



Eettisen tekoälyn mikrokurssi

OPAS

**Osaamiskokonaisuus 4 |
Läpinäkyvyys**

Hankkeen numero:
2022-1-ES01-KA220-HED-000085257

Miten tätä Opasta käytetään?

Tämä dokumentti on interaktiivinen.
Dokumentissa on linkkejä lisätietoihin.



Painike, joka vie sinut dokumentin alkuun. Tämä kuvake näkyy sivujen oikeassa yläkulmassa.



Aina kun näet tämän nuolen, se tarkoittaa, että kyseessä on **interaktiivinen väriteksti**, jota voit napsauttaa ja johon on liitetty ulkoinen linkki.

VASTUUVAPAAUSLAUSEKE: Huomaa, että emme voi taata ulkoisen sisällön, kuten videoiden, jatkuvaa saatavuutta, sillä niiden tekijät tai isäntäalustat voivat muuttaa tai poistaa niitä.

Sisältö

Klikkaa valikkoa

01. Johdanto

**02. Läpinäkyvyyden merkitys
tekoälyjärjestelmissä**

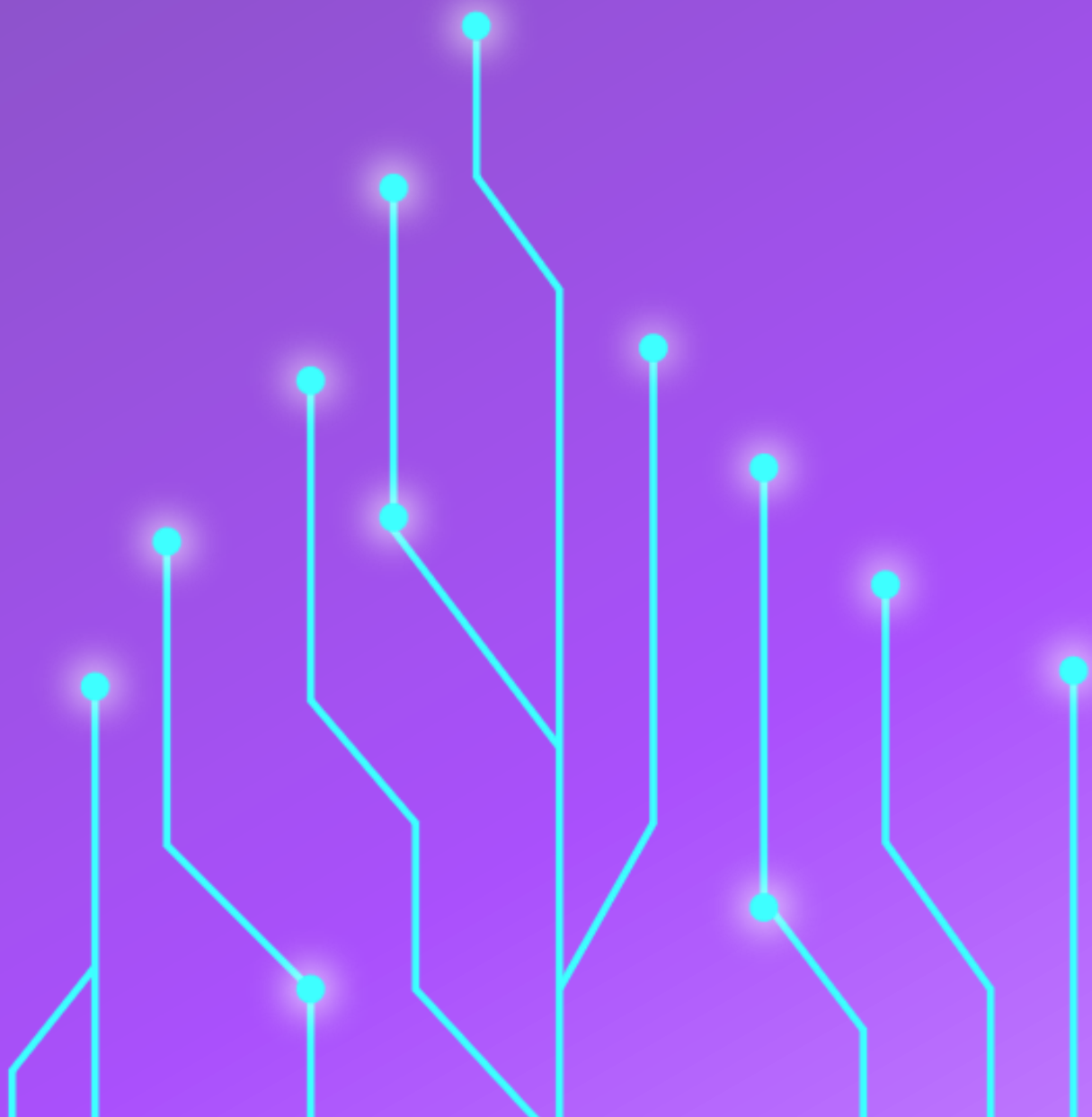
**03. Läpinäkyvyyden ja algoritmisen vinouman
välinen suhde**

