

Modul 5

Návrh založený na
hodnotách v inženýrství
umělé inteligence



Co-funded by
the European Union



Ethical Engineer

www.ethicalengineer.eu

OBSAH

- 01 Objevte zainteresované strany a hodnoty
- 02 Převést hodnoty do designu
- 03 Porovnání hodnot napříč případy
- 04 Obhajoba rozhodnutí založeného na hodnotách



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 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

Zaměření modulů

Tento modul převádí hodnoty zainteresovaných stran na požadavky, alternativy návrhu, důkazy z testování a obhajitelné inženýrské rozhodnutí.

Účel

Vybavit studenty technických oborů schopností zkoumat hodnoty, převádět je do konstrukčních kritérií a obhajovat svá rozhodnutí před zainteresovanými stranami na základě technických důkazů.

Cílová skupina

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

Navazuje na

Moduly 1 až 4: etika, transparentnost, spravedlnost, ochrana soukromí, odpovědnost, bezpečnost a dohledatelná inženýrská rozhodnutí.

[Zdroje]

- Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení.
- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>

Zaměření modulů

Tento modul převádí hodnoty zainteresovaných stran na požadavky, alternativy návrhu, důkazy z testování a obhajitelné inženýrské rozhodnutí.

Vzdělávací sekvence

Objevit hodnoty zainteresovaných stran, převést je do návrhu, porovnat případy a obhájit rozhodnutí založené na hodnotách.

Výstupy studentů

Mapa zainteresovaných stran, hierarchie hodnot, požadavky, srovnání variant a záznam o návrhu založeném na hodnotách.

[Zdroje]

- Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení.
- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>

Výstupy učení

Na konci modulu budou studenti schopni:

1. Identifikujte přímé a nepřímé zainteresované strany, rozdíly v moci a relevantní hodnoty.
2. provádějte koncepční, empirická a technická šetření v rámci Value Sensitive Design.
3. Překládejte hodnoty do požadavků, kritérií přijatelnosti a sledovatelných důkazů o návrhu.
4. porovnávejte možnosti dat, modelů, rozhraní a systémů s využitím prahových hodnot, důkazů a nejistoty.
5. Vytvořte návrhovou dokumentaci založenou na hodnotách, která zahrnuje bezpečnostní opatření, zbytková napětí a podmínky přezkumu.

[Zdroje]

- Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení.
- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>
- <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Hodnoty jsou již přítomny v inženýrských rozhodnutích

- Cíle definují úspěch. Datové sady rozhodují o tom, či zkušenosti jsou viditelné. Prahové hodnoty rozdělují riziko. Rozhraní utvářejí pozornost a kontrolu.
- Inženýrství umělé inteligence je sociotechnické: chování vychází z lidí, dat, modelů, rozhraní, organizací a provozního kontextu.
- Návrh založený na hodnotách činí tato rozhodnutí explicitními, podloženými důkazy a přezkoumatelnými. Úkolem inženýrů je ukázat, jak určitá hodnota ovlivnila rozhodnutí a jak bude požadavek ověřen.

[Zdroje]

- Původní prezentace k modulu 5 poskytnutá konsorciem projektu Ethical Engineer.
- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>

Rámce a metody jako vodítko pro inženýrská rozhodnutí

Návrh citlivý na hodnoty využívá koncepční, empirické a technické výzkumy k prozkoumání lidských hodnot v rámci celého procesu navrhování. IEEE 7000 poskytuje proces systémového inženýrství pro zjišťování hodnot a jejich promítání do konceptů, požadavků a řešení etických rizik.

Etické pokyny EU stanovují požadavky na důvěryhodnou umělou inteligenci. Zákon EU o umělé inteligenci stanovuje právní povinnosti pro definované systémy a subjekty. Tyto zdroje slouží jako vodítko pro související práci a nenahrazují důkazy zainteresovaných stran ani standardy oborového inženýrství.

Používejte aktuální znění právních předpisů a zaznamenávejte, jaká metoda, povinnost a důkazy podporují každé významné rozhodnutí.

[Zdroje]

- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>
- <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727>

Současné povinnosti a technická odpovědnost

Zákon EU o umělé inteligenci přiřazuje povinnosti podle systému, kategorie rizika a role aktéra. Technické týmy musí tyto povinnosti propojit s příslušnými právními předpisy, bezpečnostními normami a organizačními kontrolními mechanismy.

U každého významného rozhodnutí je třeba zaznamenat: osobu odpovědnou za rozhodnutí, osobu odpovědnou za důkazy, revizora, kontrolní orgán a orgán s pravomocí zastavit proces.

Odpovědnost se stává prakticky uplatnitelnou, jsou-li před nasazením systému zřejmé odpovědnost, důkazy, schvalovací procesy a postupy pro eskalaci.

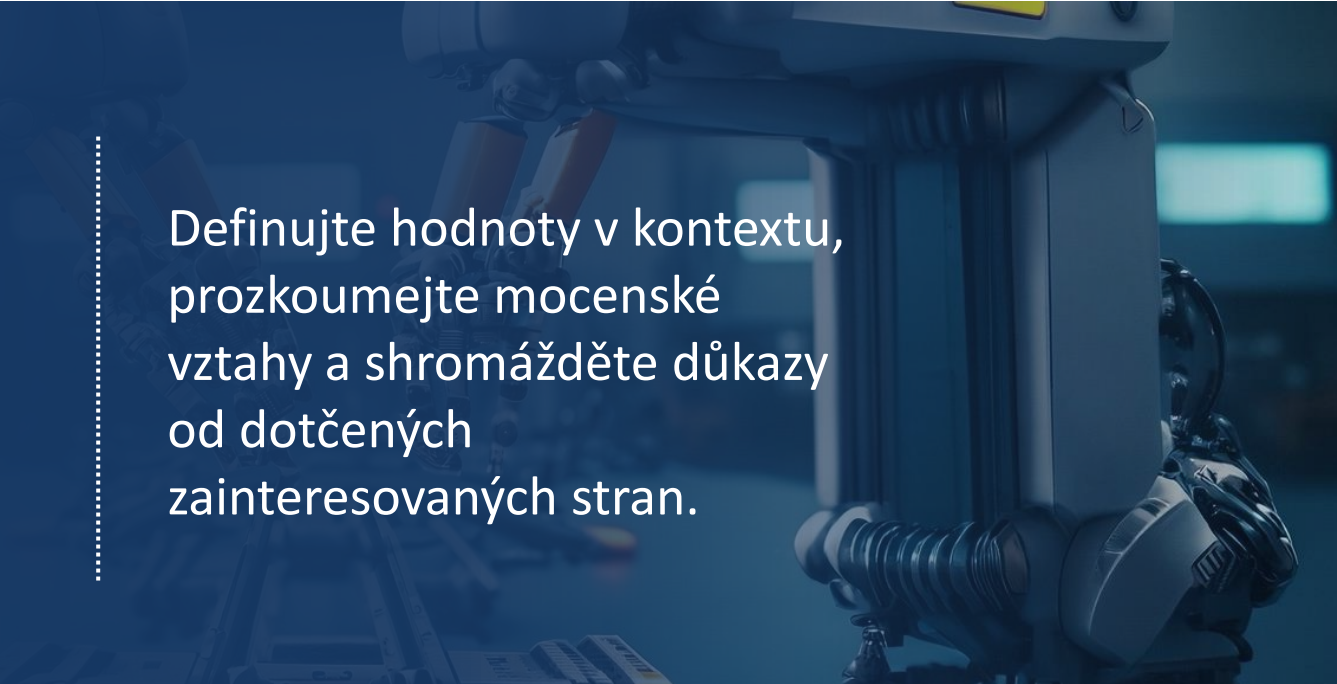
[Zdroje]

- <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727>
- <https://www.nist.gov/itl/ai-risk-management-framework>
- Nejnovější prezentace k modulu 5 poskytnutá konsorciem projektu Ethical Engineer.

01

Objevte zainteresované strany a hodnoty

Definujte hodnoty v kontextu,
prozkoumejte mocenské
vztahy a shromážděte důkazy
od dotčených
zainteresovaných stran.



Návrh založený na hodnotách propojuje etický účel s volbami systému

Hodnota vyjadřuje to, co lidé nebo instituce považují za důležité, jako je bezpečnost, důstojnost, autonomie, spravedlnost, soukromí, přístupnost, spolehlivost nebo udržitelnost. Její význam závisí na lidech, praxi a provozním kontextu.

Hodnoty ovlivňují cíle, data, funkce, rozhraní, výchozí nastavení, obchodní pravidla a údržbu. Mohou se navzájem podporovat nebo vytvářet napětí.

