exe
innovating
your world

Príručka o etike algoritmov

Úvod: Zodpovedný prístup k počítačovým algoritmom

Počítačové algoritmy pomáhajú ľudstvu

Počítače priniesli v posledných desaťročiach zázračné riešenia na množstvo veľkých problémov ľudstva. Zvýšili produktivitu práce, umožnili vzdialenú komunikáciu, zjednodušili prístup k poznatkom, priniesli zábavu. Ich prísľub zlepšovať naše životy je naďalej reálny a nadmieru lákavý.

Jadro digitálnych technológií tvoria algoritmy. Z časti sú to klasické počítačové programy, v ktorých riadok po riadku programátori počítačom hovoria, čo majú robiť. Používajú na to počítačové jazyky, ktoré sa z časti podobajú na tie ľudské a z časti skôr pripomínajú matematické vzorce.

Dnes však pribúdajú aj algoritmy, ktoré nenapísal človek – programátor, ale ktoré vytvoril sám počítač. Dokáže to napríklad tak, že mu poskytneme veľké množstvo údajov, z ktorých dokáže požadované riešenie odpozorovať a potrebný algoritmus si sám vytvoriť. Nazývame to strojové učenie, ktoré je podmnožinou oboru umelej inteligencie.

Na riešenie rôznych úloh je dnes bežné používať kombináciu počítačových algoritmov vytvorených človekom aj strojom. Všetky tieto digitálne riešenia, nech už boli vytvorené akokoľvek, voláme v tejto príručke svorne – algoritmy.

Algoritmy prinášajú aj etické otázky

Algoritmy, podobne ako iné technológie, môžu mať aj nežiadúce a negatívne dopady. Je čas si vážne klásť otázku, čo z toho, čo je vďaka digitálnym technológiám technicky možné, je aj morálne správne alebo dobré.

Schopnosť počítačových algoritmov dopĺňať, ovplyvňovať či nahrádzať ľudské chápanie a rozhodovanie prináša so sebou aj množstvo nových spoločenských otázok. Medzi takéto výzvy, s ktorými bude potrebné sa vysporiadať, patria napríklad riziká neúmyselnej diskriminácie v dôsledku neférových rozhodnutí algoritmov, či napríklad informovanosť ľudí o tom, že sa na citlivých rozhodnutiach o nich samotných podieľajú počítačové algoritmy.

Keď používame algoritmy na rozhodnutia, ktoré majú dopady na ľudí, ako zabezpečíme aby bol tento proces férový, vysvetliteľný a transparentný? Ako rozpoznať, kedy má algoritmus pri rozhodovaní nežiaduce predsudky? Kedy môžeme počítaču zveriť proces bez toho, aby sme ho kontrolovali, a kedy je dôležité ponechať aj dozor človeka? Ako vieme využiť výhod umelej inteligencie a zároveň si nenarušiť súkromie?

Ak chceme predísť rôznym nežiadúcim dôsledkom, ktoré môžu ohroziť nielen reputáciu firiem využívajúcich algoritmy umelej inteligencie, ale aj naštrbiť dôveru spoločnosti voči týmto technológiám, bude nevyhnutné, aby sme si na tieto otázky vedeli odpovedať s dostatočným predstihom.

Malo by nám ísť o to využívať potenciál plynúci z nasadenia algoritmov zodpovedne a bezpečne. Inak nám hrozí napríklad aj to, že ľudia nebudú inteligentným algoritmom plne dôverovať, nebudú ich ochotní toľko používať či rozvíjať.

Tieto otázky sa musíme naviac naučiť zodpovedať my sami, tu na Slovensku. Množstvo algoritmov k nám prichádza zo zahraničia a môžu byť vytrénované na vzorkách, ktoré u nás nemusia fungovať hodnoverne. Ak bude napríklad výrobná linka v továrni sledovať tváre pracovníkov pri stroji a upozorňovať na ich únavu, algoritmus trénovaný povedzme v Ázii môže byť u nás nepresný kvôli odlišným črtám tvárí na ktorých bol trénovaný, čo by mohlo viesť k nehodám. Iné dovezené AI technológie u nás môžu po technickej stránke fungovať bezchybne, ale nebudú kompatibilné s našimi hodnotami.

Systematické a odborné posudzovanie algoritmov pomocou tejto príručky

Táto príručka predstavuje praktický postup na systematické posúdenie etických rizík spojených s nasadzovaním algoritmov a ponúka návody na ich zmiernenie.

Jej jadro tvorí metodika vytvorená výskumníkmi z Johns Hopkins University a mestským zastupiteľstvom v San Franciscu. Jej originál sa nachádza na stránke Ethicstoolkit.ai a do nášho jazyka ju preložila a takto voľne verejnosti sprístupnila slovenská firma **exe**.

Táto príručka je určená pre každého, kto navrhuje, vyvíja, kupuje, či nasadzuje algoritmy. Je určená pre použitie algoritmov v každom odvetví priemyslu, služieb, či štátnej správy.

Táto príručka je skôr proces. V prvej časti vás prevedie sériou otázok, ktorými je možné posúdiť spôsob, závažnosť, či škálu dopadu algoritmov na našu spoločnosť a tak identifikovať možné etické riziká. V druhej časti ponúkne príručka odporučenia na zmierenie daných etických rizík.

Je v našom vlastnom záujme, aby sa technológie ako inteligentné algoritmy používali s dôverou občianskej spoločnosti. Preto potrebujeme neustále sledovať a vyhodnocovať spoločenské dopady algoritmov, a zároveň sa snažiť , aby tieto technológie boli nasadzované etickým spôsobom tak, aby z nich mal prospech čo najširší okruh ľudí. Len tak totiž žiadaný pokrok bude môcť získať podporu širokej verejnosti ako aj verejných inštitúcií.

Veríme, že tomu pomôže aj táto príručka. Budeme vďační, ak nám ju pomôžete šíriť. Prosím neváhajte sa nám ozvať s prípadnými otázkami či spätnou väzbou.

v Bratislave 13.11.2019

Miroslav Pikus, miroslav.pikus@exe.sk

Juraj Podroužek, juraj.podrouzek@pygmalios.com

http://www.e-tika.sk/

http://www.exe.sk/

1. časť: Posúdenie rizík súvisiacich s algoritmami

Prehľad

Správa algoritmov si vyžaduje porozumenie rizikám, ktoré z nich vyplývajú pre všetky zúčastnené strany. Pri používaní algoritmov je potrebné vyhodnotiť štyri hlavné skupiny rizík:

1. **Vplyv**. Skúma sa, ako algoritmus ovplyvní ľudí a majetok.
2. **Vhodné používanie**. Kontroluje sa vzťah medzi údajmi používanými v algoritme a účelom, na ktorý sa zbierali, ako aj vnímaním ich očakávaného použitia.
3. **Zodpovednosť**. Zisťuje sa, do akej miery sú ľudia zapojení do prebiehajúceho používania algoritmu, ako aj to, či sa dajú automatizované rozhodnutia každému jasne vysvetliť.
4. **Skreslenie a predpojatosť**. Skúma sa skrytý vplyv dát a ľudí, ktorí sa podieľali na vytvorení algoritmu.

V tejto časti príručky všetky uvedené riziká vyhodnotíte prostredníctvom série krokov. V každom kroku sa skúma jeden faktor. Tieto faktory sa potom zhrnú, čo zainteresovaným stranám pomôže do značnej miery pochopiť riziko. Úrovne rizika (a v niektorých prípadoch jednotlivé faktory) vám umožnia vypracovať plán na zmiernenie rizík (2. časť príručky), čiže ako budete na riadení týchto rizík spolupracovať s komunitou zainteresovaných strán.