**04. Strategiat tekoälyjärjestelmien
läpinäkyvyyden edistämiseksi**

05. Yhteenveto

01. Johdanto

Osaamiskokonaisuus 4 | Läpinäkyvyys





01. Johdanto

Tässä osaamiskokonaisuudessa oppijat saavat tietoa tekoälyjärjestelmien läpinäkyvyyden merkityksestä keskittyen peruskäsitteiden ymmärtämiseen, läpinäkyvyyden ja algoritmisen vinouman väliseen suhteeseen sekä strategioiden merkitykseen sen varmistamisessa, että tekoälyjärjestelmät ovat ymmärrettäviä, selitettävissä ja sidosryhmien saatavilla, ja ymmärtäen reaali maailman seuraukset ja arvostaen sitä, miten tärkeitä tulkittavissa olevat mallit, selkeä dokumentaatio ja tehokas viestintä voivat olla läpinäkyvyyskulttuurin edistämisessä ja algoritmisen vinouman vähentämisessä.

Kurssin osaamistavoitteet ovat seuraavat:

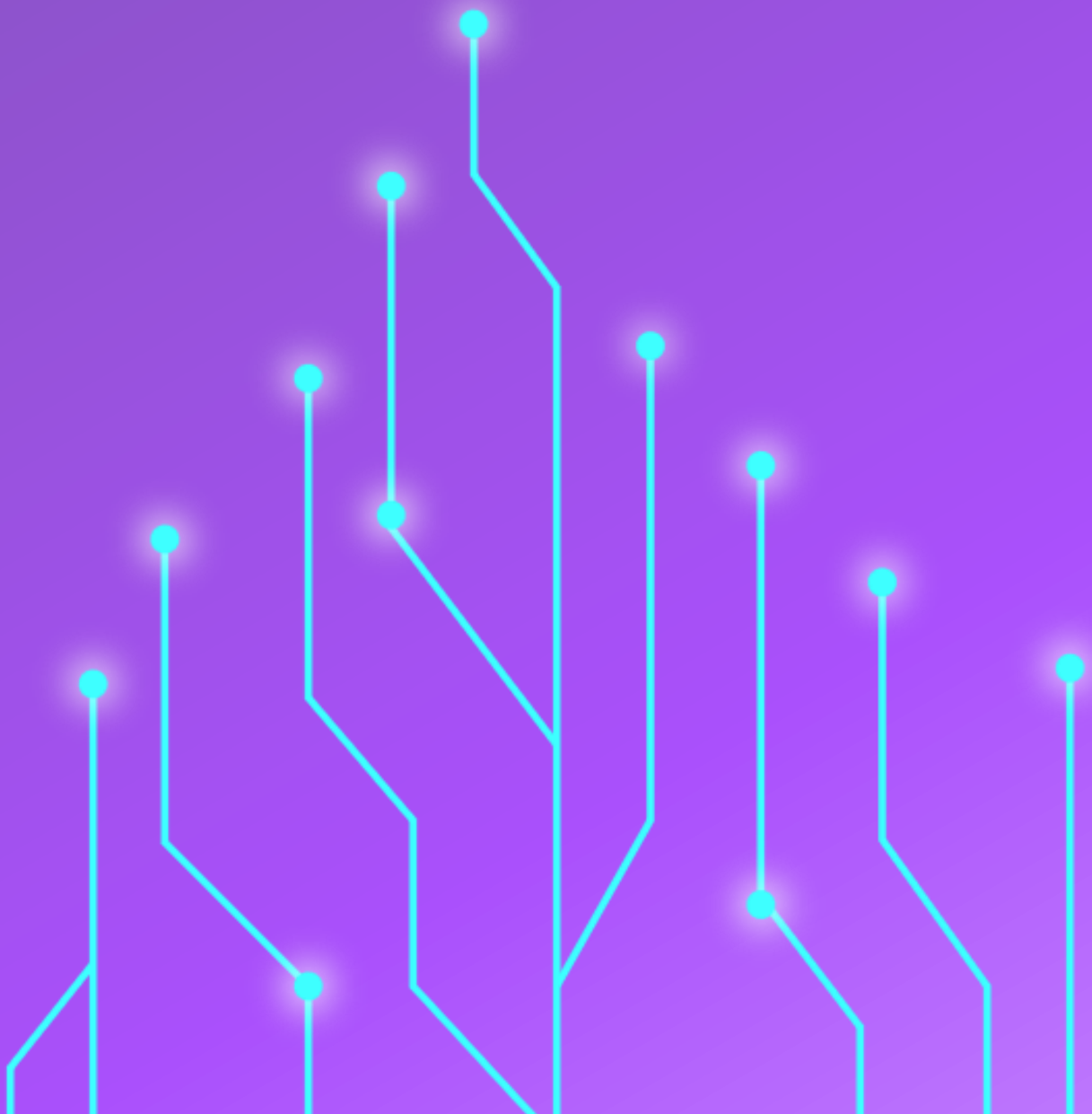
- **Tekoälyjärjestelmien läpinäkyvyyden merkitys** ja sen merkitys sen varmistamisessa, että tekoälyjärjestelmät ovat ymmärrettäviä, selitettävissä ja sidosryhmien saatavilla. Tunnistamme läpinäkyvien tekoälyjärjestelmien hyödyt ja arvostamme niiden merkitystä luottamuksen rakentamisessa ja sidosryhmien ymmärryksen mahdollistamisessa. Esimerkkinä: Syövän havaitsemiseen suunniteltu tekoälymalli voi uhata ihmishenkiä, vaikka se olisikin vain 1 % väärässä. Tällaisissa tapauksissa tekoälyn ja ihmisten on tehtävä yhteistyötä, ja tehtävä helpottuu huomattavasti, kun tekoälymalli voi selittää, miten se on päätenyt tiettyyn päätökseen. Avoimuus tekee tekoälystä tiimipelaajan.