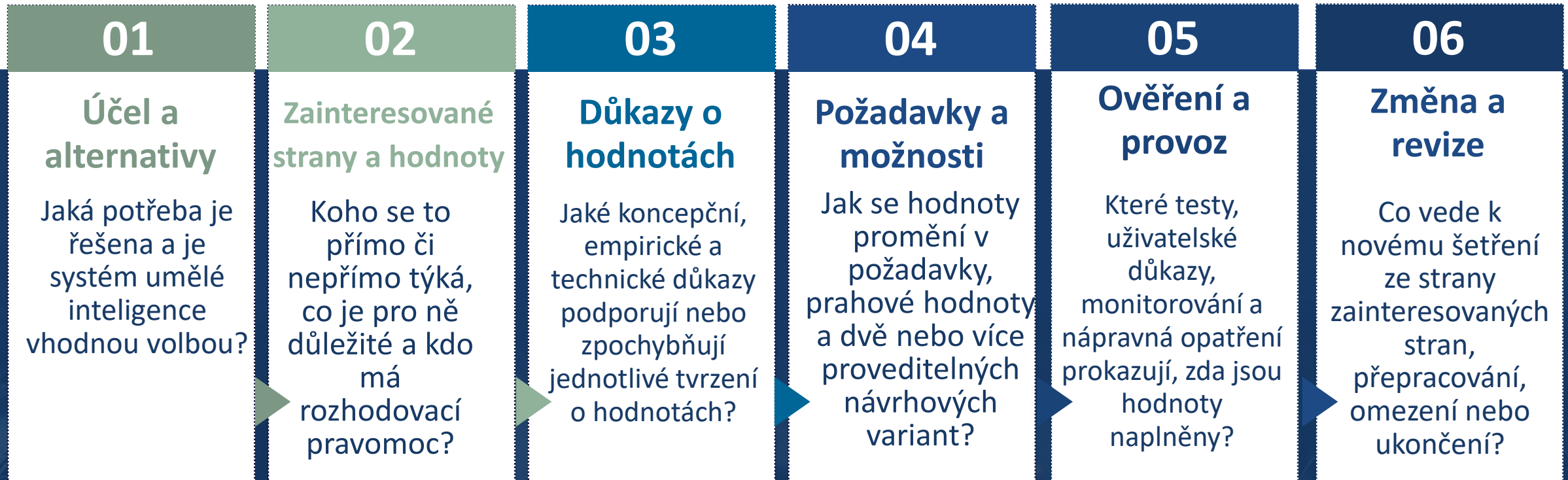
Rozhodování založené na hodnotách identifikuje zainteresované strany, důkazy a alternativy a následně promítá vybraný výklad hodnoty do požadavků, testů, vlastníků a podmínek přezkumu.

[Zdroje]

- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>

Zmapujte hodnoty v rámci celého inženýrského životního cyklu

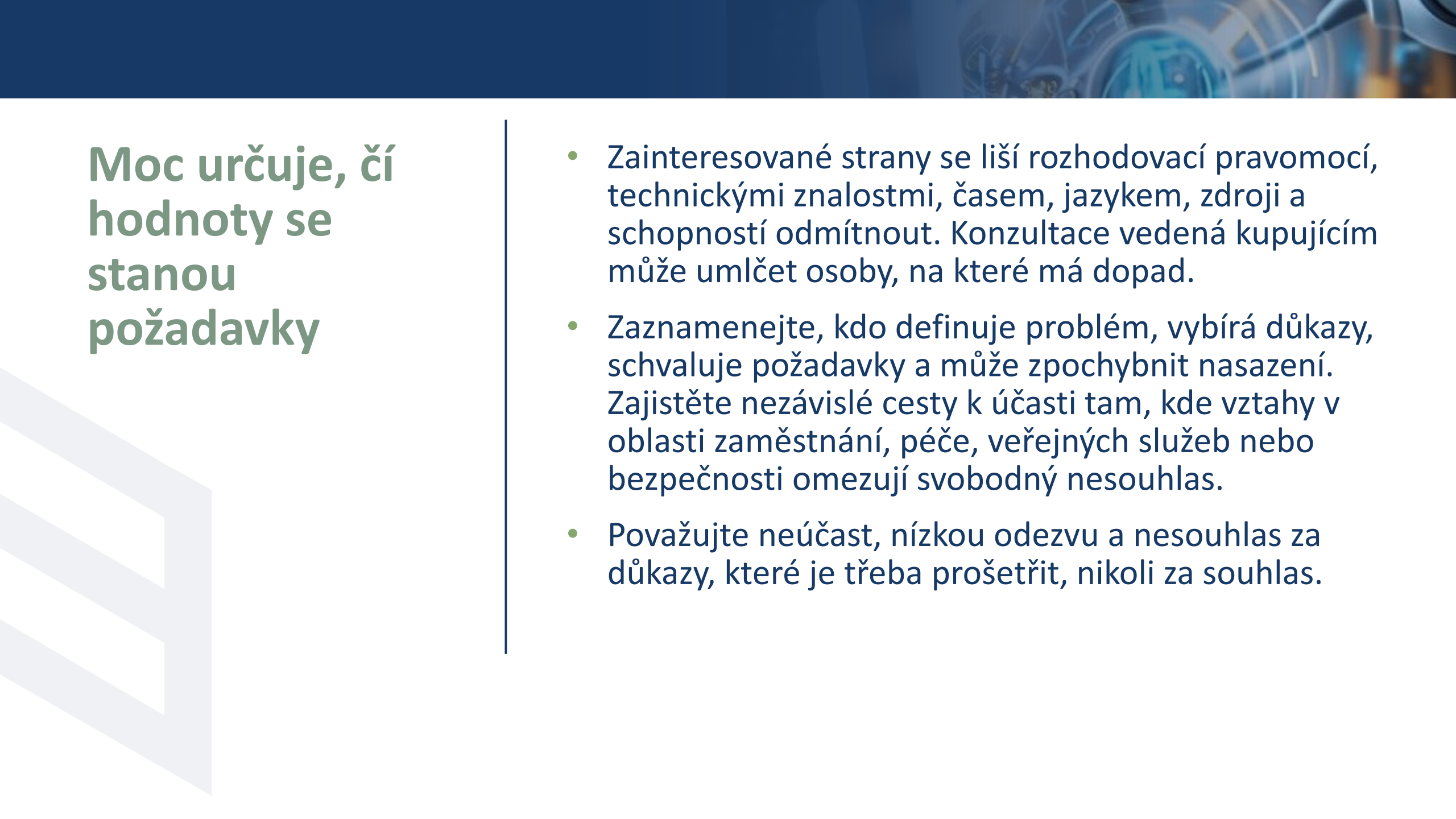
Vyberte jedno technické rozhodnutí. V každé fázi identifikujte zainteresované strany, tvrzení o hodnotách, důkazy, návrhové řešení a přezkoumání.



VÝSTUP Mapa hodnot propojující každou fázi se zainteresovanými stranami, důkazy, návrhovým řešením, kritérii přijatelnosti a spouštěčem revize.

Přímé a nepřímé zainteresované strany pocítují různé dopady

- Přímé zainteresované strany systém používají, provozují, udržují nebo přijímají rozhodnutí z něj vycházející. Nepřímé zainteresované strany mohou být ovlivněny, aniž by se systému přímo dotýkaly; patří mezi ně členové rodiny, komunity, okolní osoby, budoucí pracovníci, dodavatelé a lidé zastoupení v trénovacích datech.
- Uveďte, kdo získává schopnosti, kdo nese riziko, kdo poskytuje práci nebo data a kdo může být vyloučen. Zahrňte osoby, na které má dopad předvídatelné zneužití a selhání systému.
- Mapa zainteresovaných stran je neúplná, pokud obsahuje pouze kupující, uživatele a regulační orgány.



Moc určuje, čím hodnoty se stanou požadavky

- Zainteresované strany se liší rozhodovací pravomocí, technickými znalostmi, časem, jazykem, zdroji a schopností odmítnout. Konzultace vedená kupujícím může umlčet osoby, na které má dopad.
- Zaznamenejte, kdo definuje problém, vybírá důkazy, schvaluje požadavky a může zpochybnit nasazení. Zajistěte nezávislé cesty k účasti tam, kde vztahy v oblasti zaměstnání, péče, veřejných služeb nebo bezpečnosti omezují svobodný nesouhlas.
- Považujte neúčast, nízkou odezvu a nesouhlas za důkazy, které je třeba prošetřit, nikoli za souhlas.

Definujte každou hodnotu v konkrétním inženýrském kontextu

Bezpečnost (MSafety) může znamenat prevenci úrazů, udržování bezpečného stavu nebo podporu včasného zásahu. Autonomie se může týkat informované volby, kontroly, svobody od manipulace nebo praktické schopnosti odmítnout. Spravedlnost se může týkat výsledků, procesu, přístupu nebo nápravy.

Formulujte kontextovou definici, uveďte dotčené zainteresované strany, pravděpodobnou hrozbu a pozorovatelné důkazy. Vyhněte se používání obecných hodnotových pojmů, jako by všichni sdíleli stejný výklad.

Definice by měla být dostatečně přesná, aby mohla sloužit jako vodítko pro návrh řešení a akceptační test.

Důkazy od zainteresovaných stran musí ovlivnit volbu návrhu

V závislosti na rozhodnutí využijte rozhovory, pozorování, společný návrh, průzkumy, studie použitelnosti, stížnosti a důkazy z terénu.