Krok č. 1: Pochopenie a posúdenie vplyvu

Vplyv je spojením štyroch aspektov:

1. **Typ.** Používa sa na klasifikáciu vplyvu a definovanie jeho podstaty. Algoritmus určený na zisťovanie anomálií v genetickom kóde by napríklad patril do kategórie „život / bezpečnosť“.
2. **Stupeň.** Určuje úroveň vplyvu od zanedbateľného po závažný. Rozhodnutie o kaucii za uväzneného človeka by bolo napríklad považované za závažné.
3. **Miera.** Určuje, koľko ľudí, miest alebo vecí bude ovplyvnených.
4. **Smer.** Určuje, či je vplyv pozitívny alebo negatívny. Väčšina algoritmov bude mať pozitívny aj negatívny vplyv. Algoritmus určený na prideľovanie sociálnych bytov osobám bez domova by mal napríklad pozitívny vplyv na osoby, ktorým bolo bývanie pridelené, ale negatívny na osoby, ktoré boli vylúčené.

Pomocou krokov uvedených nižšie:

- identifikujete osoby alebo veci, ktoré budú ovplyvnené;
- analyzujete vplyv v rámci spomínaných štyroch aspektov: typu, stupňa, miery a smeru.

Postupne zistíte, že tieto kroky sa opakujú. Pri skúmaní miery vplyvu si napríklad možno uvedomíte, že ste zabudli na skupinu potenciálneho vplyvu. Preto vám predkladáme hárok, pomocou ktorého môžete túto analýzu vykonávať opakovane.

Krok č. 1.1: Opis vplyvu

Krok č. 1.1.1: Identifikovanie osôb alebo vecí, ktoré budú ovplyvnené

Pri identifikácii ovplyvnených osôb alebo vecí je užitočné zamyslieť sa nad bezprostrednosťou vplyvu:

- **Primárny.** Ide o priame ciele algoritmu, teda o ľudí, miesta alebo veci, ktorých sa algoritmus týka.
- **Sekundárny.** Ide o ľudí, miesta alebo veci, ktoré môžu pociťovať výsledky algoritmu v dôsledku toho, že ovplyvní primárne objekty, na ktoré bol zameraný.
- **Neočakávaný, resp. nezamýšľaný.** Ide o ľudí, miesta alebo veci, ktoré môžu pociťovať nezamýšľaný alebo neočakávaný vplyv algoritmu. Aj keď ich možno nepoznáte, môžete sa nad nimi zamyslieť.

V nasledujúcej tabuľke sa uvádza niekoľko príkladov primárnych, sekundárnych a neočakávaných, resp. nezamýšľaných objektov vplyvu. (Poznámka: Týmto položkám pravdepodobne budete musieť priradiť určitú úroveň dôležitosti.)

Primárny	**Sekundárny**	**Neočakávaný, resp. nezamýšľaný**
Jednotlivci	Rodina	Štvrť, škola, komunita, priatelia
Firma	Zákazníci	Štvrť, podobné firmy
Zemepisná oblasť	Obyvatelia, firmy	Realitné kancelárie, školy, návštevníci
Vybavenie	Prevádzkovatelia	Oblasti alebo miesta využívajúce vybavenie
Skupiny ľudí (napr. umelci)	Možnosti rekreácie	Kvalita života obyvateľov, hodnota majetku

Krok č. 1.1.2: Identifikácia typov vplyvu

Váš algoritmus bude mať minimálne jednu alebo viacero oblastí vplyvu. V nasledujúcej tabuľke sa opisujú rôzne typy vplyvu. Jeden typ vplyvu môže priniesť ďalší vplyv. Recenzie reštaurácie napríklad ovplyvňujú jej reputáciu, čo následne ovplyvňuje jej finančnú situáciu. Cieľom tohto kroku je zabezpečiť pochopenie podstaty vplyvov, nie ich stupňa či smeru.

Budete musieť určiť typ vplyvu pre každú skupinu určenú v kroku č. 1.1.1.

Typ	**Opis**
Prístup k tovaru, príspevkom alebo službám	Tieto typy algoritmov určujú, kto, čo alebo kde má alebo nemá prístup k tovaru, príspevkom alebo službám. Môže tam patriť prístup k poisteniu, štátnym príspevkom, možnostiam bývania, vzdelaniu, službám údržby alebo prevencie, rekreácii atď.
Finančný vplyv	Tieto typy algoritmov ovplyvňujú finančnú situáciu jednotlivcov, skupín, subjektov alebo oblastí.

Majetok alebo vybavenie	Tieto typy algoritmov ovplyvňujú kvalitu alebo hodnotu majetku alebo vybavenia.
Reputácia	Tieto typy algoritmov ovplyvňujú reputáciu jednotlivca, skupiny, subjektu alebo miesta.
Emocionálny vplyv	Tieto typy algoritmov ovplyvňujú emocionálny stav a pohodu jednotlivca alebo skupiny jednotlivcov.
Život/bezpečnosť	Tieto typy algoritmov ovplyvňujú život alebo bezpečnosť jednotlivca, skupiny, subjektu alebo miesta.
Súkromie	Tieto typy algoritmov ovplyvňujú súkromie jednotlivca alebo skupiny.
Sloboda	Tieto typy algoritmov ovplyvňujú slobodu jednotlivca, skupiny alebo subjektu.
Práva/duševné vlastníctvo	Tieto typy algoritmov ovplyvňujú práva alebo duševné vlastníctvo jednotlivca, skupiny alebo subjektu.

Krok č. 1.2: Posúdenie rozsahu vplyvu

Rozsah vplyvu je spojením stupňa a miery vplyvu.

Krok 1.2.1: Klasifikácia stupňa vplyvu

Po určení typu (typov) vplyvu algoritmu môžete klasifikovať ich relatívny vplyv. V nasledujúcej tabuľke sa opisujú úrovne vplyvu každého typu, a to od kategórie Zanedbateľný po kategóriu Závažný.

Z hľadiska smeru vplyvu považujte stupne vplyvu za neutrálne.

Typ	**Zanedbateľný**	**Malý**	**Stredný**	**Závažný**
Prístup k tovaru, príspevkom alebo službám	Bez rozdielneho prístupu k tovaru, príspevkom alebo službám	Malý rozdiel v prístupe k tovaru, príspevkom alebo službám	Stredný rozdiel v prístupe k tovaru, príspevkom alebo službám	Závažný rozdiel v prístupe k tovaru, príspevkom alebo službám
Finančný vplyv	Bez finančného vplyvu	Malý finančný vplyv	Stredný finančný vplyv	Závažný finančný vplyv
Majetok alebo vybavenie	Bez škody, zlepšenia alebo zmeny hodnoty	Malá škoda, zlepšenie alebo zmena hodnoty	Stredná škoda, zlepšenie alebo zmena hodnoty	Závažná škoda, zlepšenie alebo zmena hodnoty

