

2. modulis

Priekšstati, taisnīgums
un iekļaušana mākslīgā
intelekta sistēmās



Co-funded by
the European Union



Ethical Engineer

www.ethicalengineer.eu

SATURS

- 01 Diagnosticējiet sistēmas neobjektivitāti
- 02 Taisnīguma pārbaude kontekstā
- 03 Izstrādāt taisnīgus un iekļaujošus kontroles mehānismus
- 04 Inženierijas lēmuma pieņemšana un aizstāvēšana



Co-funded by
the European Union

Šis projekts ir finansēts ar Eiropas Komisijas atbalstu. Par šo publikāciju (paziņojumu) atbild vienīgi autors, un Komisija neuzņemas nekādu atbildību par tajā ietvertās informācijas izmantošanu. Saskaņā ar jauno Vispārīgās datu aizsardzības regulas (GDPR) regulējumu lūdzu ņemiet vērā, ka partnerība apstrādās jūsu personas datus vienīgi projekta interesēs un nolūkos, neierobežojot jūsu tiesības.



Īss pārskats par kursu „Ētiskais inženieris”

Katras moduļa galvenā satura un mācību rezultāti

1. modulis	02. modulis	03. modulis	04. modulis	05. modulis	Modulis 06
Ētika un pārredzamība	Priekšstati, taisnīgums un iekļaušana	Privātums un datu aizsardzība	Atbildība un drošība	Uz vērtībām balstīts dizains	Ilgtermiņa un ietekme
Ētikas principi un ES vērtības Pārredzamība un izskaidrojamība AI un robotikas gadījumu analīze REZULTĀTS: ētisko jautājumu karte	Aizspriedumi visā AI dzīves ciklā Taisnīguma mērķi un pierādījumi Iekļaujošas inženierijas gadījumi REZULTĀTS: aizspriedumu audits	Privātums jau projektēšanas posmā un GDPR Datu pārvaldība un PET Novērtēšana un apstrīdējamība REZULTĀTS: privātuma projektēšana	Lomas un atbildība Kļūdas, noturība un nodošana Reģistrēšana un mācīšanās no incidentiem REZULTĀTS: izsekojamības reģistrs	Ētiskās dizaina prasības Datu, modeļu un saskarnes izvēle Kompromisi un ātrās pārbaudes punkti REZULTĀTS: lēmumu matrica	Enerģija, ūdens un aparatūra Ietekme uz sabiedrību un dzīves ciklu Rādītāji, pārvaldība un pārveidošana REZULTĀTS: ilgtspējas plāns

NOSLĒGUMA DARBS: apvienot sešus rezultātus ētiskās inženierijas dokumentācijā, izmantojot vienu WP2 gadījumu vai inženierijas projektu

Moduļa ievads

Mērķis

Sagatavojiet inženierzinātņu studentus, lai viņi spētu identificēt, novērtēt un samazināt netaisnīgus rezultātus mākslīgā intelekta sistēmās, vienlaikus ieviešot iekļaujošas komandas un ieinteresēto pušu prakses.

Mērķauditorija

Augstskolu studenti inženierzinātnēs, datorzinātnēs, robotikā un saistītās disciplīnās.

Pamatojas uz

1. modulis: ētikas principi, pārredzamība un inženierzinātņu lēmumu dokumentēšana.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Moduļa ievads

Mācību secība

Diagnosticējiet neobjektivitāti, atlasiet pierādījumus, izstrādājiet kontroles pasākumus un aizstāviet inženierijas lēmumu.

Mācību rezultāti

Priekšstatu ietekmes ceļu karte, taisnīguma pārbaudes plāns, iekļaušanas pēc dizaina pārbaudes saraksts, priekšstatu ietekmes audita protokols un inženiertehniskā lēmuma kopsavilkums.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Mācību rezultāts

Līdz moduļa
beigām studenti
spēs:

1. Izskaidrot, kā problēmas formulēšana, dati, apzīmējumi, iezīmes, modeļi, atgriezeniskās saites un komandas pieņēmumi var radīt neobjektivitāti.
2. atšķirt taisnīguma mērķus un izvēlēties pierādījumus, kas piemēroti inženierijas kontekstam.
3. Novērtēt sniegumu un ietekmi attiecīgajās un savstarpēji saistītajās grupās.
4. Piemērot daudzveidības un iekļaušanas principus komandas sadarbībai, ieinteresēto pušu līdzdalībai un sistēmas izstrādei.
5. Ierosināt tehniskus, organizatoriskus, līdzdalības un cilvēku veikta pārskatīšanas kontroles pasākumus.
6. dokumentēt kompromisus, atlikušos riskus, uzraudzību un veidus, kā apstrīdēt lēmumu.

Priekšstatī klūst par inženierijas
jautājumu, kad tie maina rezultātus

- AI sistēmas klasificē, rangē, prognozē un sadala. Šīs funkcijas ietekmē piekļuvi darbam, veselības aprūpei, izglītībai, finansēm un sabiedriskajiem pakalpojumiem. Netaisnīgi rezultāti var rasties pat tad, ja modelis kopumā ir tehniski precīzs.
- Inženierijas jautājums ir konkrēts: kura sistēmas izvēle rada atšķirības attieksmē vai rezultātā, kam, kādos apstākļos un ar kādiem pierādījumiem?

Pamatota atbilde saista ietekmētos cilvēkus ar prasībām, datiem, veiktspējas testiem, cilvēku veiktu pārskatīšanu, korektīvajiem pasākumiem un uzraudzību.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Eiropas pamatnostādne s saista taisnīgumu ar inženierijas pierādījumiem

ES ētikas vadlīnijās daudzveidība, nediskriminācija un taisnīgums ir noteikti kā uzticamas mākslīgā intelekta prasības.

AI likums un ES Harta nodrošina visstingrākos juridiskos un tiesību pamatus: riska pārvaldību, datu pārvaldību, cilvēku uzraudzību, cieņu, nediskrimināciju un vienlīdzību.

