

Tervetuloa Algoritmin perusteet ja tekoälyn etiikka: Teoriasta käytäntöön -kurssille

Työkalu synkronoiduille tunneille

**CU4 | Datan oikeudenmukaisuus ja
vinoumat**

Tuki PowerPoint slidet

INDEKSI

- JOHDANTO - 3
- JOHDANTO DATAN OIKEUDENMUKAISUUTEEN JA VINOUMIIN - 7
- VINOUMIEN ERI TYYPIT TEKOÄLYJÄRJESTELMISSÄ - 20
- MENETELMÄT DATAN VINOUMIEN TUNNISTAMISEEN JA MITTAAMISEEN - 34
- STRATEGIAT VINOUMIEN TUNNISTAMISEEN JA LIEVENTÄMISEEN - 41
- OIKEUDENMUKAISUUSMITTAREIDEN YMMÄRTÄMINEN - 47
- OIKEUDENMUKAISUUS DATAN KERÄÄMISESSÄ JA ESIKÄSITTELYSSÄ - 58
- TULEVAISUUDEN SUUNTAUKSET JA KASVAVAT ONGELMAT - 67
- JOHTOPÄÄTÖKSET - 76

JOHDANTO



IMAGE SOURCE | Generated by DALL-E

Ehdotus

Aloita **interaktiivisella kyselyllä tai kysymyksellä**, jolla kartoitat osallistujien lähtötason ymmärrystä tai henkilökohtaisia kokemuksia tekoälyn oikeudenmukaisuudesta ja vinoumista. Tämä voi tehdä johdannosta kiinnostavamman ja auttaa hahmottamaan, miksi aihe on merkityksellinen.

- Esimerkki: **Kuinka hyvin tunnet tekoälyn oikeudenmukaisuudet ja vinoumat?**
 - En tunne asiaa lainkaan.
 - Olen kuullut siitä, mutta en tunne yksityiskohtia.
 - Ymmärrän perusteet, mutta haluaisin oppia lisää.
 - Tunnen aiheen hyvin ja olen työskennellyt tekoälyjärjestelmien parissa, huomioiden oikeudenmukaisuuden ja vinoumat.

TÄSTÄ OSAAMISKOKONAISUUDESTA LÖYDÄT SEURAAVAT AIHEET:

- Datan oikeudenmukaisuuden perusteet ja vinoumat tekoälyjärjestelmissä
- Erilaiset vinoumat jotka voivat vaikuttaa tekoälyjärjestelmiin
- Kuinka huomata ja mitata vinoumia datasta
- Menetelmät vinoumien vähentämiseksi ja oikeudenmukaisuuden lisäämiseksi
- Mittarin tekoälyn oikeudenmukaisuuden mittaamiseksi
- Parhaat tavat oikeudenmukaiseen datan keräämiseen ja prosessointiin
- Nousevat trendit tekoälyn oikeudenmukaisuudessa
- Oikeudenmukaisuus ja vinoumat tekoälyssä

OSAAMISKOKONAISUUDEN LOPUSSA OSAAT:

- Selittää oikeudenmukaisuuden periaatteet sekä vinoumien merkityksen tekoälyssä
- Tunnistaa useita vinouemien muotoja jotka vaikuttavat tekoälyjärjestelmiin
- Arvioida tekoälyjärjestelmiä oikeudenmukaisuusmittareilla
- Seurata parhaita käytäntöjä datan keräämisessä ja esikäsittelyssä oikeudenmukaisuuden varmistamiseksi

JOHDANTO DATAN OIKEUDENMUKAISUUTEEN JA VINOUMIIN



IMAGE SOURCE | Generated by DALL-E

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Oikeudenmukaisuus & vinoumat - karkki-esimerkki

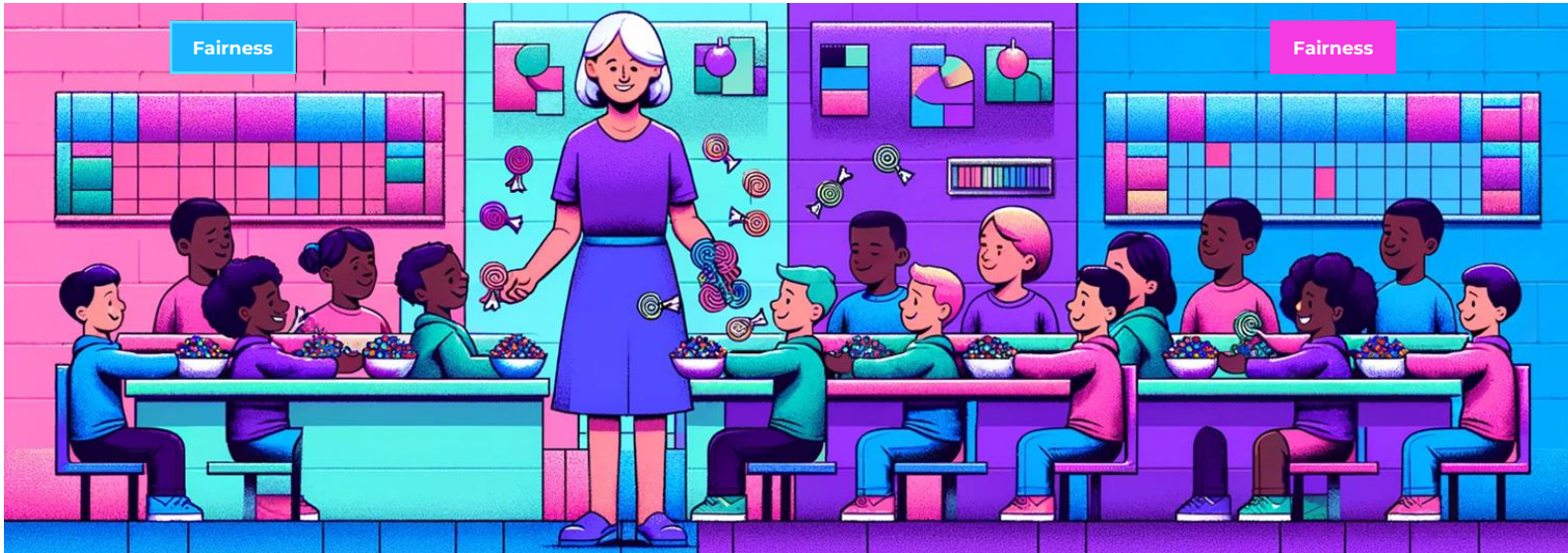


IMAGE SOURCE | Generated by DALL-E

Tutustu siihen, miten oikeudenmukaisuus ja puolueellisuus tekoälyjärjestelmissä voivat vaikuttaa päivittäisiin valintoihimme, aivan kuten opettajan säännöt karkkien jakamisesta voivat muokata luokkakokemusta.

Avataan tekoälyn oikeudenmukaisuuden käsite - yksi makea esimerkki kerrallaan!

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Oikeudenmukaisuus & vinouma - karkki-esimerkki