Reputácia	Bez zmeny reputácie	Malá zmena reputácie	Stredná zmena reputácie	Závažná zmena reputácie
Emocionálny vplyv	Bez emocionálneho vplyvu	Malý emocionálny vplyv	Stredný emocionálny vplyv	Závažný emocionálny vplyv
Život/bezpečnosť	Bez vplyvu na život, fyzickú pohodu alebo bezpečnosť	Malý vplyv na život, fyzickú pohodu alebo bezpečnosť	Stredný vplyv na život, fyzickú pohodu alebo bezpečnosť	Závažný vplyv na život, fyzickú pohodu alebo bezpečnosť
Súkromie	Bez vplyvu na súkromie	Malý vplyv na súkromie	Stredný vplyv na súkromie	Závažný vplyv na súkromie
Sloboda	Bez zmeny slobody	Malá zmena slobody	Stredná zmena slobody	Závažná zmena slobody
Práva/duševné vlastníctvo	Bez zmeny vlastníctva alebo práv duševného vlastníctva	Malá zmena vlastníctva alebo práv duševného vlastníctva	Stredná zmena vlastníctva alebo práv duševného vlastníctva	Závažná zmena vlastníctva alebo práv duševného vlastníctva

Krok č. 1.2.2: Odhad miery vplyvu

Teraz môžete posúdiť mieru vplyvu. Je ovplyvnených iba niekoľko alebo viac ľudí, vecí či miest? Pomocou nasledujúcej tabuľky odhadnite mieru vplyvu pre každú oblasť vplyvu z kroku č. 1.1.1.

Miera	**Opis**
Malá	Tento algoritmus v našej jurisdikcii ovplyvňuje veľmi málo ľudí, miest alebo vecí.
Stredná	Tento algoritmus v našej jurisdikcii ovplyvňuje značné množstvo ľudí, miest alebo vecí.
Veľká	Tento algoritmus v našej jurisdikcii ovplyvňuje takmer každého človeka, miesto alebo vec, pričom môže mať vplyv aj mimo našej jurisdikcie.

Krok č. 1.2.3: Priradenie odhadu rozsahu

Pomocou stupňa a miery vplyvu priraďte odhad rozsahu.

Odhad rozsahu		Miera vplyvu		
		Malá	Stredná	Veľká
Stupeň vplyvu	Zanedbateľný	Veľmi úzky	Veľmi úzky	Obmedzený/úzky
	Malý	Veľmi úzky	Obmedzený/úzky	Podstatný
	Stredný	Obmedzený/úzky	Podstatný	Celoplošný/široký
	Závažný	Podstatný	Celoplošný/široký	Celoplošný/široký

Krok č. 1.3: Odhad celkového smeru vplyvu

Každý typ vplyvu môže byť bez ohľadu na svoju intenzitu pozitívny, negatívny alebo obojstranný.

Napríklad ak sa rozhoduje o tom, či niekto dostane nejaký benefit, pre daného jednotlivca je to dobré. Ak sa rozhoduje o nejakej oblasti, ktorá má byť sledovaná, dopad môže byť dobrý aj zlý. V každom prípade však existujú minimálne dve skupiny, ktoré sú ovplyvnené rôzne – tí, ktorí dostanú alebo nedostanú benefit, a tí, na ktorých je alebo nie je algoritmus zacielený. Váš algoritmus teda často ovplyvní dve alebo viacero skupín v dvoch rôznych smeroch.

Napriek tomu by ste celkový smer vplyvu mali posúdiť. Pomôže vám to v ďalších častiach, keď budete zvažovať kroky, ktoré by ste mali vykonať v záujme zodpovednejšieho a etickejšieho používania algoritmu.

- **Pozitívny.** Znamená celkový pozitívny vplyv, nemá za následok rozdielnosť v prístupe (napr. vynechanie z prístupu) ani negatívne zmeny či vplyvy a neuberá inej skupine alebo oblasti.
- **Prevažne pozitívny.** Znamená pozitívny vplyv na niektorých, ale neuberá inej skupine alebo oblasti. Niekto síce nebude môcť využívať výhody, ale nikto nebude poškodený.
- **Prevažne negatívny.** Znamená negatívny vplyv na niektorých a môže ubrať inej skupine alebo oblasti.
- **Negatívny.** Znamená alebo pripisuje prevažne negatívne vplyvy, odoberá iným skupinám či oblastiam, na ktoré sa vzťahuje, alebo ich odstraňuje.

Krok č. 1.4: Priradenie celkového rizika vplyvu

Skombinujte rozsah odhadu vplyvu z kroku č. 1.2.3 s celkovým odhadom smeru z kroku č. 1.3 a na základe toho získate odhad celkového rizika vplyvu vyplývajúceho z algoritmu.

Celkové riziko vplyvu		Celkový smer			
		Pozitívny	Prevažne pozitívny	Prevažne negatívny	Negatívny
Rozsah	Veľmi úzky	Veľmi malé	Veľmi malé	Malé	Stredné
	Obmedzený/úzky	Veľmi malé	Malé	Stredné	Závažné
	Podstatný	Malé	Stredné	Závažné	Vysoké
	Celoplošný/široký	Stredné	Závažné	Vysoké	Extrémne

Krok č. 2: Posúdenie rizika vhodného používania údajov

Vhodné používanie sa v tejto časti vzťahuje na to, či by ste mali údaje používať na účely daného algoritmu. Krok č. 2 sa zaoberá tým, či sú údaje dostatočne reprezentatívne a presné nato, aby ste ich mohli použiť.

Vstupné údaje je potrebné vyhodnotiť z hľadiska konzistentnosti a kompatibilnosti, ako aj reputácie a vnímania. To nám umožní porozumieť etickému riziku, ktoré prirodzene vyplýva z používania zdrojov informácií pre zamýšľaný algoritmus.

Krok č. 2.1: Klasifikácia konzistentnosti a kompatibilnosti používania

Na aký účel boli vstupné údaje pôvodne vytvorené, zozbierané alebo získané? Do akej miery je nové používanie kompatibilné s pôvodným dôvodom zberu údajov? Pomocou nasledujúcej tabuľky obodujte konzistentnosť a kompatibilnosť svojho zamýšľaného použitia.

Konzistentnosť a kompatibilnosť	**Opis**
Áno	Údaje pre tento algoritmus používame konzistentne a kompatibilne s účelmi a kontextom, v rámci ktorých sa získali. Používanie je v súlade aj s platnými zákonmi a predpismi.
Do určitej miery	Údaje pre tento algoritmus používame do určitej miery konzistentne a kompatibilne s účelmi a kontextom, v rámci ktorých sa získali.
Nedá sa určiť	Nepoznáme účel a kontext, v rámci ktorých sa tieto údaje získali. Môžeme údajom veriť, ak nevieme, ako boli zozbierané?
Nie	Údaje pre tento algoritmus nepoužívame konzistentne a kompatibilne s účelmi a kontextom, v rámci ktorých sa získali, alebo ich používame v rozpore s nimi.

Krok č. 2.2: Klasifikácia reputácie a vnímania používania

Aké riziká z hľadiska reputácie a vnímania vyplývajú z používania týchto údajov na účely daného algoritmu? Ako naň v prípade zverejnenia zareagujú ľudia? Pomocou nasledujúcej tabuľky klasifikujte očakávanú reakciu. Vo všeobecnosti platí, že riziko z hľadiska reputácie a vnímania je pri údajoch o jednotlivcoch väčšie.