Standarts ISO 30415 sniedz vadlīnijas iekļaujošas organizācijas kultūras un lēmumu pieņemšanas veicināšanai. DigComp atbalsta digitālās informācijas, datu un tehnoloģiju kritisku izvērtēšanu.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgos vizuālos zīmolus un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

01

Priekšstatu diagnosticēšana sistēmā

Izsekot, kā mērķi, dati, apzīmējumi, modeļi un atgriezeniskās saites var ietekmēt rezultātus.

Priekšstati ir sistēmas īpašība

Sistematiskais noslēpums ir kropļojums, kas rada maldinošus vai netaisnīgus rezultātus. Tas var rasties no nelīdzsvarotiem datiem, vēsturiskas nevienlīdzības, šaurām komandas perspektīvām, ieinteresēto pušu pārstāvības trūkuma vai darbības izvēlēm.

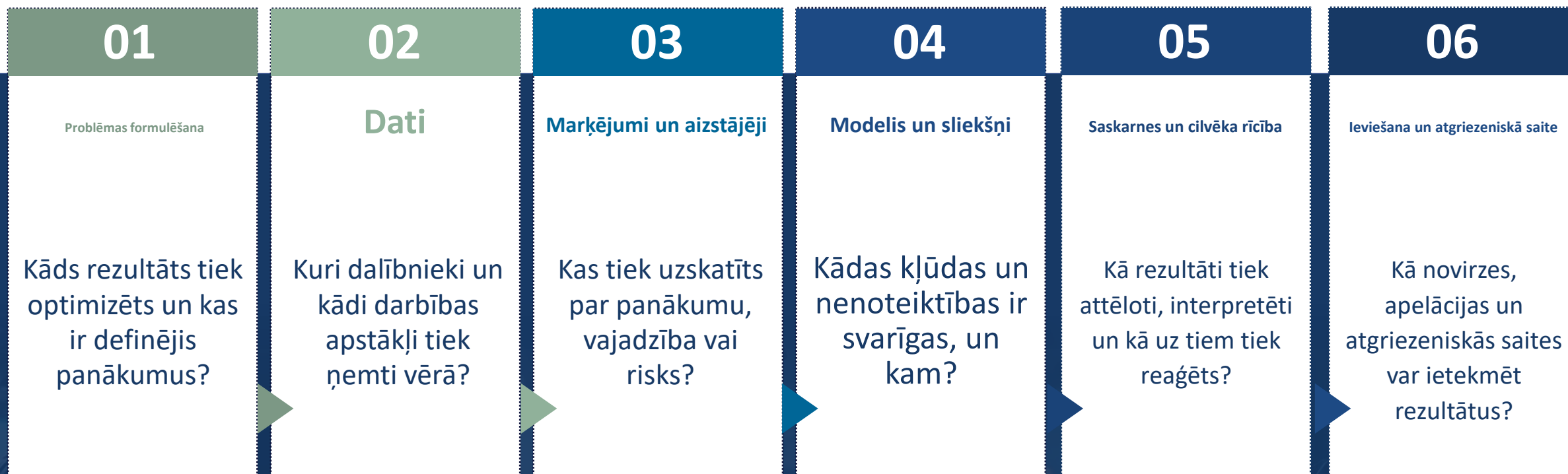
Dati var ietvert vēsturiskas nevienlīdzības, taču neobjektivitāte var ienākt arī caur mērķi, mērķa mainīgo, marķējuma noteikumiem, iezīmju dizainu, sliekšni, saskarni vai ieviešanas procesu.

[Piekluve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Priekšstati var parādīties jebkurā AI dzīves cikla posmā

Izvēlieties vienu inženierijas lēmumu. Katrā posmā identificējiet neobjektivitātes avotu, ietekmētās grupas un nepieciešamos pierādījumus.



CAPSTONE: Priekšstatu veidošanas ceļu karte, kas saista katru posmu ar pierādījumiem, ietekmētajām grupām, dažādības un iekļaušanas apsvērumiem, kā arī praktisku kontroles mehānismu.

Pabeigta neobjektivitātes analīze: pašbraucošo taksometru ieviešana

Pētījuma piemērs:
„Ethical Engineer”
pašbraucošie taksometri.

Lēmums:
Apstiprināt autonomo
taksometru izmantošanu
jauktā pilsētas rajonā. Katrā
posmā tiek reģistrēti
neobjektivitātes faktori,
ietekmētās personas,
pierādījumi un kontroles
pasākumi.

01

Problēmas formulēšana

- **Priekšstats:** Panākumus nosaka autoparka efektivitāte; retām briesmām tiek piešķirta mazāka nozīme.
- **Ietekmētie:** gājēji, velosipēdisti, pasažieri ar invaliditāti un ārkārtas dienestu darbinieki.
- **Pierādījumi/kontrole:** kopīgi definēt drošību un pieejamību; noteikt atbildību par negadījumiem.

02

Dati

- **Priekšstati:** Dati atspoguļo galvenokārt labvēlīgus laika apstākļus, kartē atzīmētus ceļus un tipiskus satiksmes modeļus.
- **Ietekmētie:** cilvēki ceļu remontdarbu zonās, apgabalos ar sliktu redzamību un slikti kartografētos rajonos.
- **Pierādījumi/kontrole:** pārbaudes aptvērums, izņēmuma gadījumu iekļaušana un darbības jomas ierobežošana.

Pabeigta neobjektivitātes analīze: pašbraucošo taksometru ieviešana

Pētījuma piemērs:
„Ethical Engineer”
pašbraucošie taksometri.

Lēmums:

Apstiprināt autonomo taksometru izmantošanu jauktā pilsētas rajonā. Katrā posmā tiek reģistrēti neobjektivitātes faktori, ietekmētās personas, pierādījumi un kontroles pasākumi.

03

Markējumi un aizstājēji

- **Priekšstati:** Gandrīz notikušie negadījumi un cilvēku iejaukšanās tiek marķēti nekonsekventi.
- **Ietekmētie:** ceļu satiksmes dalībnieki, kuru uzvedība vai pārvietošanās tiek nepietiekami uztverta.
- **Pierādījumi/kontrole:** vienoties par drošības marķējumiem; pārbaudīt anotācijas un aizstājējraudītājus.

04

Modelis un sliekšņi

- **Kļūda:** Atklāšanas un apstāšanās sliekšņi līdzsvaro komfortu un sadursmes risku.
- **Ietekmētie:** gājēji, velosipēdisti un pasažieri retos apstākļos.
- **Pierādījumi/kontrole:** salīdzināt scenāriju kļūdas; piemērot drošības rezerves un drošības robežas.