Zahrňte dotčené osoby, koncové uživatele a odborníky z technické, právní, etické a oborové oblasti. Odborníci nenahrazují ty, na které má dopad.

Zaznamenejte, co se změnilo: definici problému, požadavek, variantu, test nebo podmínky přezkumu — a kdo zůstává nezastoupen.

Napětí mezi hodnotami vyžaduje důkazy a omezení

- Hodnoty se mohou navzájem podporovat: srozumitelné vysvětlení může zlepšit autonomii, bezpečnost a důvěru. Mohou však také být v rozporu: podrobné protokolování může podporovat odpovědnost, ale zároveň zvyšovat riziko narušení soukromí; personalizace může zlepšit přizpůsobení, ale zároveň snížit standardizaci a udržitelnost.
- Nejprve definujte nepřekročitelné právní, bezpečnostní nebo práva týkající prahové hodnoty. Poté porovnejte proveditelné možnosti s využitím relevantních důkazů. Nepopisujte zamezitelnou újmu jako vyvážený kompromis.
- Zaznamenejte zbývající rozpor, dotčené zainteresované strany a důkazy, které by vedly k přehodnocení rozhodnutí.

PRAKTICKÝ PŘÍKLAD Z OBLASTI STROJÍRENSTVÍ: NÁVRH SRDEČNÍ PUMPY

Tým navrhuje jeden profil řízení podporovaný umělou inteligencí pro pulzní srdeční pumpu s cílem zjednodušit výrobu a schvalovací proces. Důkazy zainteresovaných stran ukazují, že různí pacienti potřebují ověřené provozní rozsahy a nastavení zařízení



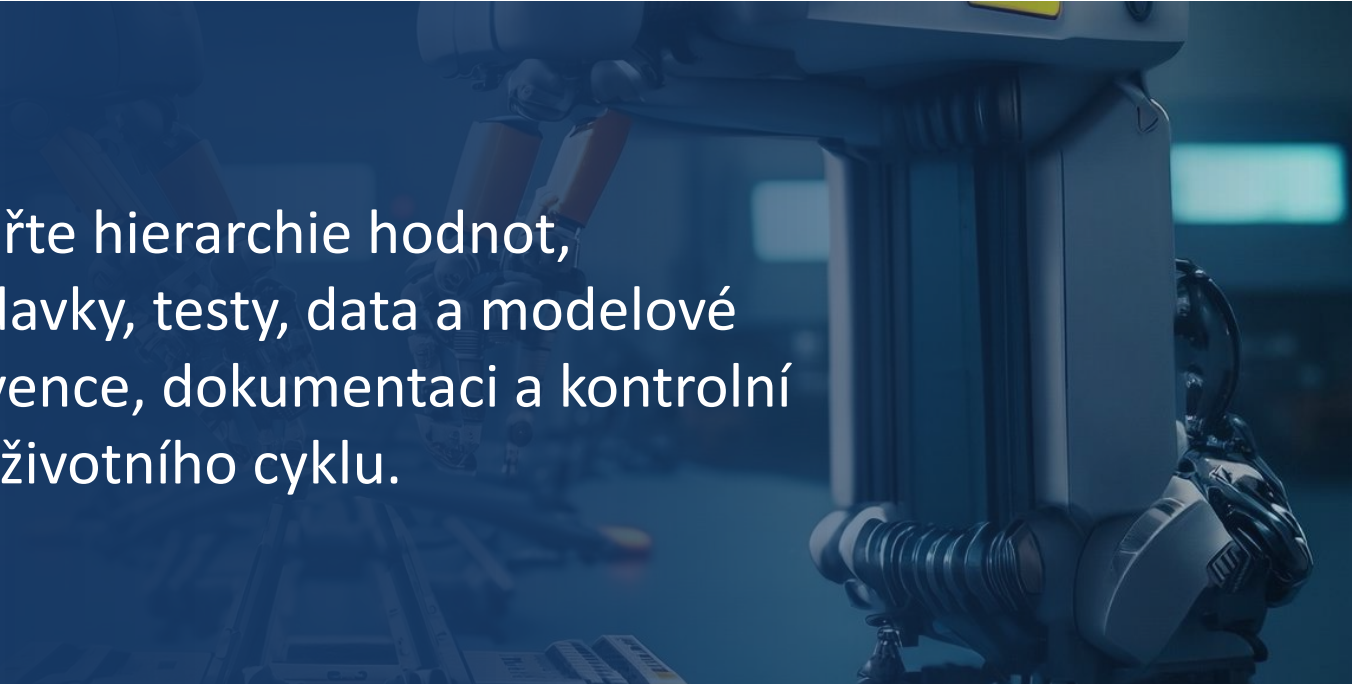
Nežádoucí událost změnila nastavení pumpy Standardizace vedla ke změně řídicí architekturu s limity specifickými pro jednotlivé pacienty s přezkoumatelnými důkazy.



02

Převést hodnoty do návrhu

Vytvořte hierarchie hodnot,
požadavky, testy, data a modelové
intervence, dokumentaci a kontrolní
body životního cyklu.



Tři typy zkoumání tvoří jednu iterativní metodu návrhu

Návrh citlivý na hodnoty kombinuje koncepční, empirické a technické výzkumy. Každý z nich může ovlivnit ostatní v závislosti na vývoji problému, důkazů a návrhu.

Tyto tři typy zkoumání opakovaně využívejte. Zjištění jedné ze zainteresovaných stran může změnit požadavek; prototyp může odhalit novou zainteresovanou stranu nebo rozpor mezi hodnotami.

Viz další snímek...

[Zdroje]

- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>

Tři vyšetřování tvoří jednu metodu iterativního návrhu

01

Koncepční

Identifikujte přímé a nepřímé zainteresované strany, definujte hodnoty v kontextu a prozkoumejte rozpory, práva a předpoklady.

02

Empirický

Zjistěte, jak lidé technologii chápou, prožívají a jak se v souvislosti s ní chovají, a to pomocí vhodných sociálních a terénních metod.

03

Technické

Analyzujte, jak stávající a alternativní návrhy podporují nebo podkopávají hodnoty, poté vytvořte prototyp a ověřte požadavky.

Výstupem je sledovatelný řetězec: definice hodnoty, podklady od zainteresovaných stran, návrhové požadavky, technické řešení, akceptační test, vlastník a spouštěč revize.

Hierarchie hodnot umožňuje přehlednost revize

- Začněte kontextovou hodnotou a poté identifikujte podpůrné normy nebo principy návrhu. Přeložte každou normu do jednoho nebo více požadavků a měřitelných kritérií přijatelnosti.
- Příklad: autonomie pracovníka → monitorování je omezeno na bezpečnost zařízení → individuální skóre produktivity jsou zakázána → ověřte, že rozhraní, databáze a zprávy je nemohou generovat.
- U každého spojení zaznamenejte důkazy, vlastníka a stav. Pokud nelze požadavek vysledovat až k hodnotě a potřebě zainteresované strany, zpochybňte jeho prioritu.

Etické uživatelské příběhy vyžadují testovatelná akceptační kritéria

- Užitečný příběh identifikuje dotčenou roli, schopnost nebo ochranu a důvod, proč je to důležité. Uvádí také podmínky zneužití, vyloučení nebo selhání.
- Příklad: Jako technik údržby potřebuji, aby systém zobrazoval spolehlivost senzorů a nedávné změny konfigurace, abych mohl posoudit, zda je výstraha důvodem k zásahu.
- Důkazy o přijatelnosti: test reprezentativního úkolu, čitelná nejistota, historie verzí, doba odezvy na výstrahu a zdokumentovaný záložní plán. Propojte příběh s požadavkem a záznamem o ověření.