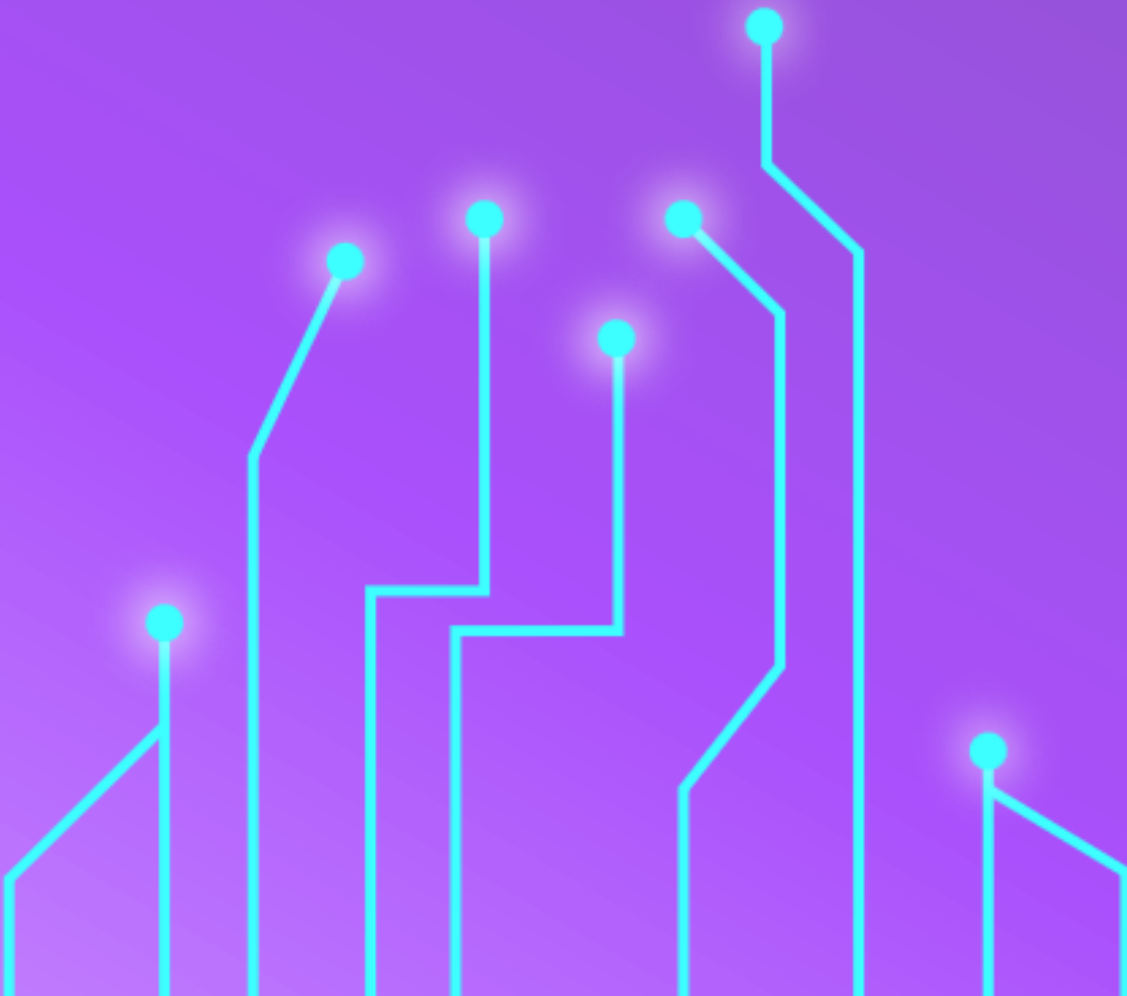


- **Läpinäkyvyyden ja algoritmisen vinouman välinen suhde**, jotta löydettäisiin yhteys läpinäkyvyyden ja algoritmisen vinouman välille, tunnistettaisiin läpinäkymättömyyden vaarat ja todettaisiin, miten läpinäkyvyyden lisääminen voi auttaa tunnistamaan, ehkäisemään ja lieventämään tekoälyjärjestelmien vinoutuneita tuloksia. Tunnistamme läpinäkyvyyden merkityksen algoritmisen vinouman käsittelyssä ja lieventämisessä. Esimerkkinä voidaan todeta, että tekoälyalgoritmit ovat usein vaikeaselkoisia siinä mielessä, että selitykset eivät ole kaikkien sidosryhmien saatavilla. Läpinäkymättömyydellä voi olla erilaisia lähteitä. Joskus instituutiot tai yritykset eivät kerro, milloin ne luottavat tekoälyjärjestelmiin tai miten nämä järjestelmät toimivat.
- **Strategiat tekoälyjärjestelmien läpinäkyvyyden edistämiseksi**, kuten selitettävien mallien käyttö, selkeä dokumentointi ja tekoälysovellusten päätöksentekoprosessista viestiminen. Selitämme, miten tärkeitä nämä strategiat ovat läpinäkyvyyden kulttuurin edistämisessä ja algoritmisen vinouman lieventämisessä. Tekoäly voi esimerkiksi vaikuttaa eri sidosryhmiin, kuten käyttäjiin, asiakkaisiin, työntekijöihin, johtajiin, sääntelyviranomaisiin tai yhteiskuntaan. Avoimuuden ja vastuullisuuden varmistamiseksi tekoälyn sidosryhmät on saatava mukaan ja vaalittava koko tekoälyn elinkaaren ajan.

02. Läpinäkyvyyden merkitys tekoälyjärjestelmissä

Osaamiskokonaisuus 4 | Läpinäkyvyys





02. Läpinäkyvyyden merkitys tekoälyjärjestelmissä

Läpinäkyvyys on yksi tekoälyjärjestelmien kehittämisen ja käyttöönoton peruseriaatteista.