Pabeigta neobjektivitātes analīze: pašbraucošo taksometru ieviešana

Pētījuma piemērs:
„Ethical Engineer”
pašbraucošie taksometri.

Lēmums:

Apstiprināt autonomo taksometru izmantošanu jauktā pilsētas rajonā. Katrā posmā tiek reģistrēti neobjektivitātes faktori, ietekmētās personas, pierādījumi un kontroles pasākumi.

05

Saskarnes un cilvēka rīcība

- **Priekšstati:** Nav skaidrs, kā rīkoties pārlicības, vadības nodošanas un ārkārtas situācijās.
- **Ietekmētie:** pasažieri ar invaliditāti, attālinātie operatori un ārkārtas situāciju reaģētāji.
- **Pierādījumi/kontrole:** testējiet pieejamus brīdinājumus; nodrošiniet iejaukšanās un eskalācijas maršrutus.

06

Ieviešana un atsauksmes

- **Priekšstats:** Labi aptvertie maršruti generē vairāk datu un šķiet drošāki.
- **Ietekmētie:** nepietiekami pārstāvētie rajoni, to iedzīvotāji un pārvietotie autovadītāji.
- **Pierādījumi/kontrole:** uzraudzīt incidentus pa apgabaliem; pārbaudīt novirzes, sūdzības un pārejas ietekmi.

Pabeigta neobjektivitātes analīze: pašbraucošo taksometru ieviešana

Pētījuma piemērs:
„Ethical Engineer”
pašbraucošie taksometri.

Lēmums

Apstiprināt tikai validētā darbības jomā,
nodrošinot pieejamus vadības elementus,
attālinātu iejaukšanos, atbildību par
incidentiem un uzraudzību reģiona līmenī.

Pārtraukt darbību, ja tiek pārkāpti drošības vai
pārklājuma nosacījumi.

[Norādījumi pasniedzējam]

- Dodiet skolēniem divas minūtes, lai salīdzinātu šo aizpildīto karti ar savu karti iepriekšējā slaidā.
- Jautājiet, kuras kontroles izmaiņas ietekmē dizainu, kuras ierobežo ieviešanu un kuras nodrošina risinājumu pēc kaitējuma rašanās.
- uzsvērt, ka galīgais lēmums ir nosacīts. Pierādījumi nosaka, kur sistēma drīkst darboties un kad tai jāpārtrauc darbība.

[Piekļuve]

- Izlasiet sešus posmus secībā un nodrošiniet saiti uz gadījuma izpēti pieejamā digitālajā formātā.

Paraugu ņemšanas, mērīšanas un marķēšanas izvēles rada dažādus riskus

Paraugu atlases neobjektivitāte rodas, ja trūkst attiecīgo personu vai darbības apstākļu. Vēsturiskā neobjektivitāte rodas, ja pagātnes nevienlīdzības tiek uzskatītas par neitrāliem pierādījumiem.

Mērījumu novirze rodas, ja mainīgais lielums grupas reģistrē atšķirīgi. Etiķešu novirze rodas, ja mērķis atspoguļo nekonsekventu vērtējumu vai nevienlīdzīgu procesu.

[Piekluve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgos vizuālos zīmolus un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Aizstājējmai nīgie var slēpt vērtību izvēles

- Aizstājējmainīgais tiek izmantots, ja vēlamā jēdzienu nevar izmērīt tieši. Izmaksas var simbolizēt veselības vajadzības. Iepriekšējā nodarbinātība var simbolizēt nākotnes sniegumu. Aizturēšana var simbolizēt likumpārkāpumu.
- Aizstājējmainīgais var būt prognozējošs, vienlaikus radot netaisnīgus rezultātus, jo to ietekmē sociālie apstākļi. Inženieriem jājauc, ko aizstājējmainīgais nozīmē katrai grupai, kādi cēloņfaktori to veido un vai paredzētais lēmums ir pamatots.

Dokumentējiet aizstājējmainīgo, tā ierobežojumus un alternatīvo mērījumu.

[Piekluve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Atsauksmju cikli var pārvērst sākotnējo tendenci par sistēmas modeli

- AI rezultāti bieži maina turpmāk vāktos datus. Papildu kontrole rada vairāk reģistrētu incidentu. Vairāk pārbaūžu atklāj vairāk kļūdu. Noraidītie pieteikuma iesniedzēji nerada turpmākus snieguma datus.
- Atsauksmju cikls var radīt iespaidu, ka modelis kļūst arvien drošāks, vienlaikus sašaurinot to, ko sistēma var novērot.

Inženiertehniskie kontroles pasākumi ietver izlases paraugus, neatkarīgus datu avotus, kontrfaktuālu pārskatīšanu, atlases rādītāju uzraudzību un iespēju cilvēkiem iesniegt trūkstošos pierādījumus.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Iekļaušana uzlabo problēmas definīciju

[Piekluve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

- Iekļaušana nozīmē, ka skartie cilvēki var ietekmēt prasības, novērtēšanu un risinājumus. Pārstāvība komandā ir svarīga, savukārt līdzdalība lēmumu pieņemšanā ir spēcīgāks inženiertehniskais kontroles līdzeklis.
- Dažādības pārvaldība nozīmē plānoto veidu, kā organizācija atzīst cilvēku atšķirības un ļauj dažādām perspektīvām ietekmēt darbu un lēmumus. Sistēmiska dažādības pārvaldība padara to redzamu pārvaldībā: ar ko tiek apspriestas lietas, kurš var apšaubīt pieņēmumus, kurš var pieprasīt pierādījumus un kurš var apstrīdēt kaitīgu mērķi.
- Identificējiet tiešos lietotājus, datus pārstāvētus cilvēkus, uz kuriem attiecas lēmumi, operatorus, uzturētājus un netieši ietekmētās kopienas. Jautājiet, kādas pieredzes trūkst un kurš var apstrīdēt ierosināto mērķi.
- Psiholoģiski droša komanda ļauj savlaicīgi ziņot par neskaidrībām, domstarpībām un iespējamo kaitējumu.