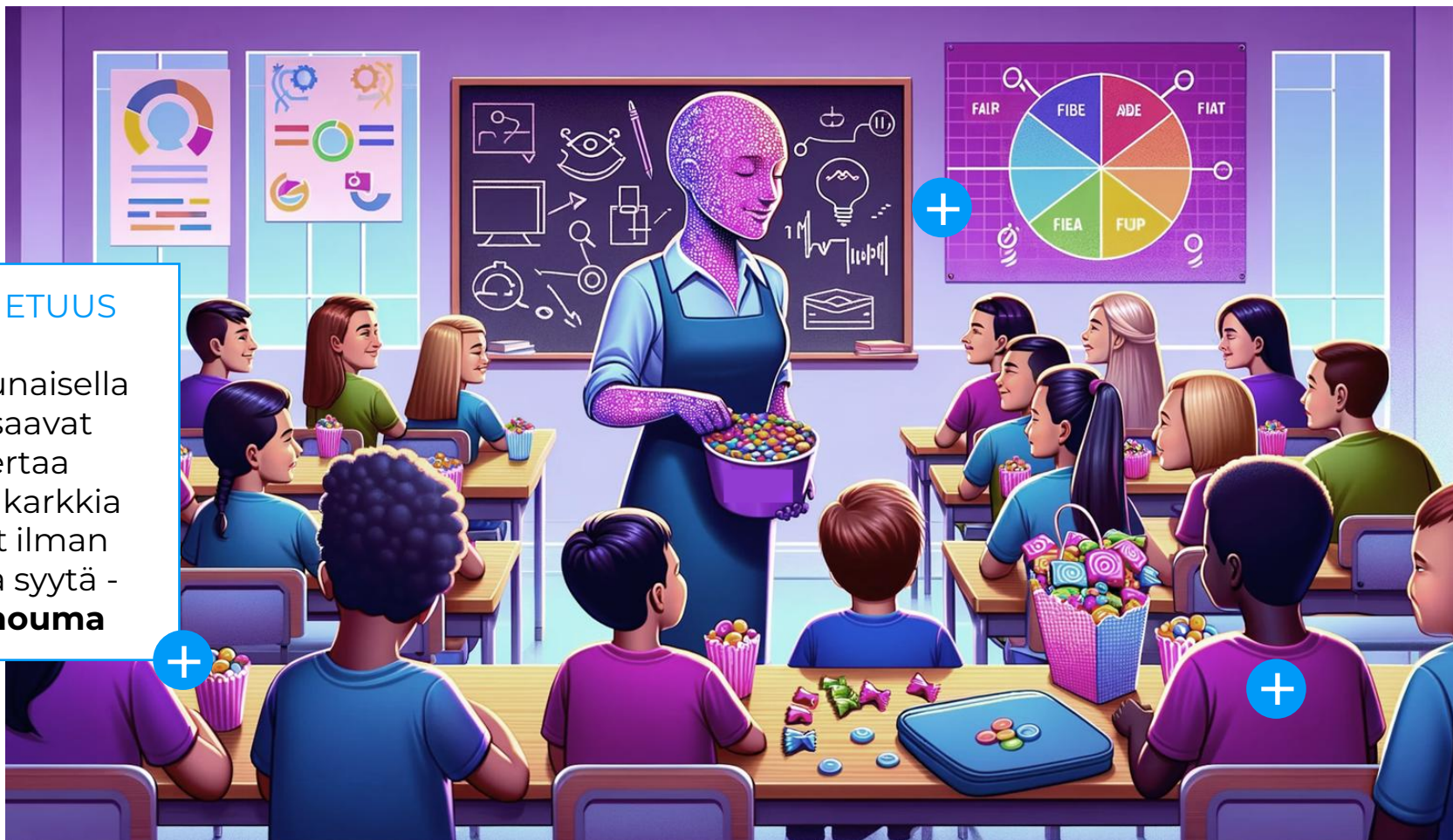
**Kuvittele tilanne, jossa mahdollisuutesi saada lisää karkkia riippuen siitä,
mitä sinulla on ylläsi!**

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Datan oikeudenmukaisuus ja vinoumat - karkki-esimerkki

EPÄREILU ETUUS

Oppilaat punaisella paidalla saavat kaksi kertaa enemmän karkkia kuin muut ilman kunnollista syytä - sitä on **vinouma**



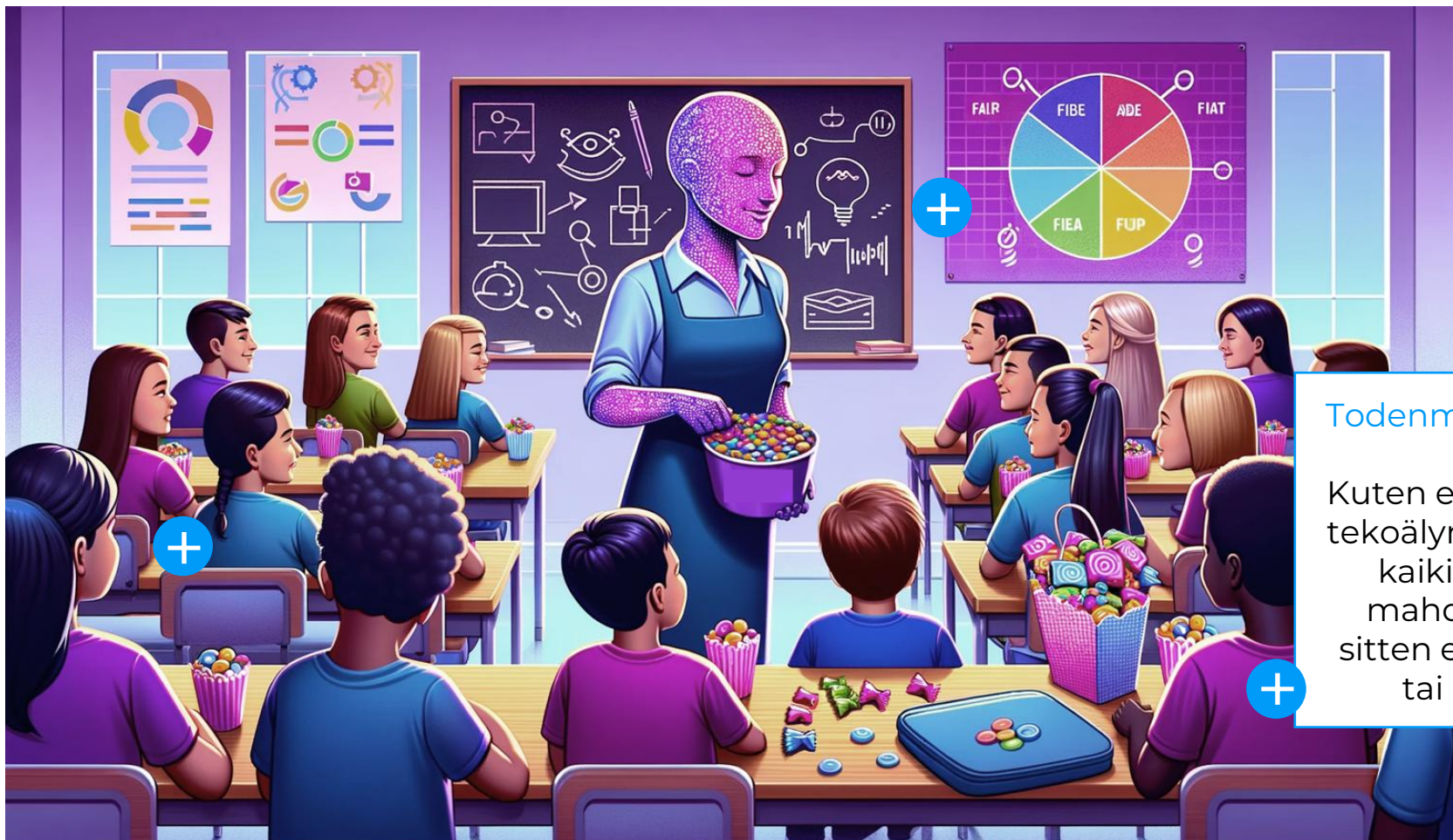
Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Oikeudenmukaisuus ja vinoumat - karkki-esimerkki



Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Oikeudenmukaisuus ja vinoumat - karkki-esimerkki



Todenmukainen tekoäly

Kuten esimerkissämme, tekoälyn tulisi varmistaa kaikille yhtäläinen mahdollisuus, oli se sitten elokuva suositus tai lainapäätös.

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Datan oikeudenmukaisuus

Datan oikeudenmukaisuus (data fairness) tarkoittaa sitä, että tekoälyjärjestelmät kohtelevat kaikkia yksilöitä tasapuolisesti ilman, että vinouma vaikuttaa lopputuloksiin. Se kattaa kaikki vaiheet datan keräämisestä, käsittelystä ja käytöstä, joiden avulla pyritään ehkäisemään syrjiviä vaikutuksia.

Datan oikeudenmukaisuus on keskeinen osa tekoälyn suunnittelua ja toteutusta, jotta se vaikuttaisi myönteisesti yhteiskuntaan. Se varmistaa, että tekoälyn tekemät päätökset ovat oikeudenmukaisia ja puolueettomia, edistään tasa-arvoa ja estäen syrjintää.

Reilut datakäytännöt johtavat luotettaviin tekoälysovelluksiin, joita voidaan laajasti hyväksyä ja jotka hyödyttävät monimuotoisia yhteisöjä. Tavoitteena on, että tekoäly palvelee kaikkia tasapuolisesti, riippumatta taustasta tai väestöryhmään kuulumisesta.

Johdanto datan oikeudenmukaisuuteen ja vinoumiin —

Mikä on vinouma?	+
Miten se ilmenee?	+
Millaisia seurauksia vinoumilla on?	+

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Mikä on vinouma?

Tekoälyn vinouma tarkoittaa systemaattisia virheitä, jotka suosivat epäoikeudenmukaisesti yhtä ryhmää muiden kustannuksella. Tämä voi johtua vinoutuneesta datasta, virheellisistä algoritmeista tai ennakkoluuloisesta koulutusprosessista.

Miten se ilmenee?