Tekoälyn läpinäkyvyydellä tarkoitetaan tekoälyjärjestelmien avoimuutta ja saavutettavuutta, jotta sidosryhmät voivat ymmärtää, miten algoritmit toimivat, miksi tietyt päätökset tehdään ja mitkä tekijät vaikuttavat niiden tuloksiin. Läpinäkyvyys käsittää useita eri näkökohtia, kuten tiedon saatavuuden tietolähteistä, algoritmisista malleista, päätöksentekoprosesseista ja mahdollisista vinoumista. Läpinäkyvät tekoälyjärjestelmät antavat sidosryhmille, kuten käyttäjille, kehittäjille, poliittisille päättäjille ja suurelle yleisölle, mahdollisuuden tarkastella ja kyseenalaistaa algoritmien tuloksia, mikä edistää luottamusta ja vastuullisuutta.

Yksi läpinäkyvien tekoälyjärjestelmien tärkeimmistä eduista on niiden ymmärrettävyys. Kun tekoälyalgoritmit ovat läpinäkyviä, sidosryhmät voivat ymmärtää, miten ne toimivat ja miksi ne tuottavat tiettyjä tuloksia. Tämän ymmärryksen ansiosta käyttäjät voivat luottaa tekoälyteknologioihin ja tehdä tietoon perustuvia päätöksiä niiden käytöstä. Esimerkiksi lääketieteellisen diagnoosin tekoälymallin yhteydessä läpinäkyvyys antaa terveydenhuollon ammattilaisille mahdollisuuden ymmärtää, miten malli on päätenyt diagnoosiinsa, jolloin he voivat validoida sen tarkkuuden ja luotettavuuden ennen hoitopäätösten tekemistä.

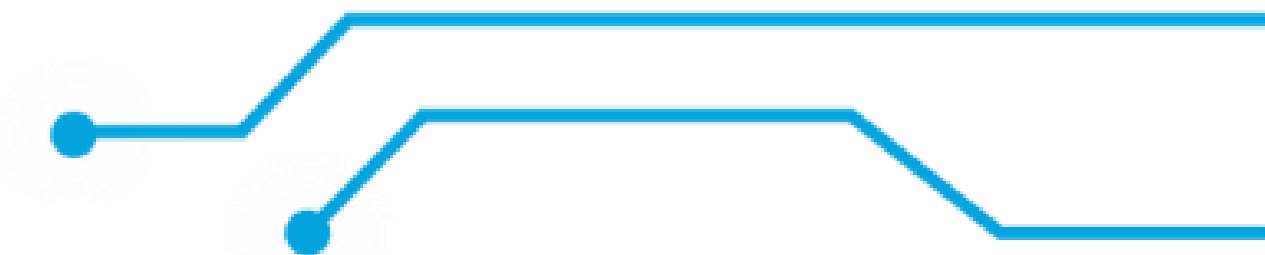


Lisäksi läpinäkyvyys helpottaa selitettävyyttä, mikä on olennaista sen varmistamiseksi, että tekoälyjärjestelmät pystyvät antamaan tulkinnanvaraisia selityksiä päätöksilleen ja toimilleen. Selitettävissä oleva tekoäly antaa sidosryhmille mahdollisuuden ymmärtää algoritmisen lopputuloksen taustalla olevia perusteluja sekä tunnistaa ja korjata vinoumat tai virheet. Esimerkiksi lainojen hyväksymiseen tarkoitetun tekoälyjärjestelmän tapauksessa läpinäkyvyys ja selitettävyys antavat lainanhakijoille mahdollisuuden ymmärtää, miksi heidän hakemuksensa hyväksyttiin tai hylättiin, ja tarjota tietoa päätöksentekoprosessista ja muutoksenhakukeinoja, jos he katsovat, että päätös oli vinoutunut tai epäoikeudenmukainen.