Zásahy do dat s sebou nesou volby hodnot

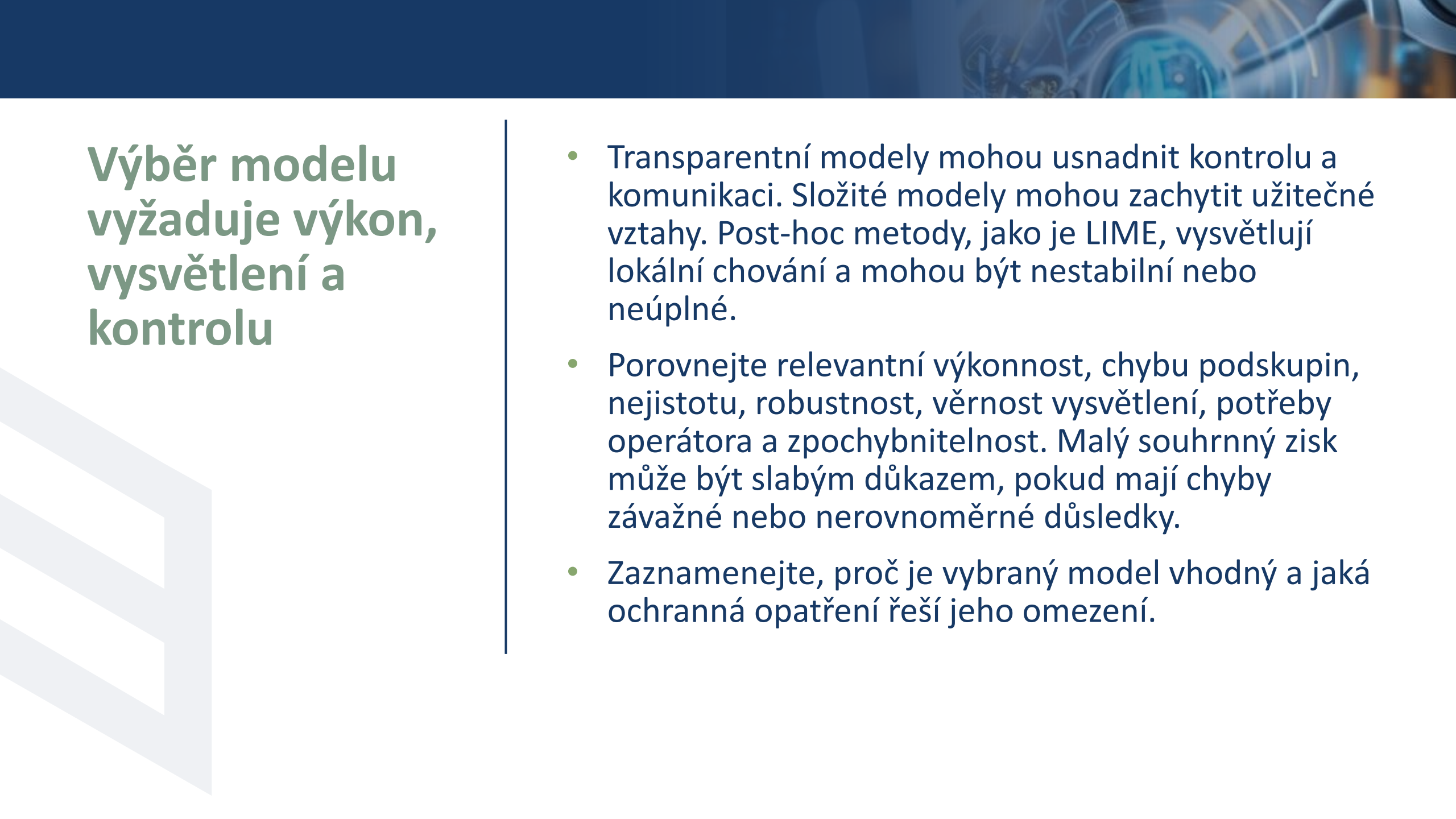
- Výběr vzorků, štítky, zástupné proměnné, pravidla pro chybějící data a augmentace určují, kterým případům se věnuje pozornost. SMOTE nebo vážení tříd mohou podporovat vzácné třídy při změně kalibrace. Přenos učení importuje předpoklady z jiné domény. PCA nebo výběr vlastností mohou zlepšit zpracovatelnost a zároveň odstranit informace relevantní pro podskupinu.
- U každého zásahu uveďte zamýšlenou hodnotu, dotčenou skupinu, mechanismus, srovnávací základnu a důkazy pro hodnocení.
- Zaznamenejte původ, omezení, zakázaná použití a rozhodnutí o zachování nebo odstranění vlastnosti.

Nerovnováha dat a SMOTE vyžadují pečlivé vyhodnocení

Vzácné třídy mohou odrážet problémy s výběrem vzorku, označováním, přístupem nebo měřením. Před výběrem technického řešení prošetřete příčinu.

Mezi možnosti patří sběr nových dat, stratifikované hodnocení, vážení tříd a syntetické nadměrné vzorkování menšin. Metoda SMOTE vytváří syntetické příklady menšin mezi pozorovanými případy; může podporovat učení při změně kalibrace nebo vytvářet nepravděpodobné případy, pokud je prostor vlastností nevhodný.

Porovnejte výchozí stav a zásah pomocí chyby podskupiny, kalibrace, nejistoty a přezkumu domény. Zaznamenejte, jakou hodnotu zásah podporuje a jaká rizika přetrvávají.



Výběr modelu vyžaduje výkon, vysvětlení a kontrolu

- Transparentní modely mohou usnadnit kontrolu a komunikaci. Složité modely mohou zachytit užitečné vztahy. Post-hoc metody, jako je LIME, vysvětlují lokální chování a mohou být nestabilní nebo neúplné.
- Porovnejte relevantní výkonnost, chybu podskupin, nejistotu, robustnost, věrnost vysvětlení, potřeby operátora a zpochybnitelnost. Malý souhrnný zisk může být slabým důkazem, pokud mají chyby závažné nebo nerovnoměrné důsledky.
- Zaznamenejte, proč je vybraný model vhodný a jaká ochranná opatření řeší jeho omezení.

Interpreovatelnost a LIME poskytují lokální důkazy

LIME aplikuje jednoduchý lokální náhradní model na jednu predikci, aby odhadl, které vlastnosti ovlivnily tento lokální výsledek.

Může pomoci inženýrům prozkoumat chování, prodiskutovat případ a identifikovat otázky pro další testování. Vysvětlení se může měnit v závislosti na výběru vzorku, parametrech a korelovaných vlastnostech.

LIME nestanovuje kauzalitu, globální chování ani spravedlnost. Otestujte stabilitu při opakovaných bězích a v podobných případech, porovnejte s odbornými znalostmi dané oblasti a zaznamenejte rozhodnutí, které vysvětlení podporuje, a omezení sdělená uživatelům.

Rozhraní utvářejí pozornost, autoritu a výzvy

Výchozí nastavení, pořadí výstrah, zobrazení spolehlivosti, popisky a načasování pracovního postupu ovlivňují důvěru.

Automatizace mění roli člověka. Když se operátoři stanou supervizory, spolehlivá automatizace může snížit bdělost a zpomalit zásah.

Otestujte, zda uživatelé dokáží rozpoznat omezení, znovu získat přehled o situaci, včas zasáhnout, přepsat rozhodnutí, požádat o přezkoumání a zotavit se po chybě.

Zaznamenejte informace, které každá cílová skupina potřebuje, a akce, které musí rozhraní podporovat.

Dostatečnost datového souboru závisí na rozhodnutí

- Neexistuje žádný univerzální poměr vzorku k vlastnosti, který by určoval dostatečnost. Důkazy závisí na úkolu, populaci, modelu, variabilitě a důsledcích chyby.
- Vyhodnoťte relevanci, původ, pokrytí podskupin a vzácných případů, nejistotu, intervaly spolehlivosti, externí validaci, posun distribuce, úniky a duplikace.
- Uveďte mezery a provozní limity. Více dat může zvýšit riziko narušení soukromí, náklady nebo historickou zaujatost; posuďte kvalitu, účel a objem.

Objem dat, dimenzionalita a transferové učení

Neexistuje žádné univerzální pravidlo pro velikost vzorku. Dostatečnost závisí na úkolu, populaci, variabilitě, modelu a důsledcích chyby.

Vysokorozměrná data mohou vést k řídkosti a přetrénování, i když se celkový vzorek jeví jako velký. Využijte učební křivky, kontrolu vlastností, regularizaci, odhady nejistoty a externí validaci.

Přenos učení a augmentace mohou snížit nároky na data a zároveň importovat předpoklady z jiné domény. Zkontrolujte vhodnost pro danou doménu, pokrytí podskupin a posun distribuce.