+

Millaisia seurauksia vinoumilla on?

+

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Mikä on vinouma?

+

Miten se ilmenee?

-

Datavinouma (data bias): Kun tekoälylle syötetty data heijastaa historiallisia epäoikeudenmukaisuuksia tai antaa puutteellisen kuvan erilaisista ryhmistä.

Algoritminen vinouma (algorithm bias): Kun päätöksiä tekevät säännöt tai mallit vahvistavat olemassa olevia ennakkoluuloja tai ohittavat vähemmistöjen erityispiirteet.

Palautevirrat (feedback loops): Kun tekoälyn vinoutuneet päätökset vaikuttavat tulevaan dataan ja vahvistavat vinoumaa entisestään.

Mitä seurauksia vinoumilla on?

+

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Mikä on vinouma?	+
Miten se ilmenee?	+
Millaisia seurauksia vinoumilla on? Yhteiskunnallinen vaikutus: Vinouma voi vahvistaa stereotypioita, ylläpitää sosiaalista epätasa-arvoa ja heikentää luottamusta keskeisiin instituutioihin. Yksilötason vaikutus: Ihmiset voivat joutua epäoikeudenmukaisen kohtelun kohteeksi tekoälyn päätösten perusteella kriittisillä elämänaalueilla, kuten rekrytoinnissa, rikosoikeudessa tai lainoissa.	-

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Tosielämän skenaario: Tekoälyn vinouma

Rekrytointityökalun vinouma

Skenaario

Amazon luopui vuonna 2018 tekoälypohjaisesta rekrytointityökalustaan, koska se osoitti harhaa naisia kohtaan. Järjestelmä koulutettiin 10 vuoden ajalta kerätyillä ansioluetteloilla, jotka olivat pääosin miesten kirjoittamia – heijastaen teknologiasektorin miesvaltaisuutta. Tekoäly oppi suosimaan miesehdokkaita ja rankaisi ansioluetteloita, joissa esiintyi sana "naisten", kuten "naisten shakkikerhon kapteeni".

Vaikutus

Tapaus osoittaa, miten tekoäly voi ylläpitää historiallisia vinoumia, jos koulutusdata on puolueellista. Se myös korostaa, kuinka tekoäly voi vaikuttaa työllistymismahdollisuuksiin ja kuinka tärkeää on edistää monimuotoisuutta ja oikeudenmukaisuutta rekrytointijärjestelmissä.

Johdanto datan oikeudenmukaisuuteen ja vinoumiin

Tosielämän skenaario: tekoälyn vinouma

Terveysthuollon algoritmien vinouma

Skenaario

Vuonna 2019 *Science*-lehdessä julkaistu tutkimus paljasti, että laajasti Yhdysvaltain terveydenhuollossa käytetty algoritmi syrji mustia potilaita. Algoritmi arvioi hoidon tarvetta hoitokustannusten perusteella ja oletti virheellisesti, että pienemmät kustannukset tarkoittavat pienempää hoidon tarvetta. Todellisuudessa mustat potilaat olivat käyttäneet vähemmän terveydenhuoltopalveluja monien esteiden vuoksi, eivät siksi että olisivat terveempiä.

Vaikutus

Tämän seurauksena mustat potilaat saivat huomattavasti vähemmän pääsyä hoito-ohjelmiin, vaikka heidän terveydellinen tarpeensa olisi ollut suurempi. Tapaus osoittaa, miten tärkeää on varmistaa tekoälyn oikeudenmukaisuus terveydenhuollossa – virheelliset päätökset voivat suoraan vaikuttaa ihmisten terveyteen ja hyvinvointiin.

VINOUMIEN ERI TYYPIT TEKOÄLYJÄRJESTELMISSÄ



IMAGE SOURCE | Generated by DALL-E

Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinoma

Napauta kääntääksesi 

Otantavinouma

Napauta kääntääksesi 

Mittausvinouma

Napauta kääntääksesi 

Algoritminen
vinouma

Napauta kääntääksesi 

Vahvistusvinouma

Napauta kääntääksesi 

Raportointivinouma

Napauta kääntääksesi 

Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinouma syntyy, kun analyysiin käytetty otos ei edusta luotettavasti koko sitä väestöä, jota sen pitäisi kuvata. Tämä johtaa vinoutuneisiin tai virheellisiin johtopäätöksiin.

Otantavinouma

Napauta kääntääksesi 

Mittausvinouma

Napauta kääntääksesi 

Algoritminen
vinouma

Napauta kääntääksesi 

Vahvistusvinouma

Napauta kääntääksesi 

Raportointivinouma

Napauta kääntääksesi 

Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinouma syntyy, kun analyysiin käytetty otos ei edusta luotettavasti koko sitä väestöä, jota sen pitäisi kuvata. Tämä johtaa vinoutuneisiin tai virheellisiin johtopäätöksiin.

Otantavinouma on valintavinouman alalaji ja ilmenee silloin, kun datan keruumenetelmä suosii tiettyjä yksilöitä tai ryhmiä muiden kustannuksella.

Mittausvinouma

Napauta kääntääksesi 

Algoritminen
vinouma

Napauta kääntääksesi 

Vahvistusvinouma

Napauta kääntääksesi 

Raportointivinouma

Napauta kääntääksesi 

Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinouma syntyy, kun analyysiin käytetty otos ei edusta luotettavasti koko sitä väestöä, jota sen pitäisi kuvata. Tämä johtaa vinoutuneisiin tai virheellisiin johtopäätöksiin.

Otantavinouma on valintavinouman alalaji ja ilmenee silloin, kun datan keruumenetelmä suosii tiettyjä yksilöitä tai ryhmiä muiden kustannuksella.

Mittausvinouma syntyy mittausprosessin virheistä tai epäjohtonmukaisuuksista, mikä johtaa epätarkkoihin tai harhaanjohtaviin tuloksiin. Syy voi olla esimerkiksi viallisissa mittalaitteissa, epäselvissä kysymyksissä tai havainnoijan puolueellisuudessa.

Algoritminen
vinouma

Napauta kääntääksesi



Vahvistusvinouma

Napauta kääntääksesi



Raportointivinouma

Napauta kääntääksesi



Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinouma syntyy, kun analyysiin käytetty otos ei edusta luotettavasti koko sitä väestöä, jota sen pitäisi kuvata. Tämä johtaa vinoutuneisiin tai virheellisiin johtopäätöksiin.

Otantavinouma on valintavinouman alalaji ja ilmenee silloin, kun datan keruumenetelmä suosii tiettyjä yksilöitä tai ryhmiä muiden kustannuksella.

Mittausvinouma syntyy mittausprosessin virheistä tai epäjohtonmukaisuuksista, mikä johtaa epätarkkoihin tai harhaanjohtaviin tuloksiin. Syy voi olla esimerkiksi viallisissa mittalaitteissa, epäselvissä kysymyksissä tai havainnoijan puolueellisuudessa.

Algoritminen vinouma ilmenee koneoppimisalgoritmeissa, kun ne syrjivät systemaattisesti tiettyjä ryhmiä tai yksilöitä esimerkiksi rodun, sukupuolen tai muiden suojeltujen ominaisuuksien perusteella, joita esiintyy koulutusdatassa.

Vahvistusvinouma

Napauta kääntääksesi



Raportointivinouma

Napauta kääntääksesi



Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinouma syntyy, kun analyysiin käytetty otos ei edusta luotettavasti koko sitä väestöä, jota sen pitäisi kuvata. Tämä johtaa vinoutuneisiin tai virheellisiin johtopäätöksiin.

Otantavinouma on valintavinouman alalaji ja ilmenee silloin, kun datan keruumenetelmä suosii tiettyjä yksilöitä tai ryhmiä muiden kustannuksella.

Mittausvinouma syntyy mittausprosessin virheistä tai epäjohtonmukaisuuksista, mikä johtaa epätarkkoihin tai harhaanjohtaviin tuloksiin. Syy voi olla esimerkiksi viallisissa mittalaitteissa, epäselvissä kysymyksissä tai havainnoijan puolueellisuudessa.

Algoritminen vinouma ilmenee koneoppimisalgoritmeissa, kun ne syrjivät systemaattisesti tiettyjä ryhmiä tai yksilöitä esimerkiksi rodun, sukupuolen tai muiden suojeltujen ominaisuuksien perusteella, joita esiintyy koulutusdatassa.

Vahvistusvinouma syntyy, kun tutkijat tai analyytikot valikoivat tai tulkitsevat dataa tavalla, joka tukee heidän ennakkokäsityksiään tai hypoteesejaan, ja samalla sivuuttavat ristiriitaisen todistusaineiston.