Lisäksi läpinäkyvyys parantaa tekoälyjärjestelmien saavutettavuutta, mikä tekee niistä osallistavampia ja oikeudenmukaisempia. Kun tekoälyalgoritmit ovat läpinäkyviä, erilaisista taustoista ja asiantuntemustasoista tulevat sidosryhmät voivat saada ja tulkita tietoa niiden toiminnasta ja tuloksista. Saavutettavuus varmistaa, että tekoälyteknologiat ovat paitsi ymmärrettäviä myös käyttökelpoisia monenlaisille käyttäjille, myös niille, joilla on vammoja tai rajalliset tekniset tiedot. Esimerkiksi kehitettäessä tekoälyllä toimivia saavutettavuustyökaluja vammaisille, avoimuus antaa käyttäjille mahdollisuuden ymmärtää, miten työkalut toimivat ja miten he voivat hyötyä niistä.

Havainnollistava esimerkki tekoälyjärjestelmien läpinäkyvyyden merkityksestä on tekoälymallien kehittäminen lääketieteelliseen diagnostiikkaan, kuten syövän havaitsemiseen. Vaikka tekoälymalli olisikin erittäin tarkka ja sen onnistumisprosentti olisi 99 prosenttia, jäljelle jäävällä yhden prosentin virhemarginaalilla voi olla hengenvaarallisia seurauksia potilaille. Tällaisissa kriittisissä