Zaznamenejte, kdo může pozastavit používání a jak mohou obyvatelé napadnout opomenutí.

[Přístupnost] Nabídněte úkol v písemné, ústní nebo vizuální podobě. Přečtěte scénář nahlas, vysvětlete zkratky a umožněte přístupné digitální vyplnění.

Individuální reflexe & Přenos

Napište krátkou
technickou
argumentaci.

1. Která volba systému vyvolala největší obavy ohledně spravedlnosti nebo inkluze?
2. Který cíl spravedlnosti jste si vybrali a co zůstává nevyřešeno?
3. Čí perspektiva nebo zkušenost v původních důkazech chyběla?
4. Které nové důkazy by vás přiměly změnit vaše rozhodnutí?
5. Které technické, organizační a inkluzivní kontrolní mechanismy by měly být součástí požadavků?
6. Jaké zbytkové riziko byste akceptovali a co by vás přimělo zastavit nasazení?
7. Jak by se vaše uvažování změnilo v jiné technické disciplíně?

Výstup:

Propojte dotčené osoby, důkazy, kontrolní opatření a odůvodněné rozhodnutí.

[Přístupnost] Použijte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Kvíz

1. Proč může vysoká celková přesnost koexistovat s nespravedlivými výsledky?
2. Jaký je rozdíl mezi zkreslením výběru a zkreslením šitéků?
3. Proč musí být metrika spravedlnosti vázána na kontext použití?
4. Co by měl inženýr nahlásit, když jsou důkazy o podskupinách omezené?
5. Uvedte dvě podmínky nezbytné pro smysluplnou lidskou kontrolu.
6. Která kontrolní metoda řeší neplatnou zástupnou proměnnou nejvíce přímo?
7. Co by měl obsahovat záznam o rozhodnutí týkající se spravedlnosti?
8. Jak mohou diverzita a inkluze zlepšit inženýrský proces?

[Přístupnost] Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.

Odpovědi na kvíz a zdůvodnění výuky

1. Celkové výsledky mohou skrývat chyby specifické pro jednotlivé skupiny a podmínky.
2. Zkreslení výběru se týká reprezentace; zkreslení označení se týká cíle nebo zaznamenaného posouzení.
3. Metriky vyjadřují různé cíle spravedlnosti a důsledky chyb.
4. Uveďte mezery a nejistoty, poté omezte nebo revidujte použití.
5. Relevantní jsou autorita, čas, informace, kompetence a záznam.
6. Přehodnoťte cíl, vlastnost nebo proces, který vedl k vytvoření náhradního ukazatele.
7. Rozhodnutí, dotčené osoby, cíl, důkazy, kontrolní mechanismy, vlastník, zbytkové riziko, monitorování a nápravná opatření.
8. Rozšířením důkazů, odhalením předpokladů, posílením přezkumu a poskytnutím dotčeným osobám možností napadnout rozhodnutí.

Kontrola výstupů učení

Před dokončením si ověřte, že jste schopni:

1. Vysledujte zaujatost v systému umělé inteligence a kontextu jeho rozhodování.
2. Vyberte a odůvodněte cíl spravedlnosti.
3. Vyhodnoťte výkon napříč relevantními a intersekcionálními skupinami, včetně nejistoty.
4. Vysvětlete, jak diverzita, participace a inkluze mění požadavky a důkazy.
5. Doporučte technické, organizační, participativní a lidské kontrolní mechanismy.
6. Obhajujte rozhodnutí „POKRAČOVAT / REVIZOVAT / ZASTAVIT“ s ohledem na zbytková rizika, monitorování a nápravná opatření

Předložené důkazy:

- Mapa cest zkreslení
- Plán testování spravedlnosti
- Inkluze již ve fázi návrhu
- Kontrolní seznam
- Audit zaujatosti
- Stručné shrnutí technického rozhodnutí.

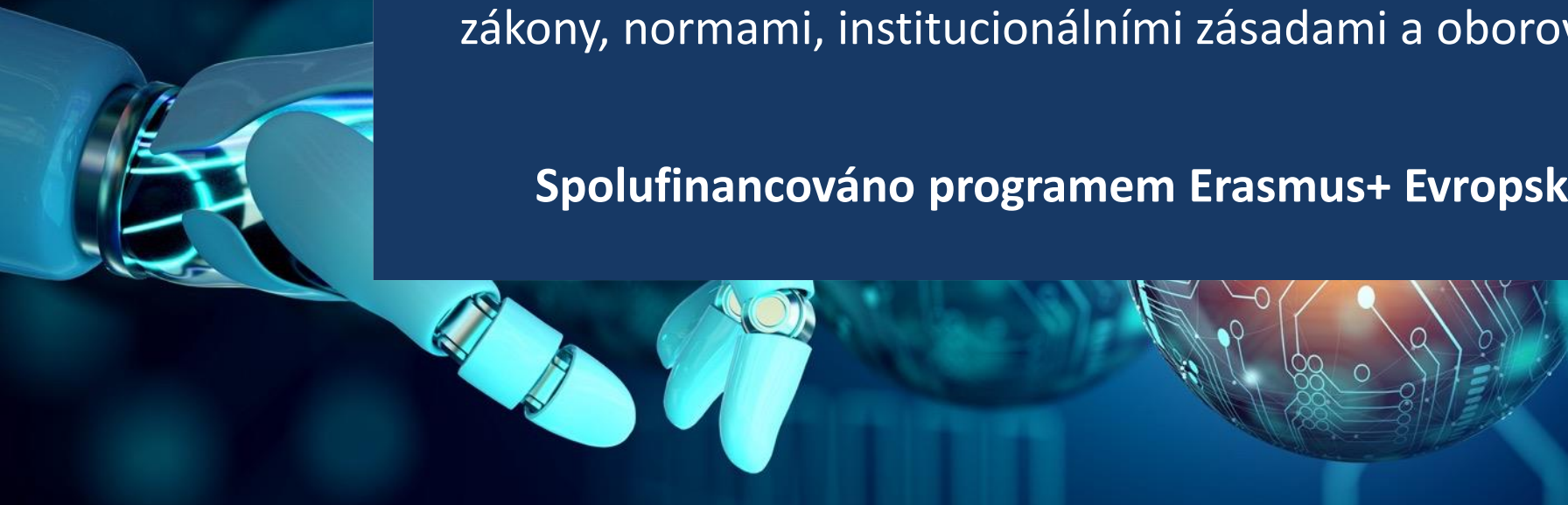
[Přístupnost] Používejte přístupné pořadí čtení, na požádání popište relevantní vizuální prvky značky a poskytněte text prezentace v přístupném digitálním formátu.



Oznámení o financování a použití ve vzdělávacích účelech

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 oborovými pokyny.

Spolufinancováno programem Erasmus+ Evropské unie.





Ethical Engineer

www.ethicalengineer.eu

Děkujeme

Dokončili jste modul 2



Co-funded by
the European Union