Raportointivinouma

Napauta kääntääksesi ↺

Vinoumien eri tyypit tekoälyjärjestelmissä

Valintavinouma syntyy, kun analyysiin käytetty otos ei edusta luotettavasti koko sitä väestöä, jota sen pitäisi kuvata. Tämä johtaa vinoutuneisiin tai virheellisiin johtopäätöksiin.

Otantavinouma on valintavinouman alalaji ja ilmenee silloin, kun datan keruumenetelmä suosii tiettyjä yksilöitä tai ryhmiä muiden kustannuksella.

Mittausvinouma syntyy mittausprosessin virheistä tai epäjohtonmukaisuuksista, mikä johtaa epätarkkoihin tai harhaanjohtaviin tuloksiin. Syy voi olla esimerkiksi viallisissa mittalaitteissa, epäselvissä kysymyksissä tai havainnoijan puolueellisuudessa.

Algoritminen vinouma ilmenee koneoppimisalgoritmeissa, kun ne syrjivät systemaattisesti tiettyjä ryhmiä tai yksilöitä esimerkiksi rodun, sukupuolen tai muiden suojeltujen ominaisuuksien perusteella, joita esiintyy koulutusdatassa.

Vahvistusvinouma syntyy, kun tutkijat tai analyytikot valikoivat tai tulkitsevat dataa tavalla, joka tukee heidän ennakkokäsityksiään tai hypoteesejaan, ja samalla sivuuttavat ristiriitaisen todistusaineiston.

Raportintivinouma viittaa siihen, että tutkimustuloksia julkaistaan tai raportoidaan valikoivasti: tilastollisesti merkittäviä tai myönteisiä tuloksia julkaistaan, kun taas ei-merkitseviä tai kielteisiä tuloksia jätetään huomiotta.

Vinoumien eri tyypit tekoälyjärjestelmissä

Esimerkit

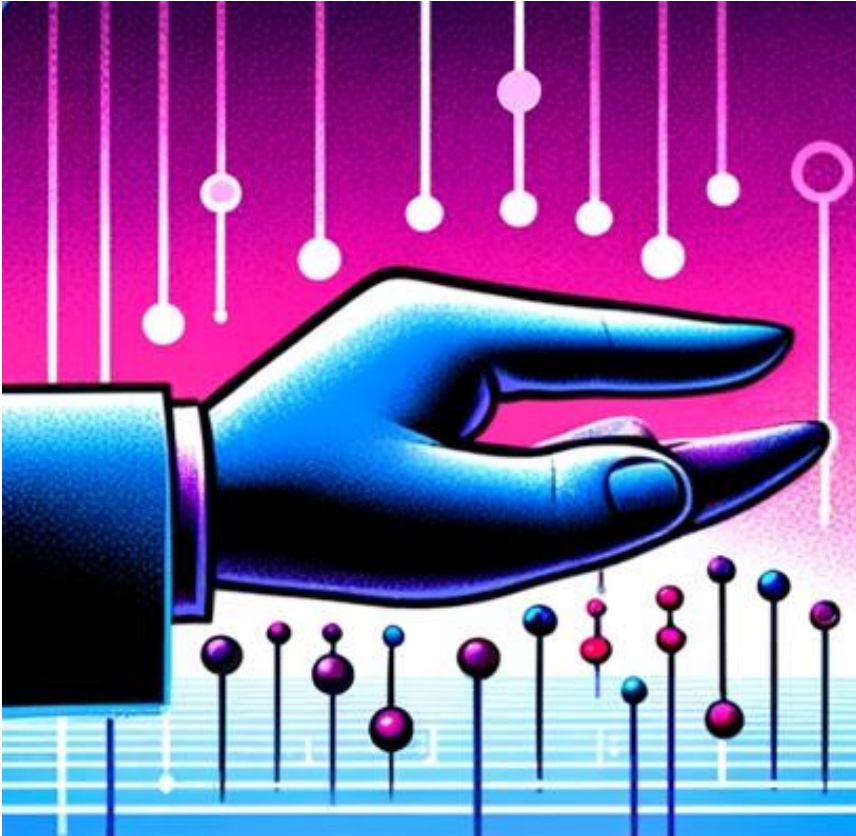


IMAGE SOURCE | Generated by DALL-E

Selection bias

Valintavinouma | Kasvojentunnistusteknologia

Kasvojentunnistusteknologioita on usein koulutettu pääasiassa länsimaalaisten kasvojen sisältämällä data-aineistoilla. Tämän monimuotoisuuden puutteen vuoksi teknologiat tekevät enemmän virheitä tunnistessaan muita etnisiä ryhmiä. Esimerkiksi MIT Media Labin tutkimuksessa havaittiin suuria tarkkuuseroja sukupuolen tunnistamisessa, erityisesti afroamerikkalaisten naisten kohdalla.

Vinoumien eri tyypit tekoälyjärjestelmissä

Esimerkit



IMAGE SOURCE | Generated by DALL-E

Sampling bias

Otantavinouma| Poliittiset kyselyvirheet

Vuoden 2016 Yhdysvaltain presidentinvaaleissa useat mielipidekyselyt aliarvioivat Donald Trumpin kannatusta. Yksi syy oli otantaharha: otosten demografinen painotus ei vastannut todellista äänestäjäkuntaa. Kaupunkiväestöä oli yliedustettuna, kun taas maaseutualueita oli aliedustettuna, mikä johti virheellisiin ennusteisiin ja vaalituloksen väärinarviointiin.

Vinoumien eri tyypit tekoälyjärjestelmissä

Esimerkit



IMAGE SOURCE | Generated by DALL-E Measurement bias

Mittausvinouma | Happisaturaatiomittarit ja ihonväri

COVID-19-pandemian aikana nousseet tutkimukset ovat osoittaneet, että pulssioksimetrit voivat näyttää harhaa arvioidessaan veren happipitoisuutta tummaihoisilla henkilöillä. Laitteiden sensorit eivät toimi yhtä tarkasti tumman ihon läpi, mikä voi johtaa vaarallisiin väärin diagnooseihin tai hoidon viivästymiseen.

Vinoumien eri tyypit tekoälyjärjestelmissä

Esimerkit



IMAGE SOURCE | Generated by DALL-E Algorithmic bias

Algoritminen vinouma | COMPAS:in rikostuomiot

Yhdysvalloissa käytetty COMPAS-ohjelmisto (Correctional Offender Management Profiling for Alternative Sanctions), jota käytetään arvioimaan rikoksenuusimisriskiä, osoittautui rasistisesti vinoutuneeksi. Analyysit osoittivat, että ohjelmisto ennusti mustien epäiltyjen uusivan rikoksia lähes kaksi kertaa todennäköisemmin virheellisesti kuin valkoisten. Tämä vaikutti tuomioihin ja takuusummiin algoritmin vinoutuneiden suositusten perusteella

Vinoumien eri tyypit tekoälyjärjestelmissä

Esimerkit



IMAGE SOURCE | Generated by DALL-E Confirmation bias

Vahvistusvinouma - Luottoluokitusmallit

Pankit ja rahoituslaitokset voivat vahingossa toistaa vahvistusharhoja käyttäessään tekoälypohjaisia luottoluokitusmalleja, jos ne perustuvat historialliseen lainadataan, jossa on ollut aiempaa syrjintää. Jos aiemmat ihmislainajat ovat suosineet tiettyjä väestöryhmiä, tekoälymalli voi oppia pitämään näitä ryhmiä vähemmän riskialttiina ja siten suosia heitä myös jatkossa.

Vinoumien eri tyypit tekoälyjärjestelmissä

Esimerkit



IMAGE SOURCE | Generated by DALL-E

Reporting bias

Raportointivinouma - Lääkkeiden haittavaikutukset verkossa

Lääkefirmat ja tutkijat, jotka käyttävät tekoälyä seurataksaan lääkkeiden haittavaikutuksia verkkofoorumeilla ja sosiaalisessa mediassa, voivat kohdata raportointivinoumaa. Vakavista oireista kärsivät potilaat ovat todennäköisemmin aktiivisia verkossa kuin ne, joilla on lieviä tai ei lainkaan oireita. Tämän seurauksena tekoäly voi yliarvioida haittavaikutusten vakavuuden tai yleisyyden, mikä vääristää lääkkeen turvallisuusprofiilia.