Praktisks piemērs no mašīnbūves: iekļaujošs sirds sūkņa dizains

Pētījuma gadījums: Ethical Engineer pulsējošā sirds sūknis.

Sākotnējais uzdevums: Izvēlēties standarta VAD ar fiksētu sistolisko tilpumu steidzamai implantācijai. Iekļaušana maina to, kas definē problēmu un nepieciešamos pierādījumus.

01

Sākotnējais uzdevums

Apstipriniet, ka standarta ierīce ir piemērota un spēj sasniegt nominālo plūsmas mērķi klīniskajā termiņā.

02

Kuri dalībnieki netika iekļauti

Pacients, ķirurgs, perfūzists, specializētā medmāsa, mehānikas vai biomedicīnas inženieris, ētikas konsultants un iepirkumu vadītājs.

03

Ko atklāja līdzdalība

Anatomija atšķiras no tipiskā diapazona. Fiksēts sistoliskais tilpums var radīt nepietiekamu plūsmu vai spiedienu. Steidzamība, izmaksas un piekrišana ietekmē īstenojamību.

Praktisks mehāniskās inženierijas piemērs: iekļaujoša sirds sūkņa projektēšana

Pētījuma gadījums: „Ethical Engineer” pulsējošā sirds sūkņa.

Sākotnējais uzdevums:
Izvēlēties standarta VAD ar fiksētu sistolisko tilpumu steidzamai implantācijai. Iekļaušana maina to, kas definē problēmu un nepieciešamos pierādījumus

04

Pārformulēta problēma

Kura no pieejamajām konfigurācijām var nodrošināt drošu fizioloģisko atbalstu šim pacientam klīniskajā laika logā?

05

Inženierijas prasības

Izmantojiet pacienta attēlveidošanas datus; simulējiet piemērotību, plūsmu, spiedienu un deformāciju; ziņojiet par nenoteiktību; salīdziniet standarta, modulāras un pielāgotas iespējas.

06

Lēmums un pārskats

Izmantojiet standarta ierīci tikai tad, ja pacienta specifiskie dati liecina par drošu pielāgojumu. Reģistrējiet atlikušo risku, informētu piekrišanu, uzraudzības un pārtraukšanas kritērijus.

Praktisks mehāniskās inženierijas piemērs: iekļaujoša sirds sūkņa projektēšana

Pētījuma gadījums: „Ethical Engineer” pulsējošā sirds sūkņa.

Iekļaušana mainīja konstrukciju

Standarta ierīces izvēles uzdevums kļuva par pacientam specifisku lēmumu par drošību, pierādījumiem un piekrišanu.

[Norādījumi pasniedzējam]

- Lūdziet apmācāmajiem noteikt precīzo brīdi, kad iekļaušana maina inženierijas uzdevumu.
- Aiciniet vienu grupu aizstāvēt standarta ierīci, vienu – pielāgošanu un vienu – atlikšanu. Katrai grupai jānorāda pierādījumi un atlikušais risks.
- Uzsvērt, ka līdzdalība rada prasības un lēmuma pieņemšanas kritērijus. Tas ir inženierijas kontroles pasākums, nevis konsultācijas, kas pievienotas procesa beigās.

[Piekļuve]

- Nolasiet katru rindu skaļi un nodrošiniet saiti uz sirds sūkņa gadījumu pieejamā digitālajā formātā.

02

Taisnīguma pārbaude kontekstā

Izvēlieties taisnīguma mērķus, pasākumus, pierādījumus un pieņemamības kritērijus inženierijas kontekstā.

Taisnīgums saista sistēmas izvēli ar nevienlīdzīgām sekām

Taisnīgums attiecas uz to, kā AI sistēma sadala ieguvumus, riskus, kļūdas, kavējumus un iespējas starp ietekmētajiem cilvēkiem. Kopējais precizitātes rādītājs var slēpt nopietnas atšķirības starp grupām vai darbības apstākļiem.

Tādējādi taisnīgums ir gan tehnisks projektēšanas mērķis, gan cilvēktiesību saistības, kas saistītas ar cieņu, nediskrimināciju un vienlīdzību. Netaisnīgums var rasties mērķa, datu, marķējumu, aizstājēju, sliekšņu, saskarnes, cilvēka rīcības vai ieviešanas procesa dēļ. Inženierijas jautājums ir, vai sistēmas izvēle rada nepamatotu nelabvēlīgu situāciju.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgos vizuālos zīmolus un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Taisnīgums saista sistēmas izvēli ar nevienlīdzīgām sekām

01

Taisnīgums mērķis

Norādiet, kāds rezultāts, iespēja vai kļūda ir jāizdala taisnīgi, un paskaidrojiet, kāpēc tas ir svarīgi šajā lēmumā.

02

Pierādījumi

Pārbaudiet sniegumu un ietekmi dažādās attiecīgajās grupās un apstākļos. Norādiet grupu skaitu, nenoteiktību un datu ierobežojumus.

03

Inženierija Reakcija

Noteikt pieņemšanas kritērijus, tehniskos un organizatoriskos kontroles pasākumus, cilvēku veikto pārskatīšanu, uzraudzību un ceļu problēmu novēršanai.

Taisnīguma mērķi var būt pretrunā viens ar otru. Inženieriem ir jāpamato izvēlētais mērķis saistībā ar lēmumu, skartajām personām un kļūdu sekām. Par biežiem mērķiem var minēt vienlīdzīgu piekļuvi, vienlīdzīgas iespējas, salīdzināmus kļūdu rādītājus, kalibrēšanu, individuālo konsekvenci un taisnīgu procesu.

**Taisnīguma
rādītājs ir
pierādījumi, nevis
lēmums**

Ziņojiet par absolūtajiem rezultātiem, nenoteiktību un grupu skaitu. Izpētiet krustpunktu grupas, ja ir pietiekami daudz pierādījumu. Apvienojiet kvantitatīvos testus ar kvalitatīviem pierādījumiem par procesu, cieņu, iekļaušanu, pieejamību un iespēju vērsties tiesā.