Reputácia a vnímanie	**Opis**
Súhlasný postoj	Väčšina bude súhlasiť s tým, ako tieto údaje používame na zamýšľané účely algoritmu. Samozrejme, tak ako pri každom verejnom úsilí, niektorí s týmto používaním súhlasiť nebudú. Používanie údajov na tento konkrétny účel je obhájiteľné a nie je ojedinelé. Údaje sú verejne dostupné.
Zmiešaný postoj	Očakávame, že niektoré skupiny ľudí budú znepokojené tým, ako tieto údaje používame na zamýšľané účely algoritmu. Ide o bežný postup, ktorý nebol právne napadnutý. Obhájiteľné, hoci bez podobných prípadov.
Odmietavý postoj	Očakávame, že väčšina ľudí bude namietať proti tomu, ako tieto údaje používame na zamýšľané účely algoritmu. Pravdepodobne to bude obhájiteľné na dosiahnutie cieľov.

Krok č. 2.3: Priradenie skóre rizika vhodného používania

Pomocou predchádzajúcich dvoch krokov priraďte skóre rizika vhodného používania.

Skóre rizika vhodného používania		**Reputácia a vnímanie**		
		Súhlasný postoj	Zmiešaný postoj	Odmietavý postoj
Konzistentnosť a kompatibilnosť	Áno	Nízke	Nízke	Stredné
	Do určitej miery	Nízke	Stredné	Vysoké
	Nedá sa určiť	Stredné	Stredné	Vysoké
	Nie	Stredné	Vysoké	Vysoké

Krok č. 3: Posúdenie rizika nesenia zodpovednosti

Zodpovednosť za používanie algoritmov možno vyvodiť preskúmaním týchto otázok:

1. Kto alebo aký subjekt prijal ktoré rozhodnutia?
2. Ako boli tieto rozhodnutia prijaté?
3. Ako tieto rozhodnutia vysvetlíme? Dajú sa tieto rozhodnutia vysvetliť?
4. Ako môžeme tieto rozhodnutia preskúmať alebo preveriť?
5. Ako môžeme tieto rozhodnutia zmeniť, ak je s nimi vyjadrený nesúhlas?
6. Týkajúce sa konkrétne algoritmu:
    a. Ako sme algoritmus pred použitím otestovali?
    b. Ako zabezpečíme funkčnosť algoritmu podľa zamýšľaného účelu?
    c. Ako budeme merať výkonnosť algoritmu?
    d. Ako po čase algoritmus upravíme?

V nasledujúcich častiach sa zameriame na prvé štyri otázky a klasifikujeme riziko nesenia zodpovednosti za algoritmus. V 2. časti tejto príručky uvádzame odporúčané postupy, ktoré treba dodržať počas vývoja algoritmu, aby sa pokryla 5. a 6. otázka.

Krok č. 3.1: Určenie skóre automatizácie

Pomocou nasledujúcej tabuľky určite úroveň automatizácie pri rozhodovacom procese alebo kroku, ktorý algoritmus ovplyvňuje.

Skóre	**Opis**
Nízke – za prítomnosti človeka	Algoritmus sa používa na stanovenie jednotlivca alebo skupiny jednotlivcov. Konečné posúdenie vykonáva človek. Algoritmus neobsahuje dôrazné odporúčania ani nevypracúva závery (napr. politické rozhodnutia, rizikové faktory atď.).
Stredné – za použitia algoritmu	Algoritmus štruktúruje, obmedzuje alebo inak navrhuje odporúčané kroky alebo rozhodnutia. Kroky alebo rozhodnutie nakoniec uskutočňuje jednotlivec alebo skupina jednotlivcov (napr. vynesenie rozsudku, kaucia atď.).
Vysoké – rozhodnutie algoritmu	Algoritmus automaticky podniká kroky alebo vykonáva rozhodnutia bez zásahu človeka alebo skupiny (napr. kamery kontrolujúce prejazd na červenú, riadenie dopravného toku, priorizácia kontroly atď.).

Krok č. 3.2: Určenie skóre prístupnosti

Skóre prístupnosti je spojením toho, aké ľahké je:

- vysvetliť algoritmus,
- preveriť a preskúmať ho.

Krok č. 3.2.1: Určenie skóre vysvetliteľnosti

Pomocou nasledujúcej tabuľky klasifikujte, nakoľko jednoduché je vysvetliť algoritmus a jeho fungovanie. (Položte si otázku: Ako dobre ho dokážem vysvetliť laikovi?)

Vysvetliteľnosť	**Opis**
Ľahká	Algoritmus sa dá jednoducho vysvetliť a nevyžaduje si hlboké znalosti o štatistike a modelovacích technikách.
Stredná	Algoritmus sa dá vysvetliť, ale vyžaduje si hlbšie znalosti alebo dôkladnejšie vysvetlenie štatistických a modelovacích techník.
Ťažká	Algoritmus je náročný alebo je dokonca nemožné ho vysvetliť, a to i používateľom s hlbokými znalosťami (napr. čierna skrinka).

Krok č. 3.2.2: Určenie skóre preveriteľnosti

Pomocou nasledujúcej tabuľky opíšte, nakoľko jednoduché je preskúmať alebo preveriť funkčnosť algoritmu, ako aj vstupy a výstupy. Ako algoritmus dosiahne jednotlivé alebo konkrétne výsledky?

Preveriteľnosť	**Opis**
Ľahká	Preverenie a preskúmanie algoritmu môžeme uskutočniť podľa potreby a máme na to prostriedky.
Stredná	V prípade potreby vieme algoritmus preveriť a preskúmať. Musíme určiť, ako zmysluplné preverenie a preskúmanie vyzerá.
Ťažká	K algoritmu a spôsobu jeho fungovania nemáme prístup. Nemáme k dispozícii žiadne použiteľné prostriedky, pomocou ktorých by sme určili, ako ho preveriť a preskúmať.

Krok č. 3.2.3: Priradenie skóre prístupnosti

Na priradenie skóre prístupnosti použite skóre vysvetliteľnosti a preveriteľnosti.

Skóre prístupnosti		Vysvetliteľnosť		
		Ľahká	Stredná	Ťažká
Preveriteľnosť	Ľahká	Prístupný	Prístupný	S istými výhradami
	Stredná	Prístupný	S istými výhradami	So značnými výhradami
	Ťažká	S istými výhradami	So značnými výhradami	So značnými výhradami

Krok č. 3.3: Priradenie rizika nesenia zodpovednosti

V tomto kroku skombinujete skóre automatizácie a prístupnosti, aby ste mohli určiť úroveň rizika nesenia zodpovednosti, ktoré so sebou prináša používanie algoritmu.

Riziko nesenia zodpovednosti		Skóre automatizácie		
		Nízke – za prítomnosti človeka	Stredné – za použitia algoritmu	Vysoké – rozhodnutie algoritmu
Skóre prístupnosti	Prístupný	Nízke	Nízke	Stredné
	S istými výhradami	Nízke	Stredné	Vysoké
	So značnými výhradami	Stredné	Vysoké	Vysoké

Krok č. 4: Posúdenie rizika spojeného s metodikou tretej strany

Keď si algoritmus zaobstaráte alebo upravujete algoritmus vytvorený inde, musíte posúdiť potenciálne riziko pri jeho vytváraní a údržbe. Aj keď používanie algoritmov tretej strany čiastočne spadá do oblasti rizika nesenia zodpovednosti, vyžaduje si ešte ďalšiu analýzu.

Krok č. 4.1 Zodpovedanie otázok o metodike tretích strán

Nižšie sa uvádza zoznam otázok, ktoré vám pomôžu posúdiť algoritmus tretej strany. Každá otázka je relevantná pre niektorú z týchto fáz: návrh, monitorovanie alebo začlenenie. Na každú z uvedených otázok treba odpovedať áno alebo nie.