MENETELMÄT DATAN VINOUMIEN TUNNISTAMISEEN JA MITTAAMISEEN



IMAGE SOURCE | Generated by DALL-E

Menetelmät datan vinoumien tunnistamiseen ja mittaamiseen

Datan auditointi	+
Eroanalyysi	+
Algoritmissen oikeudenmukaisuuden testaus	+
Herkkyysanalyysi	+

Menetelmät datan vinoumien tunnistamiseen ja mittaamiseen

Datan auditointi

Datan auditointi tarkoittaa järjestelmällistä tietoaaineistojen tarkastelua, jonka tavoitteena on tunnistaa mahdollisia vinoumia, epätasapainoja tai poikkeavuuksia, jotka voivat viitata harhaan.

Tähän prosessiin sisältyy edustuksellisen tasa-arvon tarkastaminen eri ryhmien välillä aineistossa. Tavoitteena on varmistaa, ettei tiedonkeruuprosessi järjestelmällisesti sulje pois tiettyjä väestöryhmiä tai vahvista olemassa olevia stereotypioita.

Eroanalyysi

+

Algoritmisen oikeudenmukaisuuden testaus

+

Herkkyysanalyysi

+

Menetelmät datan vinoumien tunnistamiseen ja mittaamiseen

Datan auditointi

+

Eroanalyysi

-

tarkoittaa tekoälymallien tuottamien tulosten tai ennusteiden vertaamista eri väestö- tai suojeltujen ryhmien välillä epätasa-arvoisten vaikutusten tunnistamiseksi.

Tämä edellyttää tilastollisten mittareiden, kuten virheprosenttien, hyväksymisprosenttien tai muiden tulosuuttajien erojen laskemista ryhmien välillä. Näin voidaan mitata ja visualisoida mahdollisia harhoja tekoälyn päätöksenteossa.

Algoritmissen oikeudenmukaisuuden testaus

+

Herkyysanalyysi

+

Menetelmät datan vinoumien tunnistamiseen ja mittaamiseen

Datan auditointi	+
Eroanalyysi	+
Algoritmissen oikeudenmukaisuuden testaus tarkoittaa testien tai mittareiden käyttöä tekoälyjärjestelmän päätösten oikeudenmukaisuuden arvioimiseksi. Tässä voidaan käyttää esimerkiksi oikeudenmukaisuusmittareita kuten tasavertainen mahdollisuus (equal opportunity), väestöllinen tasapaino (demographic parity) tai yksilön oikeudenmukaisuus (individual fairness). Näiden avulla arvioidaan, tekeekö algoritmi reiluja ja puolueettomia päätöksiä eri väestöryhmien välillä.	-
Herkkyysanalyysi	+

Menetelmät datan vinoumien tunnistamiseen ja mittaamiseen

Datan auditointi	+
Eroanalyysi	+
Algoritmisen oikeudenmukaisuuden testaus	+
Herkkyysanalyysi tarkoittaa tekoälypäästösten vakauden testaamista muuttamalla syöttödataa hieman ja tarkkailemalla, miten lopputulokset vaihtelevat eri ryhmien kohdalla. Muuttamalla syöttöarvoja lievästi voidaan arvioida, kuinka herkkä malli on datan muutoksille. Tämä auttaa tunnistamaan mahdollisia vinouman lähteitä ja arvioimaan mallin ennusteiden luotettavuutta ja johdonmukaisuutta erilaisilla aineistoilla.	-

Menetelmät datan vinoumien tunnistamiseen ja mittaamiseen

Interactive workshop for identifying and measuring biases in data Recruitment tool bias

Tavoite: Tämä työpaja antaa osallistujille valmiudet soveltaa datan auditointia, eroanalyysiä ja herkkyysanalyysiä oikean tietoaaineiston avulla mahdollisten harhojen tunnistamiseksi ja ymmärtämiseksi.

Työkalut ja materiaalit

Data-aineisto: Tarjotaan yksinkertainen, julkisesti saatavilla oleva tietoaaineisto, joka sisältää demografisia muuttujia (esim. *Adult Income* -aineisto UCI Machine Learning -tietovarannosta, joka sisältää muuttujia kuten ikä, koulutus, rotu, sukupuoli ja tulot).

Toiminta: Osallistujat lataavat aineiston ja suorittavat perustason datatutkimuksen (puuttuvien arvojen tarkistus ja tärkeimpien muuttujien jakaumien tarkastelu).

tavoite: Tutustuttaa osallistujat aineistoon ja valmistella data analysointia varten.

Datan auditointi: Suoritetaan datan auditointi analysoimalla, kuinka eri ryhmät ovat edustettuina aineistossa. Lasketaan esimerkiksi eri sukupuoli- ja roturyhmien osuudet. Tunnistetaan mahdollisesti aliedustetut ryhmät.

Eroanalyysi: Osallistujat soveltavat eroanalyysiä arvioidakseen, miten hyvin yksinkertainen malli (esimerkiksi logistinen regressio tulojen ennustamiseen) ennustaa eri ryhmille. Tunnistetaan merkittäviä suorituskykyeroja ryhmien välillä ja pohditaan mahdollisia syitä.

Herkkyysanalyysi: Tehdään herkkyysanalyysi muuttamalla hieman tiettyjä datapisteitä ja tarkkaillaan, miten ennusteet muuttuvat. Muokataan esimerkiksi ikää tai koulutustasoa ja tarkastellaan, miten nämä muutokset vaikuttavat mallin tekemiin ennusteisiin.

MENETELMÄT DATAN VINOUMIEN TUNNISTAMISEEN JA LIEVENTÄMISEEN



IMAGE SOURCE | Generated by DALL-E

Menetelmät datan vinoumien tunnistamisen ja lieventämiseen

Kattava edustus

Napauta kääntääksesi



Vinoumista tietoinen algoritmi

Napauta kääntääksesi



Jatkuva seuranta ja arviointi

Napauta kääntääksesi



Läpinäkyvyys ja selitettävyys

Napauta kääntääksesi



Menetelmät datan vinoumien tunnistamisen ja lieventämiseen

Kattava edustus

Monipuolinen data eri väestöryhmistä varmistaa, että tekoälyjärjestelmät koulutetaan edustavalla aineistolla, mikä vähentää harhaisten lopputulosten todennäköisyyttä. Monimuotoisuus tuo mukanaan erilaisia näkökulmia ja oivalluksia, jotka auttavat tunnistamaan ja lieventämään harhoja kehitysprosessin aikana.

Jatkuva seuranta ja arviointi

Napauta kääntääksesi 

Vinoumista tietoinen algoritmi

Napauta kääntääksesi 

Läpinäkyvyys ja selitettävyys

Napauta kääntääksesi 

Menetelmät datan vinoumien tunnistamisen ja lieventämiseen

Kattava edustus

Monipuolinen data eri väestöryhmistä varmistaa, että tekoälyjärjestelmät koulutetaan edustavalla aineistolla, mikä vähentää harhaisten lopputulosten todennäköisyyttä. Monimuotoisuus tuo mukanaan erilaisia näkökulmia ja oivalluksia, jotka auttavat tunnistamaan ja lieventämään harhoja kehitysprosessin aikana.

Jatkuva seuranta ja arviointi

Napauta kääntääksesi



Vinoumista tietoinen algoritmi

Algoritmien suunnittelu sisältää sisäänrakennettuja mekanismeja vinoumien tunnistamiseen ja torjumiseen. Tämä voi tarkoittaa esimerkiksi oikeudenmukaisuusrajoitteiden sisällyttämistä algoritmin optimointiprosessiin tai menetelmien käyttöä, kuten vastakkainen koulutus, vinoumien minimoimiseksi mallien ennusteissa

Läpinäkyvyys ja selitettävyys

Napauta kääntääksesi



Menetelmät datan vinoumien tunnistamisen ja lieventämiseen

Kattava edustus

Monipuolinen data eri väestöryhmistä varmistaa, että tekoälyjärjestelmät koulutetaan edustavalla aineistolla, mikä vähentää harhaisten lopputulosten todennäköisyyttä. Monimuotoisuus tuo mukanaan erilaisia näkökulmia ja oivalluksia, jotka auttavat tunnistamaan ja lieventämään harhoja kehitysprosessin aikana.

Jatkuva seuranta ja arviointi

Tekoälyjärjestelmien jatkuva seuranta todellisissa käyttöympäristöissä mahdollistaa vinoumien havaitsemisen niiden ilmetessä. Säännöllinen arviointi tarkoittaa tekoälyn tuottamien tulosten analysointia eri väestöryhmien välillä esiintyvien erojen tunnistamiseksi sekä algoritmien tai aineistojen päivittämistä havaittujen vinoumien korjaamiseksi viipymättä.