**Sāciet ar lēmumu un
ietekmēto
iedzīvotāju grupu**

Novērtēšanas plānā jāatspoguļo reāli apstākļi, tostarp trūkstoši dati, sensoru kvalitāte, valoda, invaliditāte, vides svārstības un tas, vai prasību apkopošanā bija pārstāvētas ietekmētās ieinteresētās puses.

Sadaliet veiktspēju un parādiet nenoteiktību

- Kopējā precizitāte var slēpt vāju sniegumu mazākai grupai. Ziņojiet par sajaukuma matricas rādītājiem, kalibrēšanu vai uzdevumam specifisko kvalitāti atbilstoši attiecīgajai grupai un darbības apstākļiem.
- Norādiet parauga lielumu, uzticamības intervālu vai citu nenoteiktības aprakstu. Ja pierādījumi ir pārāk ierobežoti, reģistrējiet nepilnības un ierobežojiet ieviešanu, nevis izmantojiet nestabilus rezultātus kā taisnīguma pierādījumu.

Atkārtojiet testus pēc modeļa, datu vai sliekšņa izmaiņām, kā arī darbības laikā.

[Piekluve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Krustpunktu analīze maina testēšanas dizainu

- Cilvēki sistēmas uztver, ņemot vērā vairākas īpašības un apstākļus vienlaikus. Modelis var darboties atbilstoši sievietēm kopumā un kādai rasu grupai kopumā, bet vienlaikus darboties slikti sievietēm šajā grupā.
- Krustpunktu analīzei nepieciešami pietiekami dati un rūpīga privātuma aizsardzība. Apvienojiet mērķtiecīgus kvantitatīvos testus ar ieinteresēto pušu apspriešanu un scenāriju analīzi.

Reģistrējiet, kuras krustpunktu kombinācijas tika novērtētas, kuras joprojām ir neskaidras un kā uzraudzība atklās kaitējumu pēc ieviešanas.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgos vizuālos zīmolus un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Sliekšņi sadala kļūdas un ieguvumus

- Rezultāts nav noteikts, kamēr nav izvēlēts sliekšnis un rīcības noteikums. Sliekšņa pārvietošana maina viltus pozitīvos rezultātus, viltus negatīvos rezultātus, darba slodzi un piekļuvi.
- Salīdziniet iespējamās sliekšņa vērtības, izmantojot ietekmēto grupu rezultātus un katras kļūdas reālās sekas. Apsveriet pārskatīšanas posmu, kurā neskaidros gadījumus novērtē cilvēks.

Lēmuma protokolā jānorāda izvēlētais sliekšnis, pieņemšanas kritēriji, pārskatīšanas kārtība, atbildīgā persona un izmaiņu iemesls.

[Piekļuve]

Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Projekta gadījums: AI darbā pieņemšanas procesos

Inženierijas fokuss: vēsturiskie dati, mērķa definīcija, aizstājējraksturlielumi, apakšgrupu testēšana, paskaidrojums, apstrīdējamība un darba prasību validācija, ņemot vērā daudzveidību.

Analizējiet darbā pieņemšanas gadījumu kā sistēmu

Pārbaudiet, vai vēsturiskā darbā pieņemšana ir derīgs spēju rādītājs. Izpētiet izglītības, karjeras pārtraukuma un valodas pazīmes kā iespējamus aizstājējus. Salīdziniet atlases un kļūdu rādītājus starp attiecīgajām grupām. Pārbaudiet, vai darbā pieņemšanas speciālisti var atcelt rezultātu, vai dažādas ieinteresētās puses palīdzēja definēt darba prasības un vai kandidāti var apstrīdēt kļūdu.

Projekta gadījums: AI sejas atpazīšanā

Atveriet „Ethical Engineer” sejas atpazīšanas gadījuma izpēti un definējiet attiecīgo kļūdu izvēlētajam lietojumam.

- Šajā projekta gadījumā tiek izskāta sejas atpazīšana, ko izmanto, lai uzlabotu drošību un optimizētu procesus, piemēram, iekāpšanu lidostā. Ieguvumi ir atkarīgi no uzticamas identitātes saskaņošanas, savukārt privātums, datu drošība un demogrāfiskie rādītāji rada ievērojamus riskus.
- **Inženierijas aspekti:** lietošanas konteksts, attēla kvalitāte, „viens pret vienu” vai „viens pret daudziem” salīdzināšana, kļūdu sekas, demogrāfiskie testi un cilvēka veikta pārbaude.

NIST pierādījumi maina lēmumu par sejas atpazīšanu

- NIST pārbaudīja lielu skaitu algoritmu un konstatēja demogrāfiskas atšķirības lielākajā daļā novērtēto algoritmu, turklāt rezultāti bija atkarīgi no algoritma, lietojuma un datiem. Tas nozīmē, ka vispārīgs apgalvojums par sejas atpazīšanu ir vājš pierādījums.
- Norādiet algoritma versiju, darbības režīmu, attēla apstākļus un iedzīvotāju grupu. Pārbaudiet kļūdu saistībā ar nodarīto kaitējumu. Kļūdaina saskaņošana novērošanas sistēmā atšķiras no neveiksmīgas verifikācijas pie vērtiem.

***Pievienojiet cilvēka veiktu pārskatīšanu,
kvalitātes pārbaudes, pārsūdzības iespēju
un incidentu uzraudzību***

[Piekluve] Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo tēmu, lai palīdzētu lasītājiem piekļūt informācijai digitālajā formātā.

Praktiskais uzdevums: Izstrādājiet taisnīguma pārbaudes plānu

- Izvēlieties darbā pieņemšanas vai sejas atpazīšanas gadījumu. Izstrādājiet plānu, ko varētu īstenot cita inženieru komanda.
- Definējiet lēmumu, populāciju, grupas, nosacījumus un kļūdu sekas. Izvēlieties divus kvantitatīvus rādītājus un vienu kvalitatīvu avotu. Norādiet datu kvalitātes pārbaudes, minimālās parauga prasības, nenoteiktības ziņošanu un pieņemšanas kritērijus.
- Pievienojiet lēmuma noteikumu par nepietiekamiem pierādījumiem, atbildīgo personu, pārskatīšanas datumu un procedūru, kādā ietekmētās personas var apstrīdēt rezultātu.
- Pievienojiet 5 punktu pārbaudes sarakstu „Inclusion-by-Design”, kas aptver dažādus datus, ieinteresēto pušu viedokli, pieejamību, dokumentāciju, testēšanu un ieviešanu.
- **Rezultāts:** vienas lappuses testa plāns, pārbaudes saraksts un divu minūšu pamatojums.