Fáza	Otázka	Bod v prípade odpovede áno
Návrh	Sme priamymi vlastníkmi algoritmu (bol vyvinutý interne, nie cez tretiu stranu).	1
Návrh	My (alebo tvorcovia) sme do návrhu algoritmu zapojili odborníkov na danú oblasť.	1
Návrh	Tvorcovia svoje predpoklady otvorene vysvetlili. Poznáme motiváciu vývojára alebo dodávateľa.	1
Návrh, monitorovanie	My (alebo tvorcovia) sme prediskutovali navrhované výstupy algoritmu s rôznymi subjektmi.	1
Návrh, monitorovanie	My (alebo tvorcovia) pravidelne kontrolujeme rozhodnutia prijímané algoritmom a upravujeme ho, aby spĺňal meniace sa potreby.	1
Návrh, monitorovanie	My (alebo tvorcovia) dokážeme algoritmus po zavedení nových premenných prestavať alebo znova vytrénovať od začiatku.	1
Monitorovanie	My (alebo tvorcovia) algoritmus pravidelne monitorujeme, aby sme sa uistili, že funguje, ako má.	1
Začlenenie	My (alebo tvorcovia) sme pred plným nasadením algoritmu a jeho využívaním na všetky rozhodnutia vyskúšali jeho pilotnú verziu alebo sme ho otestovali na podmnožine rozhodnutí z praxe.	1

Krok č. 4.2 Priradenie úrovne rizika spojeného s metodikou tretej strany

Započítajte si bod za každé tvrdenie, na ktoré ste odpovedali áno. Pomocou nasledujúcej tabuľky určite riziko spojené s metodikou tretej strany. (Poznámka: Ak sa žiadna z uvedených otázok o metodike tretej strany na vašu situáciu nevzťahuje, váš algoritmus je v zmysle tejto príručky vysoko rizikový).

Celkový počet bodov	Riziko spojené s treťou stranou
6 – 8	Nízke
3 – 5	Stredné
0 – 2	Vysoké

Krok č. 5: Posúdenie rizika historicky podmieneného skreslenia (bias)

Ak si vopred uvedomíte, že vo vašom algoritme bude prítomné skreslenie, pomôže to vám, ako aj tým, ktorí ho použijú. Vďaka tomu sa budete môcť sústrediť na minimalizáciu tohto skreslenia. Umožní vám to zamerať sa na zlepšenie súčasných postupov. Pozrite si dodatok alebo podrobné základné informácie o tejto dôležitej téme.

V tejto príručke rozlišujeme medzi:

- spoločensky podmieneným skreslením vyplývajúcim z historicky skreslených údajov (v dôsledku diskriminácie, historického odkazu, nespravodlivých politík atď.);
- technicky skreslenými údajmi, ktoré sú výsledkom neúmyselnej ľudskej chyby, problémov s kvalitou údajov alebo chýbajúcich údajov.

Algoritmus trénovaný na údajoch, ktoré sú nepresné v dôsledku ľudského pochybenia, má iný typ skreslenia ako algoritmus trénovaný na údajoch o bývaní z čias rasovej segregácie v USA. Skreslenie môže uškodiť v oboch prípadoch.

V tomto kroku určite úroveň rizika vyplývajúceho z historického zasadenia údajov použitých v algoritme. Zamyslite sa nad potenciálnym skreslením štruktúr použitých na zber údajov. Vezmite do úvahy trénovacie údaje (t. j. údaje použité pri pôvodnom trénovaní algoritmu) aj údaje vkladané do používaného algoritmu (t. j. súčasné a budúce údaje).

Riziko historicky podmieneného skreslenia	**Opis**
Nízke	Dôkladne sme preskúmali súvislosti. Údaje nie sú ovplyvnené žiadnymi zdokumentovanými ani známymi spoločenskými spormi alebo kontroverznými spoločenskými témami. Príklad: Algoritmus, na základe ktorého služba na streamovanie obsahu rozhoduje iba o možných preferenciách používateľa pri výbere filmu, pravdepodobne nebude historicky skreslený. Údaje sú nové (0- až 10-ročné).
Stredné	Do istej miery sme preskúmali súvislosti. Údaje mierne súvisia so zdokumentovanými alebo známymi spoločenskými spormi alebo kontroverznými spoločenskými témami. Príklad: Algoritmus vykonávajúci spracovanie v prirodzenom jazyku a trénovaný na starších údajoch z prieskumov o manželstve bude pravdepodobne historicky skreslený v otázke párov rovnakého pohlavia, keďže staršie stratégie zberu údajov boli ovplyvnené nespravodlivými, diskriminačnými postupmi, ktoré boli v minulosti zákonné. Údaje sú pomerne nové (11- až 25-ročné).

Vysoké	Nepreskúmali sme súvislosti. S údajmi sa spájajú negatívne historické udalosti. Údaje do veľkej miery súvisia so zdokumentovanými alebo známymi spoločenskými spormi alebo kontroverznými spoločenskými témami. Príklad: Algoritmus týkajúci sa miesta bývania trénovaný na niekoľko desiatok rokov starých údajoch o bývaní bude pravdepodobne historicky skreslený v otázke afroamerického obyvateľstva, pretože v starších údajoch sa odzrkadľuje jeho vylúčenie v dôsledku diskriminácie. Údaje sú staré (26- až viac ako 50-ročné).

Krok č. 6: Posúdenie rizika technicky podmieneného skreslenia

V tejto časti sa za technicky podmienené skreslenie považuje iba skreslenie týkajúce sa presnosti a reprezentatívnosti údajov (alebo ich nedostatku). Technicky podmienené skreslenie pri používaní algoritmov možno vyriešiť zodpovedaním týchto otázok:

1. Aká je kvalita údajov, ktoré sa majú použiť?
2. Do akej miery údaje presne odrážajú podmienky v praxi?
3. Bola počas vývoja metodika algoritmu dôkladne monitorovaná a kto toto monitorovanie vykonával?
4. Kto sa podieľal na vývoji a ako dokázal prispieť?
5. Odkiaľ pochádzajú trénovacie a ladiace údaje? Je tento zdroj vhodný pre kontext, v ktorom sa bude algoritmus používať?

Krok č. 6.1: Posúdenie rizika nereprezentatívnosti (nevyhovujúcej vzorky) a nepresnosti

K zdrojovým údajom patria trénovacie údaje (t. j. údaje použité pri pôvodnom trénovaní algoritmu) *aj* údaje vkladané do používaného algoritmu (napr. reálne, živé údaje). V oboch prípadoch musíte posúdiť skreslenie vzorky a kvalitu údajov.

Krok č. 6.1.1: Posúdenie rizika nereprezentatívnosti

V tomto kroku určite úroveň rizika nereprezentatívnosti údajov použitých v algoritme. Reprezentujú údaje vzorky vašu populáciu?

Riziko nereprezentatívnosti	**Opis**
Nízke	Údaje sú progresívne, pretože reprezentujú populáciu ako celok bez ohľadu na podskupiny.
Stredné	Údaje nadreprezentujú alebo podreprezentujú niektoré podskupiny a uvedomujeme si, kto, čo alebo kde je nad- alebo podreprezentované. Možno používame premenné, ktoré priamo nemerajú to, čo chceme zistiť (zástupné hodnoty).