Vinoumista tietoinen algoritmi

Algoritmien suunnittelu sisältää sisäänrakennettuja mekanismeja vinoumien tunnistamiseen ja torjumiseen. Tämä voi tarkoittaa esimerkiksi oikeudenmukaisuusrajoitteiden sisällyttämistä algoritmin optimointiprosessiin tai menetelmien käyttöä, kuten vastakkainen koulutus, vinoumien minimoimiseksi mallien ennusteissa

Läpinäkyvyys ja selitettävyys

Napauta kääntääksesi 

Menetelmät datan vinoumien tunnistamisen ja lieventämiseen

Kattava edustus

Monipuolinen data eri väestöryhmistä varmistaa, että tekoälyjärjestelmät koulutetaan edustavalla aineistolla, mikä vähentää harhaisten lopputulosten todennäköisyyttä. Monimuotoisuus tuo mukanaan erilaisia näkökulmia ja oivalluksia, jotka auttavat tunnistamaan ja lieventämään harhoja kehitysprosessin aikana.

Jatkuva seuranta ja arviointi

Tekoälyjärjestelmien jatkuva seuranta todellisissa käyttöympäristöissä mahdollistaa vinoumien havaitsemisen niiden ilmetessä. Säännöllinen arviointi tarkoittaa tekoälyn tuottamien tulosten analysointia eri väestöryhmien välillä esiintyvien erojen tunnistamiseksi sekä algoritmien tai aineistojen päivittämistä havaittujen vinoumien korjaamiseksi viipymättä.

Vinoumista tietoinen algoritmi

Algoritmien suunnittelu sisältää sisäänrakennettuja mekanismeja vinoumien tunnistamiseen ja torjumiseen. Tämä voi tarkoittaa esimerkiksi oikeudenmukaisuusrajoitteiden sisällyttämistä algoritmin optimointiprosessiin tai menetelmien käyttöä, kuten vastakkainen koulutus, vinoumien minimoimiseksi mallien ennusteissa

Läpinäkyvyys ja selitettävyys

Läpinäkyvyyden ja selitysten tarjoaminen tekoälypäätöksille mahdollistaa sidosryhmien ymmärtämisen, miten päätökset tehdään, ja tunnistaa mahdollisia järjestelmässä olevia vinoumia. Selitettävän tekoälyn (explainable AI) tekniikat, kuten piirteiden tärkeysjärjestyksen esittäminen tai päätösten perustelujen tarjoaminen, auttavat käyttäjiä tulkitsemaan ja arvioimaan tekoälyn tuottamien tulosten oikeudenmukaisuutta.

OIKEUDENMUKAISUUS MITTAREIDEN YMMÄRTÄMINEN



IMAGE SOURCE | Generated by DALL-E

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino	+
Yhtäläiset todennäköisyydet	+
Ehdollinen väestöllinen ero	+
Oikeudenmukaisuus tietoisuuden kautta	+
Yksilön oikeudenmukaisuus	+
Vastemuuttujan oikeudenmukaisuus	+

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino

tunnetaan myös nimellä **demografinen tasapaino (demographic parity)**. Tämä mittari arvioi, onko myönteisten lopputulosten (esim. lainojen hyväksyminen tai työtarjoukset) osuus sama eri väestöryhmien välillä.

Yhtäläiset todennäköisyydet

+

Ehdollinen väestöllinen ero

+

Oikeudenmukaisuus tietoisuuden kautta

+

Yksilön oikeudenmukaisuus

+

Vastemuuttujan oikeudenmukaisuus

+

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino	+
Yhtäläiset todennäköisyydet arkastelee, ovatko todellisten positiivisten osuus ja todellisten negatiivisten osuus samanlaisia eri väestöryhmien välillä. Se varmistaa, että mallin ennustetarkkuus on johdonmukainen kaikissa ryhmissä.	
Ehdollinen väestöllinen ero	+
Oikeudenmukaisuus tietoisuuden kautta	+
Yksilön oikeudenmukaisuus	+
Vastemuuttujan oikeudenmukaisuus	+

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino	+
Yhtäläiset todennäköisyydet	+
Ehdollinen väestöllinen ero arvioi ennustettujen tulosten eroa suojeltuun ominaisuuteen perustuen. Se mittaa ennusteiden eroja eri väestöryhmien välillä, huomioiden muut olennaiset tekijät.	
Oikeudenmukaisuus tietoisuuden kautta	+
Yksilön oikeudenmukaisuus	+
Vastemuuttujan oikeudenmukaisuus	+

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino	+
Yhtäläiset todennäköisyydet	+
Ehdollinen väestöllinen ero	+
Oikeudenmukaisuus tietoisuuden kautta	
tarkoittaa sitä, että tekoäly ottaa huomioon herkat muuttujat (kuten rodun tai sukupuolen) päätöksenteossa. Tavoitteena on tehdä oikeudenmukaisia ja puolueettomia päätöksiä huomioiden historiallinen ja yhteiskunnallinen konteksti.	
Yksilön oikeudenmukaisuus	+
Vastemuuttujan oikeudenmukaisuus	+

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino	+
Yhtäläiset todennäköisyydet	+
Ehdollinen väestöllinen ero	+
Oikeudenmukaisuus tietoisuuden kautta	+
Yksilön oikeudenmukaisuus keskittyy siihen, että samanlaisia yksilöitä kohdellaan samalla tavalla riippumatta heidän demografisista ominaisuuksistaan. Tavoitteena on minimoida epäoikeudenmukainen kohtelu, joka perustuu epäolennaisiin piirteisiin.	
Vastemuuttujan oikeudenmukaisuus	+

Oikeudenmukaisuusmittareiden ymmärtäminen

Tilastollinen tasapaino	+
Yhtäläiset todennäköisyydet	+
Ehdollinen väestöllinen ero	+
Oikeudenmukaisuus tietoisuuden kautta	+
Yksilön oikeudenmukaisuus	+
Vastemuuttujan oikeudenmukaisuus arvioi, muuttaisiko päätöksen lopputulos, jos yksilön herkkä ominaisuus (esim. rotu tai sukupuoli) vaihdettaisiin, mutta kaikki muut piirteet pidettäisiin ennallaan. Tavoitteena on varmistaa, että yksilöitä kohdellaan tasapuolisesti suojelluista ominaisuuksista riippumatta.	

Oikeudenmukaisuusmittareiden ymmärtäminen

Interaktiivinen tietovisa

Luo interaktiivinen tietovisa Mentimeter ohjelmalla

Kysymys

Mitkä ovat tärkeitä tekijöitä, kun arvioidaan, onko tekoälyjärjestelmä reilu?

Vaihtoehdot

a) Lopputulosten oikeudenmukaisuus b) Demografinen edustus c) Tasapuolinen kohtelu d) Läpinäkyvyys

Oikeudenmukaisuusmittareiden ymmärtäminen

Interaktiivinen luokkaharjoitus

Aihe: Tilastollinen tasapaino

Laske tilastollinen tasapaino

- Annettujen tietojen perusteella laske lainan hyväksymisprosentti eri ryhmille. Onko tasapaino saavutettu?
- Osallistujille annetaan hyväksymis- ja hylkäysmäärät (tai prosenttiosuudet) eri ryhmille, ja heidän tehtävänä on laskea, ovatko hyväksymisprosentit suunnilleen samat.
- Monivalintavastaukset: "Hyväksymisprosentit ovat samat; tasapaino on saavutettu." tai "Hyväksymisprosentit eroavat; tasapainoa ei ole."

Oikeudenmukaisuusmittareiden ymmärtäminen

Interaktiivinen luokkaharjoitus

Alhe Yhtäläiset todennäköisyydet

Arvioi yhtäläiset todennäköisyydet

- Oletetaan, että todellinen positiivinen osuus (lainat hyväksytty oikein) on miehillä 80 % ja naisilla 60 %. Onko yhtäläiset todennäköisyydet saavutettu?

Monivalintavastaukset: "Kyllä, todennäköisyydet ovat samat." tai "Ei, todennäköisyydet eivät ole samat."