skenaarioissa avoimuus on olennaisen tärkeää sen varmistamiseksi, että terveydenhuollon ammattilaiset ymmärtävät, miten tekoälymalli on päätenyt diagnoosiinsa, ja että he voivat tarkistaa sen tarkkuuden ennen hoitopäätösten tekemistä. Kun tekoälymalli selittää avoimesti päätöksentekoprosessinsa, siitä tulee arvokas työkalu terveydenhuollon ammattilaisille, mikä parantaa heidän kykyään diagnosoida ja hoitaa potilaita tehokkaasti.

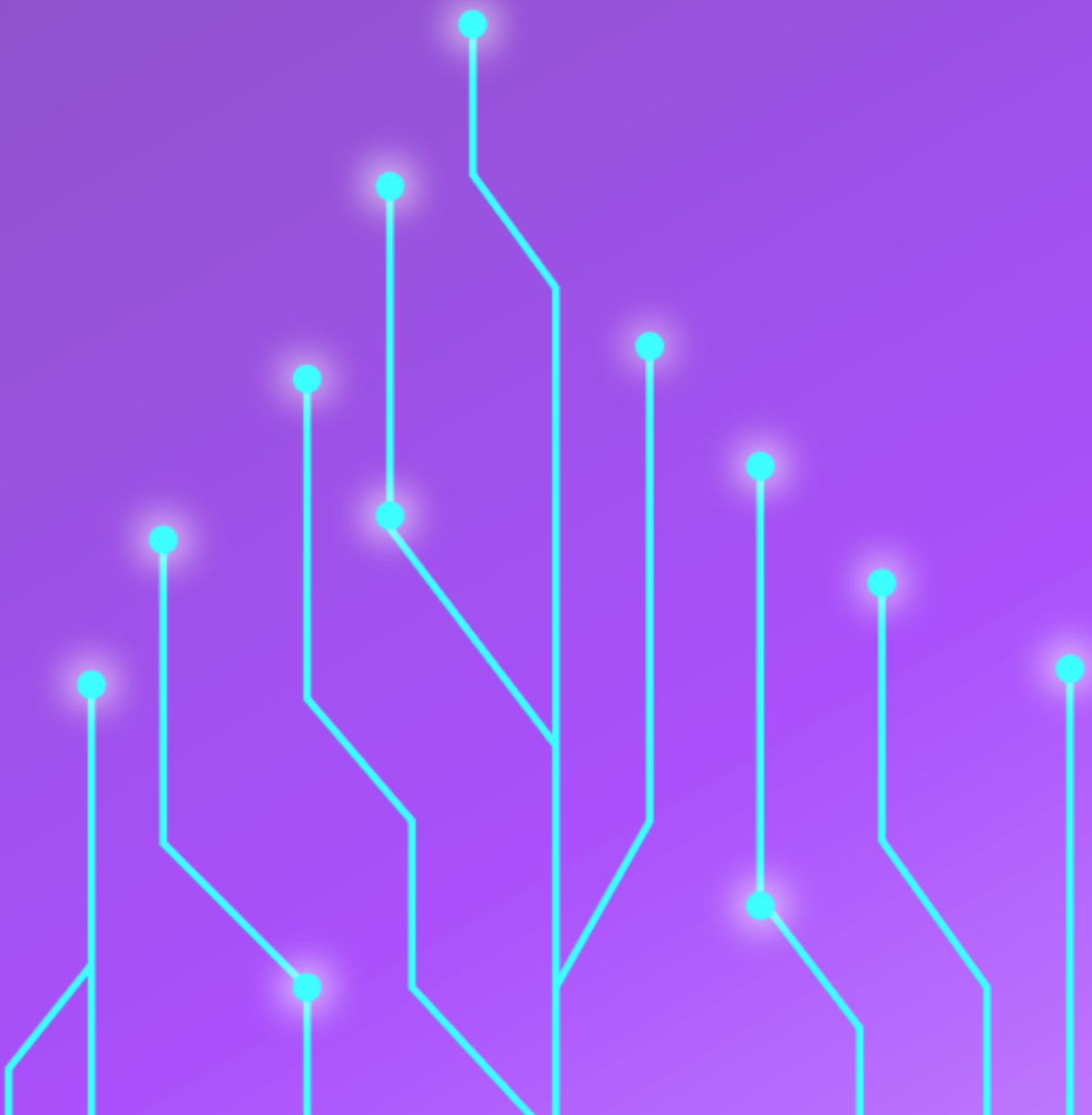
Kuten olemme jo lukeneet tästä oppaasta, algoritmisella vinoumalla tarkoitetaan tekoälyalgoritmeissa esiintyviä järjestelmällisiä virheitä tai epäoikeudenmukaisuutta, jotka johtavat syrjiviin tuloksiin tiettyjen yksilöiden tai ryhmien kannalta. Nämä vinoumat voivat johtua eri lähteistä, kuten vääristyneestä koulutusdatasta, virheellisestä algoritmisesta suunnittelusta tai järjestelmään koodatuista inhimillisistä vinoumista. Algoritmisen vinouman seuraukset voivat olla kauaskantoisia, sillä ne voivat ylläpitää eriarvoisuutta, vahvistaa stereotypioita ja heikentää luottamusta tekoälyjärjestelmiin.