Vysoké	Údaje nie sú reprezentatívne (príklad: údaje 311 sú skreslené v prípade tých, ktorí volajú na nenúdzovú linku 311). Použitie údajov môže priniesť tautologické výsledky, t. j. výsledky sa samy naplnia alebo sa môžu použiť iba na štúdium konkrétnej podskupiny. Používame premenné, ktoré nie sú vhodnými zástupnými hodnotami toho, čo sa snažíme merať. Výsledky by sa nemali použiť na dedukciu ani aplikovať na širšiu populáciu.

Hľadáte konkrétnejšie nástroje, ktoré by vám pomohli zamyslieť sa nad potenciálnym skreslením algoritmu? Neviete, ako vyhodnotiť skreslenie alebo nepresnosť? Tieto vám možno pomôžu:

- Dodatok A: Otázky týkajúce sa údajov (vypožičané od organizácie Center for Democracy and Technology)
- Reprezentatívna analýza (program Data Science for Social Good, Chicagska univerzita)
- Rámec na testovanie presnosti údajov
- Analýza údajov od centra GovEx
- Kvalita údajov od centra GovEx
- Náprava škôd spôsobených skreslením množiny údajov (MIT)

Krok č. 6.1.2: Posúdenie rizika nepresnosti

V tomto kroku určite úroveň rizika spojeného s kvalitou údajov použitých v algoritme. Zamyslite sa nad tým, ako sa údaje zbierali alebo získali, a identifikujte možné zdroje chýb vyplývajúcich z tréningu, overovania, nekonzistentnosti údajov, neštandardných metód zberu údajov atď.

Riziko spojené s kvalitou	**Opis**
Nízke	Údaje sú vysoko štruktúrované, zbierali sa na základe dôkladného overovania, trénovania a konzistentnosti. Zber údajov je automatický, vysoko štruktúrovaný a dá sa jednoducho overiť.
Stredné	Časť zberu údajov prebieha automaticky a časť zberu sa uskutočňuje manuálnym vstupom alebo iným ľudským vstupom. Overovanie je náročné alebo sa uskutočňuje, ale môžu sa vyskytnúť chyby.
Vysoké	Údaje nie sú dobre štruktúrované, overovanie sa neuskutočňuje, chýba trénovanie alebo sú metódy zberu údajov nekonzistentné.

Krok č. 6.1.3: Priradenie skóre rizika nereprezentatívnosti a nepresnosti

Skombinujte skóre rizika každej podkategórie a získate celkové skóre rizika skreslenia aj nepresnosti.

Skóre rizika nereprezentatívnosti a nepresnosti		Riziko nereprezentatívnosti		
		Nízke	Stredné	Vysoké
Riziko nepresnosti	Nízke	Nízke	Nízke	Stredné
	Stredné	Nízke	Stredné	Vysoké
	Vysoké	Stredné	Vysoké	Vysoké

Krok č. 6.2: Priradenie rizika spojeného s rozsahom trénovacích údajov

Skresliť výsledky môže aj rozsah údajov použitý na trénovanie algoritmu. Ak chcete algoritmus implementovať na Slovensku, nemá zmysel trénovať ho na medzinárodných údajoch. Ak chcete algoritmus implementovať na celonárodnej úrovni, ktorá ovplyvňuje všetkých Slovákov, nemá zmysel trénovať ho na množine údajov z Banskobystrického kraja.

Krok č. 6.2.1: Určenie skutočného zdroja trénovacích údajov

Pomocou nasledujúcej tabuľky opíšte, či sú vaše trénovacie alebo ladiace údaje lokálne alebo nelokálne.

Skutočný zdroj	Opis
Lokálny	Na trénovanie a ladenie algoritmu môžeme použiť vlastné lokálne údaje.
Nelokálny	Požičali sme si údaje niekoho iného, pracujeme v spolupráci s dodávateľom, ktorý zozbieral údaje z veľkej vzorky populácie, používame celoštátne údaje alebo používame údaje z mesta, kraja či štátu, do ktorého nepatríme.

Krok č. 6.2.2: Určenie požadovaného zdroja trénovacích údajov

Je dôležité, či sú trénovacie alebo ladiace údaje lokálne alebo nelokálne? Presnejšie povedané, pre niektoré algoritmy sú lepšie údaje z presne vymedzenej oblasti alebo konkrétne trénovacie údaje, zatiaľ čo pre iné budú vhodnejšie rozmanitejšie údaje alebo údaje so širším záberom. Príklad: Pre videostreamy s rozpoznávaním obrazu, ktoré monitorujú dopravu, bude vhodnejší zdroj údajov s oveľa širším záberom (v dôsledku vnútroštátnych noriem na označovanie ciest sa totiž určité udalosti môžu vyskytnúť len na národnej úrovni atď.), no napríklad situácie týkajúce sa konkrétnych regiónov alebo demografickej skupiny si môžu vyžadovať údaje z konkrétnej oblasti.

Požadovaný zdroj	Opis
Lokálny	Je veľmi dôležité algoritmus vytrénovať a vyladiť na lokálnych súvislostiach, pretože existujú jedinečné podmienky, ktoré by sa nemali dedukovať zo širšieho pohľadu.
Nelokálny	Nie je dôležité algoritmus vytrénovať a vyladiť na lokálnych údajoch alebo bude prínosné, ak sa použijú údaje zo širšieho kontextu.

Krok č. 6.2.3: Priradenie skóre rizika spojeného s trénovaním

Na základe odpovedí z predchádzajúcich krokov určite riziko použitia iných trénovacích údajov, než aké sú k dispozícii vo vašej jurisdikcii.

Riziko spojené s trénovaním		**Požadovaný droj**	
		Lokálny	Nelokálny
Skutočný zdroj	Lokálny	Nízke	Vysoké
	Nelokálny	Vysoké	Nízke

Krok č. 6.3: Priradenie rizika spojeného s metodikou

Skombinujte svoje skóre rizika z krokov č. 4.2 a 6.2.3, čím získate skóre rizika spojeného s metodikou.

Riziko spojené s metodikou		**Skóre rizika spojeného s trénovaním**	
		Nízke	Vysoké
Riziko spojené s metodikou tretej strany (krok č. 4.2)	Nízke	Nízke	Stredné
	Stredné	Nízke	Vysoké
	Vysoké	Stredné	Vysoké

Krok č. 6.4: Priradenie celkového rizika technicky podmieneného skreslenia

Skombinujte skóre rizika z krokov 6.1.3 a 6.3, na základe čoho priradíte celkové riziko technicky podmieneného skreslenia

Celkové riziko technicky podmieneného skreslenia		**Riziko spojené s metodikou**		
		Nízke	Stredné	Vysoké
Skóre rizika nereprezentatívnosti a nepresnosti	Nízke	Nízke	Nízke	Stredné
	Stredné	Nízke	Stredné	Vysoké
	Vysoké	Stredné	Vysoké	Vysoké

2. časť: Riadenie rizík spojených s algoritmami

V prvej časti príručky ste posudzovali rôzne rizikové faktory pre súbor algoritmov, ktorý chcete implementovať. Na základe výsledkov, ktoré ste získali, v tejto časti identifikujete vhodné mechanizmy na zmiernenie niektorých rizík.

Upozornenie:

- Jednotlivé zmierňovacie mechanizmy môžu byť užitočné v prípade viacerých rizík.
- Pre niektoré riziká alebo úrovne neexistujú konkrétne zmierňovacie mechanizmy.
- Niektoré podprvky rizika sú zahrnuté, iné nie.
- Každá úroveň rizika vychádza z predchádzajúcich zmierňovacích mechanizmov. Ak teda máte napríklad **vysoký** rizikový faktor, mali by ste uplatniť aj zmierňovacie mechanizmy pre **nízky** a **stredný** rizikový faktor.