OIKEUDENMUKAISUUS DATAN KERÄÄMISESSÄ JA ESIKÄSITTELYSSÄ



IMAGE SOURCE | Generated by DALL-E

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset
datalähteet

Napauta kääntääksesi ↺

Osallistava otanta

Napauta kääntääksesi ↺

Datan puhdistus

Napauta kääntääksesi ↺

Vinoumien tunnistus

Napauta kääntääksesi ↺

Oikeudenmukaisuus
ominaisuuksien
suunnittelu

Napauta kääntääksesi ↺

Läpinäkyvyys ja
dokumentaatio

Napauta kääntääksesi ↺

Jatkuva valvonta

Napauta kääntääksesi ↺

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Napauta kääntääksesi ↺

Datan puhdistus

Napauta kääntääksesi ↺

Vinoumien tunnistus

Napauta kääntääksesi ↺

Oikeudenmukaisuus
ominaisuuksien
suunnittelu

Napauta kääntääksesi ↺

Läpinäkyvyys ja
dokumentointi

Napauta kääntääksesi ↺

Jatkuva valvonta

Napauta kääntääksesi ↺

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Käytä osallistavia otantamenetelmiä varmistaaksesi, että kaikki väestöryhmät ovat riittävästi edustettuina aineistossa. Tämä voi tarkoittaa esimerkiksi kerrostettua otantaa tai vähemmistöryhmien ylliedustamista epätasapainojen korjaamiseksi.

Datan puhdistus

Napauta kääntääksesi ↺

Vinoumien tunnistus

Napauta kääntääksesi ↺

Oikeudenmukaisuus ominaisuuksien suunnittelu

Napauta kääntääksesi ↺

Läpinäkyvyys ja dokumentointi

Napauta kääntääksesi ↺

Jatkuva valvonta

Napauta kääntääksesi ↺

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Käytä osallistavia otantamenetelmiä varmistaaksesi, että kaikki väestöryhmät ovat riittävästi edustettuina aineistossa. Tämä voi tarkoittaa esimerkiksi kerrostettua otantaa tai vähemmistöryhmien ylliedustamista epätasapainojen korjaamiseksi.

Datan puhdistus

Puhdista ja esikäsittele data huolellisesti poistaaksesi mahdolliset harhat tai epäjohdonmukaisuudet. Tämä sisältää virheiden tunnistamisen ja korjaamisen, puuttuvien arvojen käsittelyn sekä tietomuotojen standardoinnin.

Vinoumien tunnistus

Napauta kääntääksesi ↻

Oikeudenmukaisuus ominaisuuksien suunnittelu

Napauta kääntääksesi ↻

Läpinäkyvyys ja dokumentointi

Napauta kääntääksesi ↻

Jatkuva valvonta

Napauta kääntääksesi ↻

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Käytä osallistavia otantamenetelmiä varmistaaksesi, että kaikki väestöryhmät ovat riittävästi edustettuina aineistossa. Tämä voi tarkoittaa esimerkiksi kerrostettua otantaa tai vähemmistöryhmien ylliedustamista epätasapainojen korjaamiseksi.

Datan puhdistus

Puhdista ja esikäsittele data huolellisesti poistaaksesi mahdolliset harhat tai epäjohtonmukaisuudet. Tämä sisältää virheiden tunnistamisen ja korjaamisen, puuttuvien arvojen käsittelyn sekä tietomuotojen standardoinnin.

Vinoumien tunnistus

Ota käyttöön menetelmiä vinoumien havaitsemiseksi ja vähentämiseksi datassa. Tämä voi sisältää vinouma-auditointeja tai automatisoituja työkaluja, joilla tunnistetaan eroja eri väestöryhmien välillä.

Oikeudenmukaisuus
ominaisuuksien
suunnittelu

Napauta kääntääksesi ↺

Läpinäkyvyys ja
dokumentointi

Napauta kääntääksesi ↺

Jatkuva valvonta

Napauta kääntääksesi ↺

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Käytä osallistavia otantamenetelmiä varmistaaksesi, että kaikki väestöryhmät ovat riittävästi edustettuina aineistossa. Tämä voi tarkoittaa esimerkiksi kerrostettua otantaa tai vähemmistöryhmien ylliedustamista epätasapainojen korjaamiseksi.

Datan puhdistus

Puhdista ja esikäsittele data huolellisesti poistaaksesi mahdolliset harhat tai epäjohtonmukaisuudet. Tämä sisältää virheiden tunnistamisen ja korjaamisen, puuttuvien arvojen käsittelyn sekä tietomuotojen standardoinnin.

Vinoumien tunnistus

Ota käyttöön menetelmiä vinoumien havaitsemiseksi ja vähentämiseksi datassa. Tämä voi sisältää vinouma-auditointeja tai automatisoituja työkaluja, joilla tunnistetaan eroja eri väestöryhmien välillä.

Oikeudenmukaisuus ominaisuuksien suunnittelu

Harkitse piirteiden valinnan ja muokkauksen vaikutuksia oikeudenmukaisuuteen. Vältä sellaisten muuttujien käyttöä, jotka voivat ylläpitää stereotyyppioita tai syrjintää, ja pyri valitsemaan merkityksellisiä ja syrjimättömiä ominaisuuksia.

Läpinäkyvyys ja
dokumentointi

Napauta kääntääksesi ↺

Jatkuva valvonta

Napauta kääntääksesi ↺

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Käytä osallistavia otantamenetelmiä varmistaaksesi, että kaikki väestöryhmät ovat riittävästi edustettuina aineistossa. Tämä voi tarkoittaa esimerkiksi kerrostettua otantaa tai vähemmistöryhmien ylliedustamista epätasapainojen korjaamiseksi.

Datan puhdistus

Puhdista ja esikäsittele data huolellisesti poistaaksesi mahdolliset harhat tai epäjohtonmukaisuudet. Tämä sisältää virheiden tunnistamisen ja korjaamisen, puuttuvien arvojen käsittelyn sekä tietomuotojen standardoinnin.

Vinoumien tunnistus

Ota käyttöön menetelmiä vinoumien havaitsemiseksi ja vähentämiseksi datassa. Tämä voi sisältää vinouma-auditointeja tai automatisoituja työkaluja, joilla tunnistetaan eroja eri väestöryhmien välillä.

Oikeudenmukaisuus ominaisuuksien suunnittelu

Harkitse piirteiden valinnan ja muokkauksen vaikutuksia oikeudenmukaisuuteen. Vältä sellaisten muuttujien käyttöä, jotka voivat ylläpitää stereotyyppioita tai syrjintää, ja pyri valitsemaan merkityksellisiä ja syrjimättömiä ominaisuuksia.

Läpinäkyvyys ja dokumentointi

Dokumentoi tiedonkeruu- ja esikäsittelyprosessit läpinäkyvyyden ja vastuun varmistamiseksi. Tämä sisältää esimerkiksi kaikki päätökset, jotka voivat vaikuttaa oikeudenmukaisuuteen, kuten datan siivouksen tai piirteiden muokkauksen aikana tehdyt valinnat.

Jatkuva valvonta

Napauta kääntääksesi ↺

Oikeudenmukaisuus datan keräämisessä ja esikäsittelyssä

Monipuoliset datalähteet

Kerää dataa monipuolisista lähteistä varmistaaksesi väestön kattavan edustuksen. Tämä auttaa estämään tiettyjen väestöryhmien aliedustusta tai vinoumia.

Osallistava otanta

Käytä osallistavia otantamenetelmiä varmistaaksesi, että kaikki väestöryhmät ovat riittävästi edustettuina aineistossa. Tämä voi tarkoittaa esimerkiksi kerrostettua otantaa tai vähemmistöryhmien yliedustamista epätasapainojen korjaamiseksi.

Datan puhdistus

Puhdista ja esikäsittele data huolellisesti poistaaksesi mahdolliset harhat tai epäjohtonmukaisuudet. Tämä sisältää virheiden tunnistamisen ja korjaamisen, puuttuvien arvojen käsittelyn sekä tietomuotojen standardoinnin.

Vinoumien tunnistus

Ota käyttöön menetelmiä vinoumien havaitsemiseksi ja vähentämiseksi datassa. Tämä voi sisältää vinouma-auditointeja tai automatisoituja työkaluja, joilla tunnistetaan eroja eri väestöryhmien välillä.

Oikeudenmukaisuus ominaisuuksien suunnittelu

Harkitse piirteiden valinnan ja muokkauksen vaikutuksia oikeudenmukaisuuteen. Vältä sellaisten muuttujien käyttöä, jotka voivat ylläpitää stereotyyppioita tai syrjintää, ja pyri valitsemaan merkityksellisiä ja syrjimättömiä ominaisuuksia.

Läpinäkyvyys ja dokumentointi

Dokumentoi tiedonkeruu- ja esikäsittelyprosessit läpinäkyvyyden ja vastuun varmistamiseksi. Tämä sisältää esimerkiksi kaikki päätökset, jotka voivat vaikuttaa oikeudenmukaisuuteen, kuten datan siivouksen tai piirteiden muokkauksen aikana tehdyt valinnat.

Jatkuva seuranta

Seuraa jatkuvasti datan keruu- ja esikäsittelyprosessia uusien vinoumien tai erojen havaitsemiseksi ja korjaamiseksi. Päivitä säännöllisesti aineistoa ja esikäsittelymenetelmiä oikeudenmukaisuuden säilyttämiseksi ajan myötä.