03

Dizaina izstāde

un iekļaujošās kontroles

Pārvērtiet secinājumus
tehniskās, organizatoriskās,
līdzdalības un uzraudzības
kontroles pasākumos.

Sāciet riska mazināšanas pasākumus pirms modeļa apmācības

Pirms modelēšanas izstrādāji pārbaudes sarakstu „Iekļaušana jau projektēšanas posmā”. Dokumentējiet datu izcelsmi, mērķi, piekrišanu vai juridisko pamatu, zināmos trūkumus, paredzēto lietojumu un pārstāvības ierobežojumus. Iekļaujiet ieinteresēto personu pārskatīšanu gadījumos, kad mērķis, aizstājējs vai paredzētais lietojums var ietvert sociālu nevienlīdzību.

Izmantojiet paraugu ņemšanu un svēršanu tikai tad, ja tās risina identificēto mehānismu.

[Piekluve]

- Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgos vizuālos zīmolus un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Apmācības un pēcapstrādes kontroles mehānismiem nepieciešams skaidri formulēts taisnīguma mērķis

- Modeļa ierobežojumi, pārsvēršana, atkārtota paraugu ņemšana, reprezentācijas apguve un sliekšņa pielāgošana var mainīt grupu rezultātus. To piemērotība ir atkarīga no izvēlētā taisnīguma mērķa un riska konteksta.
- Pēc katras izmaiņas salīdziniet veikspēju, taisnīgumu, nenoteiktību un izturību. Pārbaudiet, vai vienas grupas ieguvumi nerada nepieņemamus zaudējumus citai grupai vai neapdraud drošību.

Reģistrējiet kompromisu un saglabājiet atsauces līmeni, lai intervences ietekme paliktu redzama.

Cilvēka veiktajai pārskatīšanai jābūt izstrādātai kā reālai kontrolei

Izstrādājiēt pārskatīšanas izraisītāju, pierādījumu parādīšanu, lēmuma pieņemšanas iespējas, eskalācijas ceļu un reģistrēšanu. Pārbaudiet pārskatītāju konsekvenci un darba slodzi. Uzraudzības komandām jābūt atbilstošām tehniskām, nozares un praktiskās pieredzes perspektīvām. Sniedziet skartajām personām skaidru paziņojumu un pieejamu ceļu uz labojumu vai pārsūdzību.

Daudzveidīgām komandām nepieciešamas lēmumu pieņemšanas tiesības un pierādījumi

Komandas daudzveidība var atklāt „aklos punktus”, taču pārstāvība pati par sevi negarantē ietekmi. Standarts ISO 30415 definē iekļaušanu kā kultūru, taisnīgumu, cieņu un atbildību; inženierijas praksē tas jāpārvērš lēmumu pieņemšanas tiesībās. Jānosaka, kurš var apšaubīt mērķi, apturēt ieviešanu, pieprasīt apakšgrupas pierādījumus un apstiprināt atlikušo risku.

Ārējais gadījums: „Rite Aid” sejas atpazīšana

- ASV Federālā tirdzniecības komisija apgalvoja, ka Rite Aid izmantoja sejas atpazīšanu bez saprātīgām procedūrām, lai novērstu kaitējumu patērētājiem. Sūdzībā tika aprakstīti viltus pozitīvie brīdinājumi un kļūdas testēšanā, attēla kvalitātē, personāla apmācībā, uzraudzībā un riska novērtēšanā.
- **Inženierijas mācība:** apakšgrupas darbība ir viena no kontroles sistēmas sastāvdaļām. Attēla kvalitāte, datu bāzes noteikumi, operatora rīcība, eskalācija un incidentu pierādījumi arī ietekmē kaitējuma rašanos.

Jautājiet, kāda kontroles sistēma būtu bijusi jāievieš, lai novērstu ieviešanu, un kādi pierādījumi būtu bijuši jāizmanto, lai apturētu sistēmas darbību.

Pārvērtiet „Rite Aid” mācības kontroles pasākumos

- **Taisnīguma problēmai reti ir viens vienīgs tehnisks risinājums.**

Rite Aid gadījums parāda, kā kaitējumu var izraisīt vairāki savstarpēji saistīti trūkumi — slikta ievades kvalitāte, vāji saskaņošanas noteikumi, operatoru rīcība, nepietiekama apmācība un pārvaldības trūkums. Tāpēc efektīvām kontroles sistēmām jādarbojas visā sistēmā.

- **Attēla kvalitāte**

Noraidiet ievadi, kas neatbilst pārbaudītajai kvalitātes robežvērtībai.

- **Saskaņošanas pierādījumi**

Attiecīgā gadījumā norādiet saskaņošanas ticamību, attēla kvalitātes robežvērtības un alternatīvus kandidātus.

Pārvērtiet „Rite Aid” mācības par kontroles pasākumiem

○ **Operatora rīcība**

Pirms jebkādas nelabvēlīgas rīcības pieprasiet neatkarīgu pārskatīšanu. Brīdinājums pats par sevi nekad nedrīkst noteikt iznākumu

○ **Apmācība**

Pārbaudiet personāla kompetenci, izmantojot reālistiskus scenārijus un atkārtotas pārbaudes.

○ **Uzraudzība**

Uzskaitiet viltus trauksmes, skartās grupas, sūdzības un jaunus kaitējuma modeļus.

Pārvērtiet „Rite Aid” gūtās atziņas kontroles pasākumos

○ Pārvaldība

Prasiet apstiprinājumu pirms lietošanas paplašināšanas un definējiet skaidrus iemeslus, kad lietošana jāaptur.

○ Inženierijas princips

Katram kontroles pasākumam ir nepieciešams izmērāms pieņemamības kritērijs, atbildīgais īpašnieks un pierādījumi, ka tas darbojas

Dilemma: taisnīga sadale, ja trūkst datu

Lēmums

TURPINĀT, PĀRSKATĪT vai PĀRTRAUKT. Noteikt taisnīgus noteikumus neskaidriem gadījumiem, minimālo pierādījumu apjomu, cilvēka veiktu pārskatīšanu, pieejamu pārsūdzības iespēju, uzraudzību, atbildīgo īpašnieku un to, kā pārskatīšanas procesā tiek pārstāvēti pieaugušie, starptautiskie vai studijas pārtraukušie studenti.