Dokumentace propojuje návrh, důkazy a přezkum

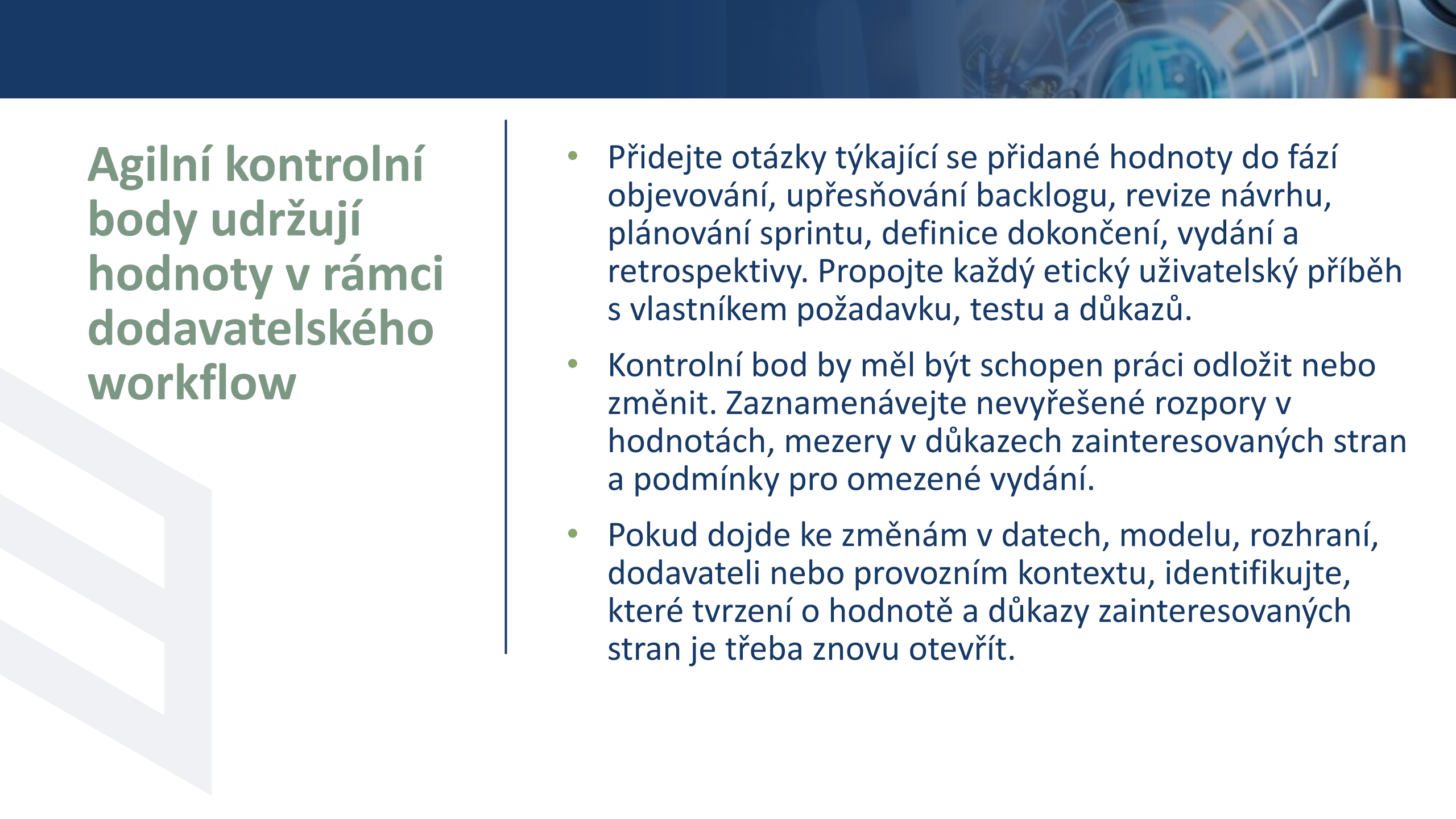
- Datový list sady dat zaznamenává motivaci, složení, sběr, zpracování, doporučené použití a omezení. Kartu modelu zaznamenává zamýšlené použití, hodnocení, výsledky podskupin, omezení a etické aspekty. Kartu systému propojuje komponenty, rozhraní, uživatele, provozní kontext a bezpečnostní opatření.
- Dokumentaci využívejte při revizích návrhů, zadávání zakázek, řízení změn a poučení se z incidentů. Udržujte ji ve verzích a propojte ji s důkazním materiálem.
- Dokumentace podporuje odpovědnost, když mění rozhodnutí, odhaluje mezeru nebo umožňuje přezkum.

Datové listy, karty modelů a karty systémů

Datové listy datových sad zaznamenávají motivaci, složení, sběr, zpracování, zamýšlené použití a omezení.

Karty modelu zaznamenávají zamýšlené použití, hodnocení, výsledky podskupin, omezení a etické aspekty. Karty systému propojují komponenty, rozhraní, uživatele, kontext a bezpečnostní opatření.

Záznamy udržujte ve verzích a propojte je s důkazy, schváleními a řízením změn. Tam, kde se uplatňuje zákon EU o umělé inteligenci, propojte je s technickou dokumentací, řízením rizik a důkazy o monitorování po uvedení na trh.



Agilní kontrolní body udržují hodnoty v rámci dodavatelského workflow

- Přidejte otázky týkající se přidané hodnoty do fází objevování, upřesňování backlogu, revize návrhu, plánování sprintu, definice dokončení, vydání a retrospektivy. Propojte každý etický uživatelský příběh s vlastníkem požadavku, testu a důkazů.
- Kontrolní bod by měl být schopen práci odložit nebo změnit. Zaznamenávejte nevyřešené rozpory v hodnotách, mezery v důkazech zainteresovaných stran a podmínky pro omezené vydání.
- Pokud dojde ke změnám v datech, modelu, rozhraní, dodavateli nebo provozním kontextu, identifikujte, které tvrzení o hodnotě a důkazy zainteresovaných stran je třeba znovu otevřít.

Monitorování musí detekovat odchylky v hodnotách i v modelu

- Technický výkon může zůstat stabilní, zatímco se mění sociální kontext. Nová populace, pracovní postupy, pobídky, politika nebo sekundární využití mohou podkopat původní hodnotovou nabídku.
- Sledujte relevantní výsledky, stížnosti, výjimky, vyloučení, provizorní řešení, incidenty a distribuční účinky. Zajistěte přístupnou cestu pro zpětnou vazbu a nápravu.
- Pro každé měřítko definujte vlastníka, frekvenci, prahovou hodnotu a opatření. Jakmile důkazy překročí spouštěcí hodnotu, znovu zhodnoťte zainteresované strany a požadavky.

Před váženým bodováním možností aplikujte povinné prahové hodnoty

Začněte s právními, bezpečnostními, právy a profesními omezeními. Odstraňte všechny varianty, které nesplňují povinnou prahovou hodnotu, pokud nelze návrh revidovat.

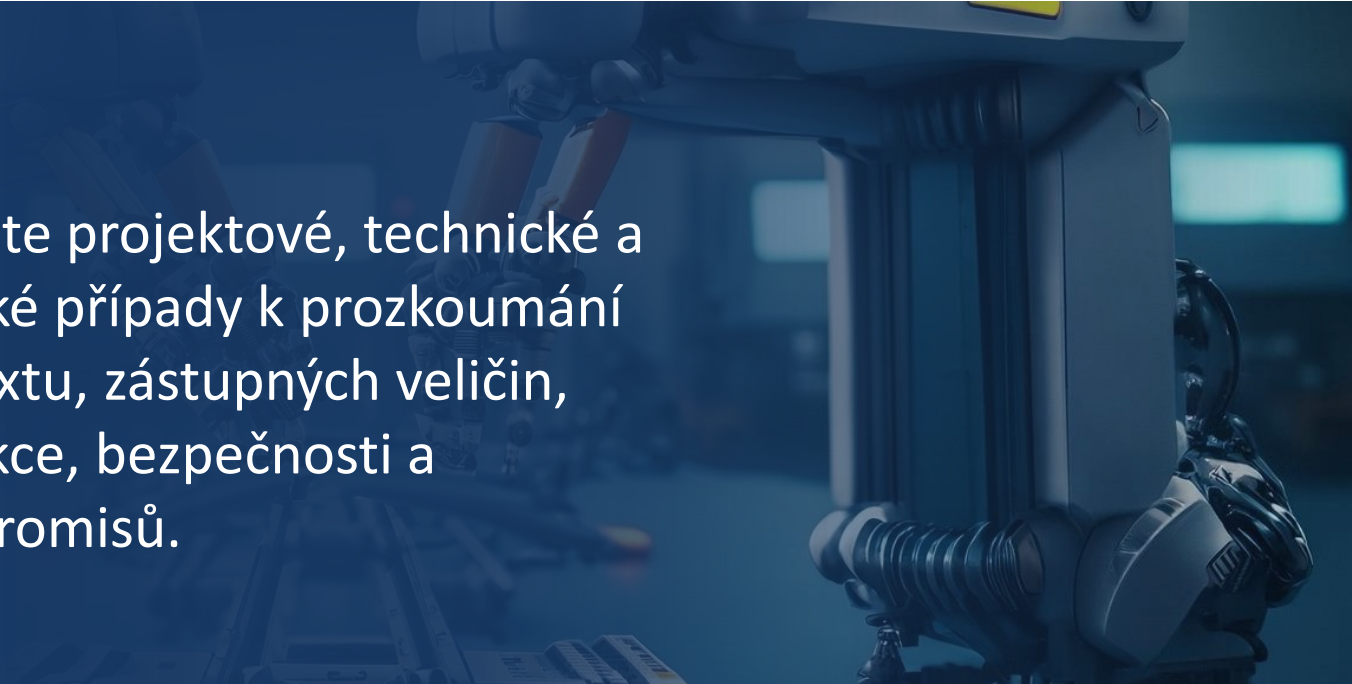
Propojte návrh založený na hodnotách s řízením rizik: identifikujte újmu nebo ohrožení hodnot, vyhodnoťte důkazy a důsledky, vyberte kontrolní opatření, ověřte jeho účinnost, zdokumentujte zbytkové riziko a přezkoumejte spouštěcí prahy.

Poté porovnejte proveditelné varianty pomocí kritérií, jako jsou bezpečnost, autonomie, spravedlnost, soukromí, vysvětlitelnost, výkon, udržitelnost, náklady a udržitelnost. Vážený součet nemůže vyvážit závažné porušení práv nebo bezpečnostní selhání.

03

Porovnání hodnot napříč případy

Využijte projektové, technické a klinické případy k prozkoumání kontextu, zástupných veličin, dedukce, bezpečnosti a kompromisů.