TULEVAISUUDEN SUUNTAUKSET JA KASVAVAT ONGELMAT



IMAGE SOURCE | Generated by DALL-E

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja
tulkittavuus

Napauta kääntääksesi ↺

Vastahyökkäykset ja
puolustus

Napauta kääntääksesi ↺

Oikeudenmukaisuus
useilla ulottuvuuksilla

Napauta kääntääksesi ↺

Algoritmissen
vinouman
lieventäminen

Napauta kääntääksesi ↺

Eettinen harkinta ja
hallinta

Napauta kääntääksesi ↺

Ihmislähtöinen
suunnittelu

Napauta kääntääksesi ↺

Oikeudenmukaisuus
nousevissa
teknologioissa

Napauta kääntääksesi ↺

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja tulkittavuus

Kun tekoälyjärjestelmät kehittyvät yhä monimutkaisemmiksi, tarve läpinäkyvyydelle ja selitettävyydelle kasvaa. Tulevaisuudessa voidaan odottaa uusien menetelmien ja standardien kehittämistä, jotka auttavat selittämään tekoälyn päätöksiä ja tekemään ne ymmärrettäviksi eri sidosryhmille.

Vastahyökkäykset ja puolustus

Napauta kääntääksesi ↺

Oikeudenmukaisuus useilla ulottuvuuksilla

Napauta kääntääksesi ↺

Algoritmisen vinouman lieventäminen

Napauta kääntääksesi ↺

Eettinen harkinta ja hallinta

Napauta kääntääksesi ↺

Ihmislähtöinen suunnittelu

Napauta kääntääksesi ↺

Oikeudenmukaisuus nousevissa teknologioissa

Napauta kääntääksesi ↺

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyyys ja
tulkittavuus

Vastahyökkäykset ja puolustus

Vihamieliset hyökkäykset, joissa tekoälyä pyritään huijaamaan hienovaraisilla syötedatan muutoksilla, uhkaavat sekä järjestelmien oikeudenmukaisuutta että luotettavuutta. Tulevaisuuden tutkimus voi keskittyä kehittämään tehokkaita puolustuskeinoja tällaisia hyökkäyksiä vastaan tekoälyn luotettavuuden

Oikeudenmukaisuus
useilla ulottuvuuksilla

Napauta kääntääksesi ↺

Algoritmissen
vinouman
lieventäminen

Napauta kääntääksesi ↺

Eettinen harkinta ja
hallinta

Napauta kääntääksesi ↺

Ihmislähtöinen
suunnittelu

Napauta kääntääksesi ↺

Oikeudenmukaisuus
nousevissa
teknologioissa

Napauta kääntääksesi ↺

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja
tulkittavuus

Vastahyökkäykset ja
puolustus

Oikeudenmukaisuus useilla ulottuvuuksilla

Tekoälyn oikeudenmukaisuus ei rajoitu pelkästään demografisiin tekijöihin, vaan siihen voivat vaikuttaa myös esimerkiksi sosioekonominen asema, vammaisuus tai kielitaito. Tulevaisuuden kehityksessä saatetaan painottaa moniulotteisten oikeudenmukaisuusmittarien ja -algoritmien luomista, jotka huomioivat risteävät syrjinnän muodot.

Algoritmisen
vinouman
lieventäminen

Napauta kääntääksesi ↺

Eettinen harkinta ja
hallinta

Napauta kääntääksesi ↺

Ihmislähtöinen
suunnittelu

Napauta kääntääksesi ↺

Oikeudenmukaisuus
nousevissa
teknologioissa

Napauta kääntääksesi ↺

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja
tulkittavuus

Vastahyökkäykset ja
puolustus

Oikeudenmukaisuus
useilla ulottuvuuksilla

Algorithmic bias mitigation

Algoritmisen vinouman vähentämiseen tähtäävät tekniikat tulevat pysymään keskiössä. Tutkimus keskittyy kehittämään oikeudenmukaisempia malleja, jotka vähentävät puolueellisuutta eri päätöksenteon alueilla.

Eettinen harkinta ja
hallinta

Napauta kääntääksesi ↺

Ihmislähtöinen
suunnittelu

Napauta kääntääksesi ↺

Oikeudenmukaisuus
nousevissa
teknologioissa

Napauta kääntääksesi ↺

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja
tulkittavuus

Vastahyökkäykset ja
puolustus

Oikeudenmukaisuus
useilla ulottuvuuksilla

Algoritmisen
vinouman
lieventäminen

Eettinen harkinta ja hallinta

Tekoälyn eettiset kysymykset – kuten oikeudenmukaisuus ja syrjimättömyys – säilyvät tärkeinä keskustelunaiheina. Kehityssuunta voi johtaa sääntelyohjeiden ja -standardien luomiseen, jotta tekoälyjärjestelmät ovat entistä vastuullisempia ja läpinäkyvämpiä.

Ihmislähtöinen
suunnittelu

Napauta kääntääksesi 

Oikeudenmukaisuus
nousevissa
teknologioissa

Napauta kääntääksesi 

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja
tulkittavuus

Vastahyökkäykset ja
puolustus

Oikeudenmukaisuus
useilla ulottuvuuksilla

Algoritmisen
vinouman
lieventäminen

Eettinen harkinta ja
hallinta

Ihmislähtöinen suunnittelu

Tulevaisuuden tekoäly voi korostaa ihmislähtöistä suunnittelua, jossa käyttäjien palaute ja tarpeet huomioidaan järjestelmien oikeudenmukaisuuden ja käytettävyyden parantamiseksi. Tämä voisi tarkoittaa tekoälyn yhteiskehittämistä eri sidosryhmien kanssa.

Oikeudenmukaisuus
nousevissa
teknologioissa

Napauta kääntääksesi ↺

Tulevaisuuden suuntaukset ja kasvavat ongelmat

Selitettävyys ja
tulkittavuus

Vastahyökkäykset ja
puolustus

Oikeudenmukaisuus
useilla ulottuvuuksilla

Algoritmissen
vinouman
lieventäminen

Eettinen harkinta ja
hallinta

Ihmislähtöinen
suunnittelu

Oikeudenmukaisuus nousevissa teknologioissa

Kun tekoälyä sovelletaan yhä useammilla alueilla –
kuten itseajavat autot, terveydenhuollon
järjestelmät ja kasvojentunnistus –
oikeudenmukaisuuden ja vinoumien
huomioiminen muuttuu entistä tärkeämmäksi.
Tulevaisuudessa saatetaan kehittää uusia
oikeudenmukaisuuden periaatteita vastaamaan
näiden teknologioiden erityispiirteitä.

JOHTOPÄÄTÖKSET



IMAGE SOURCE | Generated by DALL-E

JOHTOPÄÄTÖKSET

- 1. Monimutkaisuuden ymmärtäminen:** Tekoälyn oikeudenmukaisuus ja vinoumat ovat moniulotteisia haasteita, jotka edellyttävät kokonaisvaltaisia strategioita ja jatkuvaa huomiota.
- 2. Toimintastrategiat:** Vinoumien tunnistamiseen ja lieventämiseen tekoälyjärjestelmissä on olemassa useita keinoja – aina monipuolisesta edustuksesta datan keruussa algoritmisten harhojen vähentämiseen.
- 3. Tuleviin haasteisiin varautuminen:** Tekoälyn kehittyessä on tärkeää pysyä ajan tasalla uusista ilmiöistä ja haasteista, kuten häiritsevistä hyökkäyksistä, moniulotteisesta oikeudenmukaisuudesta ja eettisistä kysymyksistä.
- 4. Kohti oikeudenmukaisempaa tulevaisuutta:** Asettamalla ihmiskeskeinen suunnittelu, läpinäkyvyys ja vastuu etusijalle voimme rakentaa tekoälyjärjestelmiä, jotka edistävät oikeudenmukaisuutta, osallisuutta ja luottamusta.

KIITOS



Universitat
de les Illes Balears

ISQe
ENGAGING PEOPLE



**Euroopan unionin
osarahoittama**

Euroopan unionin rahoittama. Esitetyt näkemykset ja mielipiteet ovat ainoastaan tämän tekstin laatijoiden näkemyksiä eivätkä välttämättä vastaa Euroopan unionin tai Euroopan koulutuksen ja kulttuurin toimeenpanovirasto (EACEA) kantaa. Euroopan unioni ja EACEA eivät ole vastuussa niistä.

2022-1-ES01-KA220-HED-000085257