[Piekluve] Piedāvājiēt uzdevumu rakstiskā, mutiskā vai vizuālā formā. Nolasiet scenāriju skaļi, paskaidrojiet akronīmus un nodrošiniēt pieejamu digitālu uzdevuma izpildi.

04

Veikt un aizstāvēt inženierijas lēmumu

Veiciet neobjektivitātes pārbaudi, reģistrējiet atlikušo risku un aizstāviet lēmumu, izmantojot pierādījumus.

Projekta gadījums A: Pārmērība datos

- Piemērs „Ētiskais inženieris” parāda, kā nereprezentatīvi vai vēsturiski neobjektīvi dati var radīt neprecīzas prognozes un nevienlīdzīgus rezultātus. Izmantojiet to, lai identificētu konkrēto datu ģenerēšanas procesu, nevis uzskatītu neobjektivitāti par vispārēju īpašību.
- Noteiet, kas ir izlaists, kurš mainīgais ir izkropļots un kā tas ietekmē lēmumu. Ierosiniet vienu datu kontroles pasākumu, vienu sistēmas kontroles pasākumu un vienu uzraudzības pasākumu.

Atveriet gadījuma izpēti „Priekšstatī datos” un reģistrējiet atlikušo nenoteiktību.

Projekta gadījums B: AI vēsturisko fotogrāfiju novērtēšanā

- Šajā gadījumā tiek apskatīta AI sistēma, kas prognozē, kuras vēsturiskās fotogrāfijas var piesaistīt uzmanību vai radīt ieņēmumus. No interneta iegūtie dati var veicināt popularitāti, vienlaikus atņemot jutīgām fotogrāfijām to vēsturisko kontekstu.
- Inženierijas dilemma attiecas uz mērķi, apmācības signālu, kopienu pārstāvību un kuratoru izmantoto saskarni.

Atveriet vēsturisko fotogrāfiju novērtēšanas gadījumu. Izlemiet, kā kultūras konteksts, ieinteresēto pušu pārskatīšana un ierobežojumi automatizētajai rangu noteikšanai būtu jāmaina dizainu.

Pirms izvēlēties risinājumu,
salīdziniet gadījumus

Pielāgojiet kontroli neobjektivitātes mehānismam, nevis apzīmējumam „neobjektivitāte”.

- **Priekšstati datos sākas ar pierādījumu kvalitāti:** kas ir izlaists, kurš mainīgais ir izkropļots un kā tas ietekmē lēmumu.
- **Vēsturisko fotogrāfiju novērtēšana rada vērtību konfliktu:** popularitāte, ieņēmumi un kultūras nozīme var norādīt uz dažādām tendencēm.

Pirms kontroles pasākuma izvēles uzdot četrus jautājumus:

- Kura sistēmas izvēle rada nelabvēlīgu situāciju?
- Kādi pierādījumi liecinātu, ka kaitējums ir reāls?
- Kura kontroles metode maina mehānismu: mērķis, dati, sliekšnis, saskarne, pārskatīšana vai ieviešanas ierobežojums?
- Kāds atlikušais risks paliek pēc kontroles pasākuma piemērošanas?

[Piekluve] - Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaida tekstu pieejamā digitālajā formātā.

Komandas uzdevums: Veiciet neobjektivitā tes pārbaudi

Izvēlieties vienu „Ethical Engineer” gadījumu. Piešķiriet sistēmu, datu, testēšanas, ieinteresēto personu un reģistrētāja lomas.

1. noteikt paredzēto lietojumu un kritisko lēmumu.
2. izveidojiet neobjektivitātes ceļu.
3. izvēlieties taisnīguma mērķi un pierādījumus.
4. ierosiniet tehniskos un organizatoriskos kontroles pasākumus.
5. norādiet trūkstošos pierādījumus, atlikušo risku un risinājumu.
6. Nolemjiet: turpināt (GO), pārskatīt (REVISE) vai pārtraukt (STOP).

Rezultāts

Aizpildīts revīzijas protokols un divu minūšu garš, uz pierādījumiem balstīts lēmuma kopsavilkums.

[Piekluve] – Piedāvāji uzdevumu rakstiskā, mutiskā vai vizuālā formā. Nolasiet scenāriju skaļi, paskaidrojiet akronīmus un nodrošiniet pieejamu digitālu izpildi.

No analīzes līdz pamatotam inženierijas lēmumam

Rūpīga neobjektivitātes pārbaude pārvērš pierādījumus par pamatotu argumentu



Darba lapa: Lēmuma protokols par neobjektivitāti un taisnīgumu

Sistēma un paredzētais lietojums:

Kritiskais lēmums un skartās personas:

Priekšstatu mehānisms / ceļš:

Taisnīguma mērķis un tā atbilstība:

Pierādījumi, grupas, apstākļi un nenoteiktība:

Iekļaušana / ieinteresēto pušu pierādījumi:

[Pieklūve] Uzdevumu piedāvāji rakstiskā, mutiskā vai vizuālā formā. Nolasiet scenāriju skaļi, paskaidrojiet akronīmus un nodrošiniet pieejamu digitālu uzdevuma izpildi.

Turpinājums nākamajā slaidā

Tehniskā kontrole un pieņemšanas tests:

Pārvaldība / iekļaušanas kontrole un atbildīgā persona:

Lēmums: TURPINĀT PĀRSKATĪT PĀRTRAUKT

Atlikušais risks un uzraudzības izraisītājs:

Risinājums / Apelācijas kārtība:

Nesaskaņa vai alternatīvs viedoklis:

[Piekluve] Uzdevumu piedāvāji rakstiskā, mutiskā vai vizuālā formā. Nolasiet scenāriju skaļi, paskaidrojiet akronīmus un nodrošiniet pieejamu digitālu uzdevuma izpildi.