Projektový případ: AI a nábor zaměstnanců

- Případ náboru v rámci projektu „Ethical Engineer“ zkoumá, jak může automatizace ovlivnit přístup k práci, reprodukovat historické vzorce a přesouvat rozhodovací pravomoc mezi uchazeči, náboráři a dodavateli.
- Technické zaměření: účel, důkazy, relevantní atributy, přístupnost, vysvětlení, pravomoc náboráře, výzva a monitorování.
- Otevřete projektový případ náboru. Určete jedno hodnotové tvrzení organizace a důkazy potřebné k jeho ověření.

Projektová případová studie: hodnocení historických fotografií

- Případová studie Ethical Engineer zkoumá hodnocení historických fotografií s podporou umělé inteligence. Kulturní význam, původ a nejistota mohou být sníženy, pokud model zachází s obrázkem jako s izolovanými daty.
- Technické zaměření: kontext zdroje, zastoupené komunity, citlivé závěry, odborná interpretace, nejistota, přístup a opravy.
- Otevřete případ historických fotografií. Definujte, které zainteresované strany by měly ovlivňovat zamýšlené použití a které závěry by měly být zakázány nebo omezeny.

Projektová případová studie: umělá inteligence v analýze dat

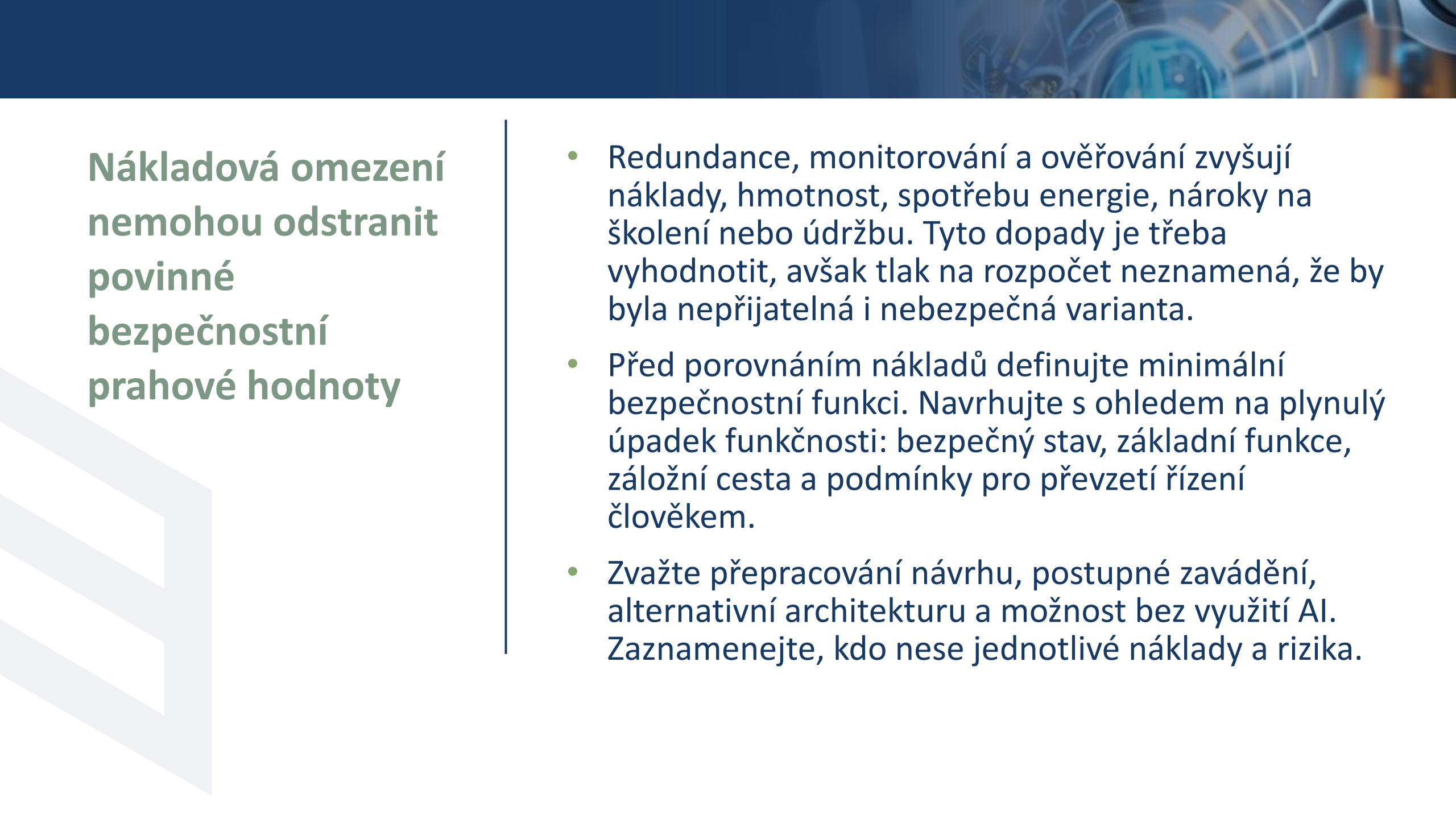
- Případová studie „Ethical Engineer“ v oblasti analýzy dat zkoumá, jak může technicky platná analýza nabýt nových důsledků, když jsou data kombinována, odvozována, vizualizována nebo znovu použita v rámci jiného rozhodnutí.
- Technické zaměření: účel, původ, zástupné proměnné, agregace, interpretace, přístup, sekundární využití a otázka, kdo může zpochybnit závěr.
- Otevřete projektovou případovou studii o analýze dat. Identifikujte napětí mezi hodnotami, které vzniká v důsledku jedné volby dat nebo vizualizace, a navrhnete ověřitelné konstrukční řešení.

Projektová případová studie: Umělá inteligence v pulzních srdečních pumpách

- Případová studie „Ethical Engineer“ týkající se srdečních pump se zabývá standardizací, přístupem, personalizovanou péčí a bezpečností. Návrh, který zjednodušuje výrobu a řízení, nemusí zohlednit fyziologické rozdíly.
- Technické zaměření: cílová populace, hodnoty pacientů, bezpečné mechanické a regulační limity, pravomoc klinického lékaře, alarmy, důkazy a odpovědnost za úpravy.
- Otevřete případ pulzní srdeční pumpy. Porovnejte pevný, konfigurovatelný a adaptivní návrh s využitím povinných bezpečnostních prahových hodnot.

Stejná metoda odhaluje různé hodnotové rozpory

- Při náboru se klade důraz na spravedlivý přístup, důstojnost, vysvětlení a výzvy. Historické fotografie přidávají kulturní kontext, původ a uctívou interpretaci. Analýza dat zdůrazňuje účel, dedukci, volbu zástupných proměnných a sekundární využití. Srdeční pumpy začleňují bezpečnost, přístup, personalizaci a klinickou autoritu do mechanického systému.
- Pro každý případ definujte technické rozhodnutí, zainteresované strany, kontextové hodnoty, mezeru v důkazech, povinnou prahovou hodnotu a proveditelné možnosti.
- Porovnejte, které metody jsou přenositelné a které požadavky závisí na dané oblasti.



Nákladová omezení nemohou odstranit povinné bezpečnostní prahové hodnoty

- Redundance, monitorování a ověřování zvyšují náklady, hmotnost, spotřebu energie, nároky na školení nebo údržbu. Tyto dopady je třeba vyhodnotit, avšak tlak na rozpočet neznamena, že by byla nepřijatelná i nebezpečná varianta.
- Před porovnáním nákladů definujte minimální bezpečnostní funkci. Navrhujte s ohledem na plynulý úpadek funkčnosti: bezpečný stav, základní funkce, záložní cesta a podmínky pro převzetí řízení člověkem.
- Zvažte přepracování návrhu, postupné zavádění, alternativní architekturu a možnost bez využití AI. Zaznamenejte, kdo nese jednotlivé náklady a rizika.

Případy selhání odhalují hodnoty zakotvené v předchozích rozhodnutích

Společnost Uber ukazuje, jak mohou harmonogram, řízení testování a předpoklady o lidském monitorování ovlivnit bezpečnost. Boeing 737 MAX ukazuje důsledky závislosti na senzorech, pravomoci v oblasti automatizace, informací a organizačního tlaku. Klinický případ astmatu ukazuje, jak se model může naučit spíše postupy péče než fyziologii.

Zjistěte, v jakém předchozím rozhodnutí o návrhu mohly lepší důkazy, námitky zainteresovaných stran nebo povinná prahová hodnota změnit výsledek.

Nepoužívejte postup jako náhradu za přepracování, pokud nebezpečí spočívá v architektuře. Analýza selhání by měla vést k revidovaným požadavkům a plánu ověřování.

Případ klinické AI: astma a zkreslení postupu péče

Zdálo se, že nemocniční model rizik považoval astma za ochranný faktor pro pacienty s pneumonií. V kontextu školení byli pacienti s astmatem často dříve zařazeni na jednotku intenzivní péče. Model se naučil postup péče a jeho výsledky, nikoli bezpečné fyziologické pravidlo.