Pokyny: Vyberte a implementujte zmierňovací mechanizmus, ktorý zodpovedá jednotlivým vybratým úrovniam rizika (úrovne vybraté v jednotlivých krokoch nájdete v dokumente *Príručka o etike a algoritmoch: 1. časť*). Zmierňovacie mechanizmy, ktoré platia pre váš scenár, si môžete podčiarknuť, zakrúžkovať alebo označiť. Táto časť príručky je rozdelená na dve časti: **spárovanie rizika so zmierňovacím mechanizmom** (rýchly prehľad stratégií) a **podrobnosti o zmierňovacích mechanizmoch** (podrobné vysvetlenia daných stratégií).

Spárovanie rizika so zmierňovacím mechanizmom

Pozrite sa na krok č. 1.3 o odhade rozsahu:

Ak ste vybrali možnosť **Veľmi úzky** alebo **Obmedzený/úzky**, zapojte dotknuté komunity (zmierňovací mechanizmus č. 1).

Ak ste vybrali možnosť **Podstatný**, použite verejné monitorovanie výkonnosti (zmierňovací mechanizmus č. 2).

Ak ste vybrali možnosť **Celoplošný/široký**, vytvorte radu pre inštitucionálne skúmanie[1] (zmierňovací mechanizmus č. 3) alebo inú verejnú poradnú skupinu s rozhodovacou právomocou v otázkach programu (zmierňovací mechanizmus č. 4).

Pozrite sa na krok č. 1.4 o klasifikácii celkového rizika vplyvu:

Ak ste vybrali možnosť **Veľmi malé**, **Malé** alebo **Mierne**, zapojte dotknuté komunity (zmierňovací mechanizmus č. 1).

[1] Ďalšie informácie sa nachádzajú v definícii rady pre inštitucionálne preskúmanie vypracovanej Oregonskou štátnou univerzitou a v pokynoch na registráciu rady pre inštitucionálne preskúmanie vydaných Ministerstvom zdravotníctva a sociálnych služieb USA.

Ak ste vybrali možnosť Závažné, použite verejné monitorovanie výkonnosti (zmierňovací mechanizmus č. 2).

Ak ste vybrali možnosť Vysoké alebo Extrémne , vytvorte radu pre inštitucionálne preskúmanie (zmierňovací mechanizmus č. 3) alebo inú verejnú poradnú skupinu s rozhodovacou právomocou v otázkach programu (zmierňovací mechanizmus č. 4).

Pozrite sa na krok č. 2.3 o vhodnom používaní údajov:

Ak ste vybrali možnosť Nízke alebo Stredné , iniciujte dialóg s verejnosťou o nových použitiach údajov, ktoré sa vzťahujú na algoritmy (zmierňovací mechanizmus č. 5).

Ak ste vybrali možnosť Vysoké , nájdite alebo vytvorte alternatívne zdroje údajov, ktorými nahradíte tie nevhodné (zmierňovací mechanizmus č. 6).

Pozrite sa na krok č. 3.3 o nesení zodpovednosti:

Ak ste vybrali možnosť Nízke alebo Stredné, pomocou automatizovaných testovacích nástrojov pravidelne vyhodnocujte výkonnosť algoritmu (zmierňovací mechanizmus č. 7), zabezpečte uplatnenie mechanizmu posúdenia človekom (zmierňovací mechanizmus č. 8) a požadujte zásah človeka pred vykonaním každého rozhodnutia algoritmu (zmierňovací mechanizmus č. 9).

Ak ste vybrali možnosť Vysoké, zabezpečte, aby sa pri ladení algoritmu použili výsledky mechanizmu posúdenia človekom (zmierňovací mechanizmus č. 8), zabezpečte, aby sa v prípade každého rozhodnutia vždy zaznamenali relevantné vstupy a stav počítača (zmierňovací mechanizmus č. 10), a pravidelne vyhodnocujte rozhodnutia prijímané na základe zásahu človeka (zmierňovací mechanizmus č. 11).

Pozrite sa na krok č. 4.2 o tretej strane:

Ak ste vybrali možnosť Nízke alebo Stredné, preneste riziko spojené so zodpovednosťou na dodávateľa (zmierňovací mechanizmus č. 12) a zaveďte nezávislé monitorovanie prostredníctvom interných zdrojov alebo inej tretej strany (zmierňovací mechanizmus č. 13).

Ak ste vybrali možnosť Vysoké, začleňte stimuly pre dodávateľov, aby ste dosiahli požadované výstupy (zmierňovací mechanizmus č. 14).

Pozrite sa na krok č. 5.1 o historicky podmienenom skreslení:

Ak ste vybrali možnosť Nízke alebo Stredné, vylaďte algoritmus tak, aby sa systematicky minimalizoval vplyv skreslenia alebo sa kompenzovali chýbajúce údaje (zmierňovací mechanizmus č. 15).

Ak ste vybrali možnosť **Vysoké**, údaje nepoužívajte(zmierňovací mechanizmus č. 6) a nájdite alternatívne zástupné hodnoty so správnym skreslením (zmierňovací mechanizmus č. 16).

Pozrite sa na krok č. 6.1.3 o skreslení a nepresnosti:

Ak ste vybrali možnosť **Vysoké**, spustite projekt na zlepšenie riadenia údajov (zmierňovací mechanizmus č. 16) alebo nájdite iný zdroj údajov (zmierňovací mechanizmus č. 6).

Pozrite sa na krok č. 6.2.3 o trénovacích údajoch:

Ak ste vybrali možnosť **Vysoké**, nájdite vhodnejší zdroj údajov (zmierňovací mechanizmus č. 6).

Pozrite sa na krok č. 6.3 o nevhodnej metodike:

Ak ste vybrali možnosť **Vysoké**, najmite si preverovateľov algoritmov, aby preskúmali vplyv faktorov, premenných alebo kovariantov (zmierňovací mechanizmus č. 17).

Pozrite sa na krok č. 6.4 o celkovom skreslení:

Ak ste vybrali možnosť **Nízke** alebo **Stredné**, jasne definujte prostriedky merania skreslenia a priebežne program monitorujte, aby sa zabezpečilo, že sa skreslenie nezvyšuje (zmierňovací mechanizmus č. 18).

Ak ste vybrali možnosť **Vysoké**, porovnajte existujúce skreslenie s predpokladaným (zmierňovací mechanizmus č. 19).

Pozrite sa na krok č. 7.0 o celkovom riziku:

Ak ste vybrali možnosť **Nízke**, zabezpečte, aby programoví manažéri pochopili a schválili rizikový profil

(zmierňovací mechanizmus č. 20).

Ak ste vybrali možnosť **Stredné**, zabezpečte, aby používatelia systému (a dotknutí jednotlivci) vedeli, že rozhodnutia sa prijímajú automaticky (zmierňovací mechanizmus č. 21), a na jedno rozhodnutie použite viaceré algoritmy, pričom uprednostnite rozhodnutia, ktoré vedú k požadovaným výstupom (zmierňovací mechanizmus č. 22).