Lomu spēle: Iepirkumu pārskatīšanas komisija

- Pilsēta plāno iegādāties mākslīgā intelekta sistēmu, kas nosaka prioritātes mājokļu pārbaudēm. Piegādātājs nodrošina vispārēju precizitāti, bet nesniedz rezultātus par apakšgrupām vai apkaimēm. Apgabalos, kur pastāv valodas barjeras un zema uzticēšanās, vēsturiskie dati par sūdzībām ir nepilnīgi.
- **Lomas:** iepirkumu vadītājs, mājokļu inženieris, pārdevējs, kopienas pārstāvis, datu aizsardzības speciālists un drošības vadītājs.

Vienojieties par pierādījumiem un līguma nosacījumiem, kas nepieciešami pirms izmēģinājuma posma. Norādiet vienu nosacījumu, kad jāveic labojumi (REVISE), un vienu, kad jāpārtrauc (STOP). Fiksējiet, kurš var apturēt sistēmas lietošanu un kā iedzīvotāji var apstrīdēt kādu izlaidumu.

[Piekluve] Piedāvāji uzdevumu rakstiskā, mutiskā vai vizuālā formā. Nolasiet scenāriju skaļi, paskaidrojiet akronīmus un nodrošiniet pieejamu digitālu aizpildīšanu.

Individuālā pārdomāšan & Pārnese

Uzrakstiet īsu
inženierzinātnisk
u argumentāciju.

1. Kura sistēmas izvēle radīja vislielākās bažas par taisnīgumu vai iekļaušanu?
2. Kādu taisnīguma mērķi jūs izvēlējāties un kas paliek neatrisināts?
3. Kuras personas perspektīva vai pieredze nebija iekļauta sākotnējos pierādījumos?
4. Kādi jauni pierādījumi liktu jums mainīt savu lēmumu?
5. Kādi tehniskie, organizatoriskie un iekļaušanas kontroles pasākumi ir jāiekļauj prasībās?
6. Kādu atlikušo risku jūs pieņemt, un kas liktu jums pārtraukt ieviešanu?
7. Kā mainītos jūsu argumentācija citā inženierzinātņu nozarē?

Rezultāts:

Saistiet skartās personas, pierādījumus, kontroles pasākumus un pamatotu lēmumu.

[Piekluve] Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgos vizuālos zīmolus un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.

Tests

1. Kāpēc augsta kopējā precizitāte var pastāvēt vienlaikus ar netaisnīgiem rezultātiem?
2. Kāda ir atšķirība starp paraugu atlases novirzi un marķējuma novirzi?
3. Kāpēc taisnīguma rādītājam ir jābūt saistītam ar lietošanas kontekstu?
4. Ko inženierim jāziņo, ja pierādījumi par apakšgrupām ir ierobežoti?
5. Nosauciet divus nosacījumus, kas nepieciešami nozīmīgai cilvēka veiktai pārskatīšanai.
6. Kura kontrole visvairāk tieši risina problēmu ar nederīgu aizstājējmainīgo?
7. Kas jāiekļauj taisnīguma lēmuma protokolā?
8. Kā daudzveidība un iekļaušana var uzlabot inženierijas procesu?

Atbilžu uz jautājumiem un mācību pamatojums

1. Kopējie rezultāti var slēpt grupai un apstākļiem raksturīgas kļūdas.
2. Paraugu atlases neobjektivitāte attiecas uz pārstāvniecību; marķējuma neobjektivitāte attiecas uz mērķi vai reģistrēto vērtējumu.
3. Rādītāji atspoguļo dažādus taisnīguma mērķus un kļūdu sekas.
4. Ziņojiet par atšķirībām un nenoteiktību, pēc tam ierobežojiet vai pārskatiet izmantošanu.
5. Nozīme ir autoritātei, laikam, informācijai, kompetencei un reģistrācijai.
6. Pārskatiet mērķi, pazīmi vai procesu, kas radīja aizstājējradītāju.
7. Lēmums, ietekmētās personas, mērķis, pierādījumi, kontroles pasākumi, atbildīgā persona, atlikušais risks, uzraudzība un korektīvie pasākumi.
8. Paplašinot pierādījumus, atklājot pieņēmumus, pastiprinot pārskatīšanu un nodrošinot ietekmētajām personām iespējas apstrīdēt lēmumus.

Mācību rezultātu pārbaude

Pirms pabeigšanas pārliedzinieties, ka varat:

1. Izsekojiet aizspriedumus caur AI sistēmu un tās lēmumu kontekstu.
2. Izvēlieties un pamatojiet taisnīguma mērķi.
3. Novērtējiet sniegumu attiecīgajās un savstarpēji saistītajās grupās, ņemot vērā nenoteiktību.
4. Paskaidrojiet, kā daudzveidība, līdzdalība un iekļaušana maina prasības un pierādījumus.
5. Ieteikiet tehniskos, organizatoriskos, līdzdalības un cilvēku veikto pārskatīšanas kontroles pasākumus.
6. Aizstāviet lēmumu „TURPINĀT / PĀRSKATĪT / PĀRTRAUKT” ar atlikušo risku, uzraudzību un korektīvajiem pasākumiem.

Iegūtie pierādījumi:

- Neobjektivitātes ceļa karte • Taisnīguma pārbaudes plāns • Iekļaušana jau projektēšanas posmā
- Pārbaudes saraksts • Neobjektivitātes audits • Inženierijas lēmuma kopsavilkums.

[Piekluve] Izmantojiet pieejamu lasīšanas secību, pēc pieprasījuma aprakstiet attiecīgo vizuālo zīmolu un nodrošiniet slaidu tekstu pieejamā digitālajā formātā.



Paziņojums par finansējumu un izmantošanu izglītības nolūkos

Šis atklātais izglītības resurss atbalsta mācīšanu un diskusijas. Tas nav uzskatāms par juridisku, normatīvu, drošības vai profesionālu inženierijas konsultāciju. Pielietojot tā saturu, ievērojiet spēkā esošos likumus, standartus, iestāžu politiku un nozarei specifiskos norādījumus.

Līdzfinansē Eiropas Savienības programma „Erasmus+“.





Ethical Engineer

www.ethicalengineer.eu

Paldies

Jūs esat pabeidzis 2. moduli



Co-funded by
the European Union