Tento vzorec by se mohl stát nebezpečným při použití v prostředí s odlišnými postupy při přijímání pacientů. Inženýři by měli zkoumat kauzální předpoklady, procesy péče, přenositelnost a změnu domény.

Revize požadavku: klinická doporučení s významným dopadem vyžadují před použitím přezkoumání dané oblasti, analýzu postupu péče a validaci v zamýšleném prostředí.

Případ klinické AI: náklady na zdravotní péči použité jako náhradní ukazatel

Algoritmus pro populační zdraví používal předpovězené náklady na zdravotní péči jako náhradní ukazatel zdravotních potřeb. Nerovný přístup a výdaje znamenaly, že náklady neodrážely nemoc rovnoměrně napříč skupinami. Lidé s podobnými zdravotními potřebami mohli obdržet odlišná skóre rizika a mít odlišný přístup k podpoře. Provedte audit cíle, náhradního ukazatele, štítků a procesu přidělování. Porovnejte výsledky podle relevantních skupin a zeptejte se dotčených zainteresovaných stran, zda cíl odráží zamýšlenou hodnotu. Přepracujte cíl pomocí přímých měřítek potřeb, otestujte efekt přidělování a zaznamenejte zbývající nejistotu.

Technické zásahy mění několik hodnot najednou

- Metoda SMOTE nebo vážení tříd může zlepšit pozornost věnovanou vzácným třídám, ale zároveň ovlivnit kalibraci. Přenos učení může snížit objem dat a výpočetní náročnost, ale zároveň importovat nevhodné reprezentace. Augmentace může zlepšit robustnost nebo vytvořit nerealistické případy. PCA může snížit složitost, ale zároveň skrýt vlastnosti relevantní pro určitou podskupinu. LIME může pomoci s lokálním vysvětlením, ale zůstává pouze přibližným.
- Porovnejte každý zásah s výchozí situací. Otestujte zamýšlený přínos, vedlejší účinky, výkon v podskupinách, nejistotu a provozní limity.
- Nástroj je jednou z možností návrhu, nikoli důkazem, že daná hodnota byla dosažena.

Aplikovaný úkol:

Proved'te strukturované zkoumání hodnot

1. Vyberte si nábor zaměstnanců, historické fotografie, analýzu dat nebo srdeční pumpy.
2. Definujte technické rozhodnutí a alternativu bez využití umělé inteligence.
3. Zmapujte přímé a nepřímé zainteresované strany, mocenské vztahy a možné vyloučení.
4. Definujte dvě relevantní hodnoty v daném kontextu a identifikujte mezi nimi napětí.
5. Uved'te koncepční, empirické a technické důkazy.
6. Vytvořte dvě proveditelné návrhové varianty a povinné prahové hodnoty.
7. Vyberte variantu, požadavek, akceptační test, vlastníka a spouštěč revize.

Výstup: jednostránková analýza hodnot a dvouminutová obhajoba návrhu.

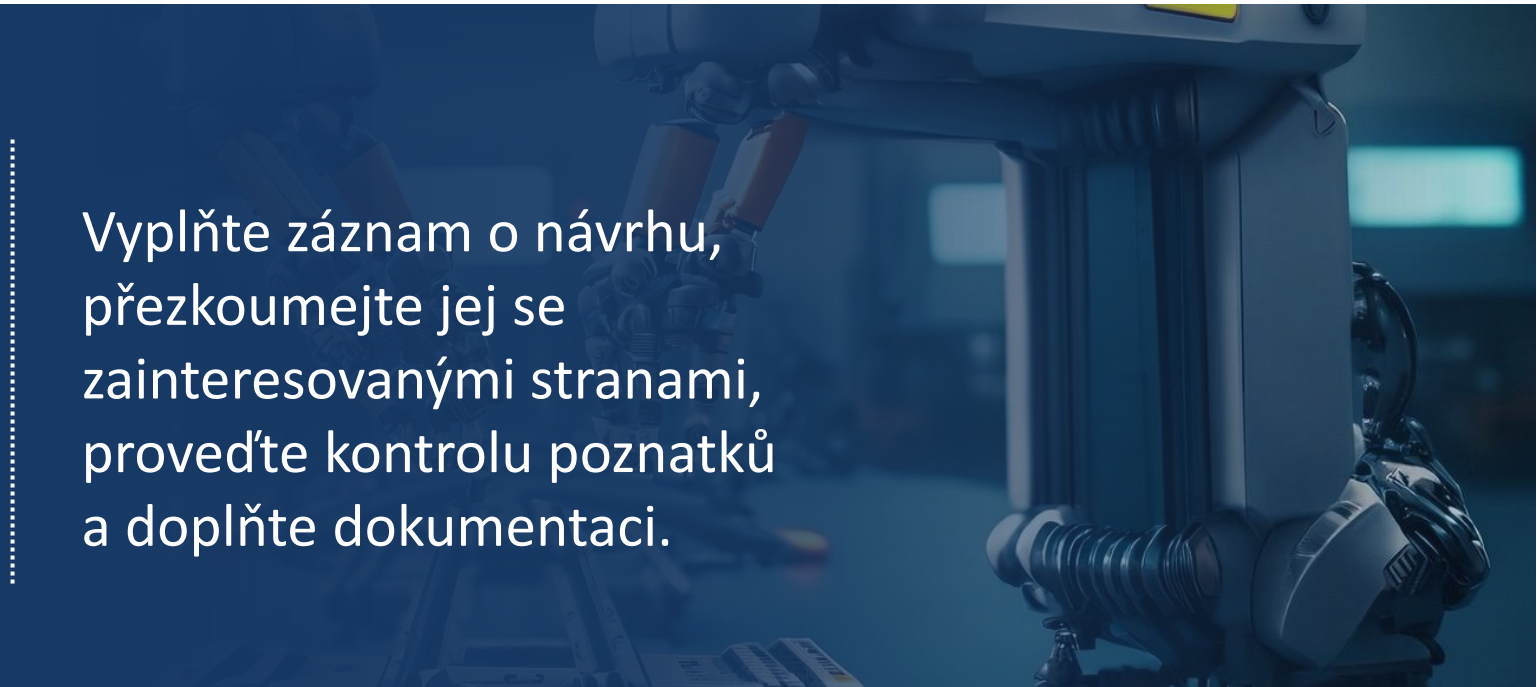
Dilema: skupiny zainteresovaných stran se neshodnou na přijatelném návrhu

- Systém na pracovišti podporovaný umělou inteligencí může snížit fyzickou zátěž. Vedení si cení údajů o produktivitě pro investiční rozhodnutí. Zaměstnanci podporují bezpečnostní výstrahy a odmítají individuální monitorování. Dodavatel tvrdí, že systém nemůže tyto dvě funkce oddělit bez dodatečných nákladů.
- Rozhodněte se pro GO, REVISE nebo STOP. Určete, komu chybí důkazy, které hodnoty jsou v rozporu a který požadavek je závazný. Porovnejte přepracovaný návrh, ohraničený pilotní projekt a alternativu bez AI.
- Zaznamenejte nejsilnější odlišný názor a důkazy, které by mohly změnit rozhodnutí.

04

Obhajte rozhodnutí založené na hodnotách

Vyplňte záznam o návrhu,
přezkoumejte jej se
zainteresovanými stranami,
proved'te kontrolu poznatků
a doplňte dokumentaci.



Případ A: Průmyslový exoskeleton podporovaný umělou inteligencí

- Továrna navrhuje exoskeleton vybavený senzory, který má snížit počet úrazů, zlepšit přizpůsobení a plánovat údržbu. Dodavatel také nabízí individuální hodnocení produktivity na základě pohybu, zátěže a doby nečinnosti. Zaměstnanci mají obavy z dopadu na zdraví a z disciplinárního využití.
- Porovnejte tři možnosti: cloudovou osobní optimalizaci s analýzou produktivity; bezpečnost a přizpůsobení přímo na zařízení s pseudonymními událostmi údržby; nebo neadaptivní mechanický systém.
- Před hodnocením proveditelných možností definujte povinné požadavky na bezpečnost, ochranu soukromí a zapojení pracovníků.

Dokončený záznam o návrhu založeném na hodnotách: průmyslový exoskelet

ROZHODNUTÍ REVIZE: zvolte podporu bezpečnosti a přizpůsobení přímo na zařízení; vyloučte individuální hodnocení produktivity.

DŮKAZY ZŮČASTNĚNÝCH STRAN: Zaměstnanci, zástupci pro bezpečnost, údržbáři a vedoucí pracovníci upřednostňují správné nasazení, bezpečné zatížení, jasná varování a absenci disciplinárních postihů.

PRAHOVÉ HODNOTY: Splnění mechanických bezpečnostních limitů; žádné využití pro účely produktivity nebo docházky; zástupci zaměstnanců schvalují rozsah dat; místní ruční ovládání zůstává k dispozici.