Ak ste vybrali možnosť **Vysoké**, odložte implementáciu, kým sa riziká neznížia alebo kým prínosy podstatne neprevážia nad rizikami (zmierňovací mechanizmus č. 23), vytvorte radu pre inštitucionálne preskúmanie s rozhodovacou právomocou v otázkach programu (zmierňovací mechanizmus č. 3) a zabezpečte, že výskumníci budú implementáciu pravidelne vyhodnocovať (

zmierňovací mechanizmus č. 11) a predkladať správy rade pre inštitucionálne preskúmanie (zmierňovací mechanizmus č. 3).

Podrobnosti o zmierňovacích mechanizmoch

Zmierňovací mechanizmus č. 1. Efektívne zapojenie komunity je zamerané na ľudí, stojí na partnerstvách a zohľadňuje právomoci. Zapojenie komunity by malo byť spoločenské (pomocou existujúcich sociálnych sietí a vzťahov), technologické (zručnosti, nástroje a digitálne priestory), fyzické (občania) a založené na rovnosti (berie do úvahy právomoci a zodpovednosť za ne). Zapojenie dotknutých komunít v súvislosti s verejne dostupnými údajmi by mohlo vyzerať napríklad takto: verejnosť sa bude podieľať na vypracovaní pravidiel, budú sa priorizovať verejne dostupné údaje, verejnosť vytvorí inovatívne nástroje z nespracovaných údajov a potom bude pracovať s údajovými aplikáciami a vizualizačnými nástrojmi.

Zmierňovací mechanizmus č. 2. Účelom verejného monitorovania výkonnosti je identifikovať oblasti dobrej výkonnosti a oblasti, v ktorých možno výkonnosť zlepšiť. Informácie o výkonnosti by mali byť sústredené (na ciele a služby agentúry), vhodné (a užitočné pre zainteresované strany, ktoré ich budú pravdepodobne používať), vyvážené (poskytujúce predstavu o tom, čo agentúra robí, a týkajúce sa všetkých podstatných oblastí práce), stabilné (aby ich neovplyvnili organizačné zmeny alebo odchod jednotlivcov), začlenené (do organizácie) a efektívne z hľadiska nákladov (vyvažujúce prínos informácií s nákladmi na ich získanie).

Zmierňovací mechanizmus č. 3. Rada pre inštitucionálne preskúmanie je tradičný výbor, ktorého úlohou je skúmať a schvaľovať žiadosti týkajúce sa výskumných projektov. Rada pre inštitucionálne preskúmanie môže existovať aj v neakademických kruhoch a jej členovia môžu predstavovať potrebný krok pred implementáciou algoritmu.

Zmierňovací mechanizmus č. 4. Verejné poradné skupiny sa zvyčajne skladajú z kľúčových zainteresovaných strán projektu, ako aj zo zástupcov všeobecnej verejnosti, ktorí majú prispievať k vývoju projektu.

Zmierňovací mechanizmus č. 5. Dialóg s verejnosťou o nových použitiach údajov sa môže iniciovať jednoducho vypracovaním prieskumu a zapojením obyvateľov doň, zasielaním týždenných alebo dvojtýždenných správ alebo newsletterov, ktoré obyvateľov informujú o nových použitiach, organizovaním stretnutí na mestských úradoch s cieľom diskutovať o údajoch, zverejnením verejne dostupných údajov online alebo prevádzkovaním verejného Githubu.

Zmierňovací mechanizmus č. 6. Pokúste sa predísť vzniku kontroverzie tak, že nezačnete projekt s údajmi, ktoré môžu uškodiť. Nájdite alebo vytvorte nové zdroje údajov. Môžete napríklad zostaviť inventár údajov, v ktorom možno nájsť vhodnejšie údaje, preskúmať danú tému online a nájsť nové údaje tam alebo zozbierať nové údaje.

Zmierňovací mechanizmus č. 7. Systematické kontroly možno do životného cyklu algoritmu začleniť pomocou automatizácie testovacích nástrojov na vyhodnocovanie výkonnosti algoritmu (napr. konfúznych matíc pri vyhodnocovaní klasifikačných modelov). Ak uvedený klasifikačný model nesprávne klasifikuje 70 % prípadov, automatizovaný testovací nástroj môže byť naprogramovaný, aby červenými písmenami zobrazil hlásenie STOP.

Zmierňovací mechanizmus č. 8. Skvelým doplnkom projektu, ktorého súčasťou je algoritmus alebo algoritmy, môže byť mechanizmus posúdenia človekom, prostredníctvom ktorého môže človek uplatniť svoj úsudok. Ak sa informácie z tohto mechanizmu použijú pri vylaďovaní algoritmu, môže to byť pre projekt ozaj prínosné.

Zmierňovací mechanizmus č. 11. Rozhodnutia prijímané na základe zásahu človeka pravidelne vyhodnocujte, aby ste odhalili neželané skreslenie, ktorého sa môže posudzovateľ dopustiť.

Zmierňovací mechanizmus č. 13. Prenesenie monitorovania algoritmu na interný zdroj alebo inú tretiu stranu prináša do metodiky algoritmu ďalšiu úroveň subjektivity.

Zmierňovací mechanizmus č. 15. Chýbajúce údaje môžu byť príčinou štatistickej nepresnosti v každom projekte. V súvislosti s algoritmami môžu chýbajúce údaje výrazne zhoršiť už i tak škodlivé skreslenie. Ak viete, že váš algoritmus používa údaje, ktoré pozostávajú zväčša z chýbajúcich hodnôt, uistite sa, že dokáže tieto hodnoty systematicky zohľadňovať. Príklad: Ak máte malú množinu údajov, môžete sa rozhodnúť priradiť váhu niektorým demografickým údajom, aby bolo možné presnejšie odzrkadliť všeobecnú populáciu.

Zmierňovací mechanizmus č. 16. Realizácia projektu zlepšenia riadenia údajov môže pozostávať z: vytvorenia štruktúry riadenia údajov, vypracovania pravidiel pre verejne dostupné údaje, spustenia nového inventára údajov, vytvorenia portálu s verejne dostupnými údajmi, zaviazania sa k dodržiavaniu nového postupu zverejňovania údajov alebo noriem týkajúcich sa údajov, systematického testovania kvality údajov, dodržiavania nových pravidiel uchovávania údajov či zásad ochrany osobných údajov a zásad zabezpečenia, zapojenia komunity v oblasti údajov alebo náboru nových zamestnancov a talentov.

Zmierňovací mechanizmus č. 17. Veľmi užitočný môže byť nábor preverovateľov algoritmov (štatistikov, dátových analytikov, odborníkov na dátovú vedu, počítačových výskumníkov atď.), ktorých úlohou je preskúmať vplyv určitých faktorov, premenných alebo kovariantov. Možno vám pomôže, ak sa títo preverovatelia budú k vášmu algoritmu pravidelne vracať a vykonávať kontrolu systémov. K akému záveru dospeli po prísnom preverení vášho algoritmu?

Zmierňovací mechanizmus č. 18. Jednoznačnosť a stanovenie zámerov sú veľmi dôležité za každých okolností. V súvislosti s algoritmom jasne definujte prostriedky merania skreslenia a potom program priebežne monitorujte, aby sa zabezpečilo, že sa skreslenie nezvyšuje.

Zmierňovací mechanizmus č. 20. Zabezpečte, aby programoví manažéri pochopili rizikový profil, ktorý váš algoritmus odráža, a mohli ho odsúhlasiť. Dokážu vysvetliť každé riziko a chápu, koho alebo čo toto riziko alebo tieto riziká môžu ovplyvniť?