POŽADAVKY A TESTY: Zpracovat kloubové úhly a zatížení lokálně; odstranit surové pohybové vyrovnávací paměti; pseudonymizovat události údržby; otestovat mezní hodnoty zatížení, načasování výstrah, záložní řešení, použitelnost a obsah zpráv.

KONTROLA: Bezpečnostní inženýr odpovídá za mezní zatížení; vedoucí pro data ověřuje zakázané použití; zástupce zaměstnanců kontroluje zprávy. Znovu otevřít po změně funkce, dat, pracoviště nebo dodavatele.

Případ B: řízení poptávky po dálkovém vytápění

- Obec plánuje řízení dodávek dálkového vytápění pomocí umělé inteligence s využitím údajů o počasí, senzorů v budovách a prognóz poptávky. Tento návrh může snížit spotřebu paliva, zatímco konzervativní řízení může zajistit komfort v domech se špatnou tepelnou izolací. Nájemníci si nemohou vybrat senzory v budově a někteří nemají možnost podávat stížnosti přímo.
- Porovnejte energetickou účinnost, tepelný komfort, cenovou dostupnost, ochranu soukromí, odolnost a veřejnou odpovědnost. Zahrňte alternativu založenou na pravidlech a řízenou obsluhou.
- Před nasazením definujte minimální teplotní a obnovovací prahy, omezení účelu senzorů, možnost ručního zásahu obsluhy, veřejné reportování a postup pro nápravu.

Týmová aktivita:

Laboratoř pro rozhodování o návrhu založeném na hodnotách

1. Vyberte exoskeleton, dálkové vytápění nebo případovou studii. Určete systémového inženýra, vedoucího zainteresovaných stran, vedoucího pro důkazy, zástupce provozu, kritického recenzenta a zapisovatele.
2. Definujte rozhodnutí, hranice a alternativu bez využití umělé inteligence.
3. Proved'te koncepční, empirické a technické analýzy.
4. Přeložte dvě hodnoty na požadavky a akceptační testy.
5. Stanovte povinné prahové hodnoty a poté porovnejte dvě nebo tři proveditelné možnosti.
6. Rozhodněte se pro GO, REVISE nebo STOP a zaznamenejte bezpečnostní opatření, vlastníky, zbývající napětí a podmínky přezkumu.
7. Prezентуйте dvouminutovou obhajobu a po připomínkách zainteresovaných stran proveďte revizi.

Silná analýza založená na hodnotách mění technický návrh

- Silná analýza definuje rozhodnutí, hranice systému a alternativu bez AI. Mapuje přímé i nepřímé zainteresované strany, mocenské vztahy a vyloučení.
- Hodnoty jsou definovány v kontextu. Důkazy zainteresovaných stran mění požadavek, variantu, test nebo podmínku přezkumu.
- Každý požadavek má kritérium přijatelnosti, vlastníka důkazů a umístění v životním cyklu. Povinné prahové hodnoty jsou odděleny od vážených kritérií.
- Konečný záznam zachycuje nejistotu, nesouhlas, zbytkové napětí, monitorování, nápravná opatření a spouštěče přezkumu.

Pracovní list: Záznam o návrhu založeném na hodnotách

System, zamýšlené použití a alternativa bez AI:

Přímé a nepřímé zainteresované strany, moc a vyloučení:

Kontextové definice hodnot a napětí:

Koncepční důkazy a předpoklady:

Empirické důkazy, účastníci a omezení:

Technické možnosti a dopady na hodnoty:

[Zdroje]

Společnost Ethical Engineer schválila žádost KA220-HED, WP3 Výuka a učení.

https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf. <https://standards.ieee.org/standard/7000-2021.html>

Pokračování na dalším snímku



Povinné prahové hodnoty:

Požadavky, akceptační testy a vlastníci:

Rozhodnutí: POKRAČOVAT PŘEPRAVIT ZASTAVIT

Ochranná opatření, přetrvávající napětí a nesouhlas:

Monitorování, nápravná opatření a spouštěče přezkumu:

[Zdroje]

Etický inženýr schválil žádost KA220-HED, WP3 Výuka a učení.

https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf. <https://standards.ieee.org/standard/7000-2021.html>

Role play: přezkum návrhu pod tlakem nákladů a harmonogramu

- Dodavatel exoskeletu může dodržet termín dodávky pouze v případě, že zůstane aktivována cloudová analytika. Vedení tvrdí, že pozdější zavedení zásad může zabránit zneužití. Zástupci zaměstnanců požadují technické oddělení před zahájením jakéhokoli pilotního projektu. Bezpečnostní tým potřebuje funkci adaptivního přizpůsobení pro úkoly s vysokým rizikem.
- Role: technický manažer, bezpečnostní inženýr, zástupce zaměstnanců, vedoucí datového týmu, dodavatel a nezávislý recenzent.
- Rozhodněte, které hodnoty jsou povinné, zda jsou zásady dostačující, která varianta je proveditelná a kdo může odložit uvedení do provozu. Zaznamenejte jeden nepřijatelný argument a jeden platný důvod pro omezený pilotní projekt.

Individuální reflexe a přenos

- Napište krátký technický argument.
- Který zainteresovaný subjekt nebo hodnota zpočátku chyběla?
- Který důkaz změnil definici problému nebo návrh?
- Který požadavek umožnil hodnotu otestovat?
- Která prahová hodnota zabránila nepřijatelnému kompromisu?
- Kdo nese zbývající zátěž nebo riziko?
- Jaká změna by vedla k přehodnocení rozhodnutí?

Přeneste tuto metodu do jiné technické disciplíny a uveďte jednu změnu týkající se zainteresovaných stran, důkazů nebo ověření.

Kvíz

1. Jak se liší přímý zainteresovaný subjekt od nepřímého zainteresovaného subjektu?
2. Jaké jsou tři typy analýz v rámci Value Sensitive Design?
3. Proč by měly být hodnoty definovány v kontextu?
4. Co spojuje hodnotu s inženýrským akceptačním testem?
5. Proč se před váženým bodováním uplatňují povinné prahové hodnoty?
6. Uveďte jedno riziko technického zásahu, jako je SMOTE, PCA nebo post hoc vysvětlení.
7. Jaké důkazy ukazují, že účast zainteresovaných stran ovlivnila návrh?

[Zdroje]

- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>
- Původní sada materiálů k modulu 5 poskytnutá konsorciem projektu Ethical Engineer.

Odpovědi na kvíz a pedagogické zdůvodnění

1. Přímé zainteresované strany se systémem přímo interagují nebo od něj přijímají rozhodnutí; nepřímé zainteresované strany jsou ovlivněny širšími důsledky.
2. Koncepční, empirické a technické výzkumy.
3. Bezpečnost, autonomie nebo spravedlnost mohou mít pro různé lidi, postupy a systémy odlišný význam.
4. Sledovatelná hierarchie hodnot: hodnota, norma, požadavek, kritérium přijatelnosti a důkaz.
5. Vážený součet nesmí vyvážit závažné selhání v oblasti práva, bezpečnosti nebo práv.
6. Mezi platné odpovědi patří změna kalibrace, ztráta informací, importovaná zaujatost, nestabilita nebo zavádějící vysvětlení.
7. Zaznamenaná změna problému, požadavku, možnosti, testu nebo rozhodnutí spojená s důkazy zainteresovaných stran.

Kontrola výstupů učení

Před dokončením si ověřte, zda dokážete:

1. Zmapujte přímé a nepřímé zainteresované strany, moc a vyloučení.
2. Provádějte koncepční, empirická a technická šetření.
3. Převést hodnoty na požadavky a kritéria přijatelnosti.
4. Porovnejte možnosti dat, modelů, rozhraní a systémů s využitím prahových hodnot a důkazů.
5. Vytvořte návrh založený na hodnotách, včetně zbytkových napětí a revize.

Důkazy:

Mapa zainteresovaných stran, hierarchie hodnot, srovnání variant a návrhová dokumentace.

[Zdroje]

- Ethical Engineer schválil žádost KA220-HED, WP3 Výuka a učení.
- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>

Příspěvek do dokumentace kurzu

Přidejte tyto výstupy z modulu 5 do svého etického inženýrského spisu:

1. Mapa zainteresovaných stran, moci a vyloučení.
2. Kontextové definice hodnot a napětí mezi hodnotami.
3. Koncepční, empirické a technické důkazy.
4. Hierarchie hodnot s požadavky a akceptačními testy.
5. Důkazy o zásazích do dat, modelů a vysvětlení.
6. Povinné prahové hodnoty a srovnání možností.
7. Rozhodnutí o návrhu založená na hodnotách, ochranná opatření a spouštěče přezkumu.

Propojte tyto důkazy s prací v oblasti spravedlnosti, ochrany soukromí a odpovědnosti z modulů 2 až 4.

[Zdroje]

- Aplikace KA220-HED schválená organizací Ethical Engineer, WP3 Výuka a učení.
- https://nissenbaum.tech.cornell.edu/papers/Value_Sensitive_Design.pdf
- <https://standards.ieee.org/standard/7000-2021.html>



Oznámení o 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 5



Co-funded by
the European Union