03. Läpinäkyvyyden ja algoritmisen vinouman välinen suhde

Osaamiskokonaisuus 4 | Läpinäkyvyys





03. Läpinäkyvyyden ja algoritmisen vinouman välinen suhde

Läpinäkymättömyys eli avoimuuden puute pahentaa algoritmiseen vinoumaan liittyviä riskejä.

Tekoälyalgoritmit ovat usein vaikeaselkoisia, mikä tarkoittaa, että niiden päätösten ja toimien selitykset eivät ole helposti kaikkien sidosryhmien saatavilla. Tämä vaikeaselkoisuus voi johtua monista eri syistä, kuten institutionaalisesta salassapidosta, yritysten luottamuksellisuudesta tai teknisestä monimutkaisuudesta. Kun sidosryhmät eivät saa tietoa tekoälyjärjestelmistä, ne eivät pysty arvioimaan algoritmien tulosten oikeudenmukaisuutta, luotettavuutta tai eettisiä vaikutuksia, mikä johtaa vastuullisuuden puuttumiseen ja mahdolliseen vahinkoon.

Läpinäkyvyys on ratkaiseva vastalääke tekoälyjärjestelmien läpinäkymättömyydelle, sillä se antaa sidosryhmille mahdollisuuden tarkastella ja kyseenalaistaa algoritmien päätöksiä ja vähentää näin algoritmisen vinouman riskejä. Lisäämällä avoimuutta tekoälyn kehittäjät ja ammattilaiset voivat tarjota sidosryhmille tietoa siitä, miten tekoälyjärjestelmät toimivat, miksi tietyt päätökset tehdään ja mitkä tekijät vaikuttavat niiden tuloksiin. Läpinäkyvät tekoälyjärjestelmät antavat sidosryhmille mahdollisuuden tunnistaa ja puuttua vinoumiin, validoida algoritmien tarkkuus ja asettaa kehittäjät vastuuseen tekoälyteknologian eettisestä ja oikeudenmukaisesta käytöstä.



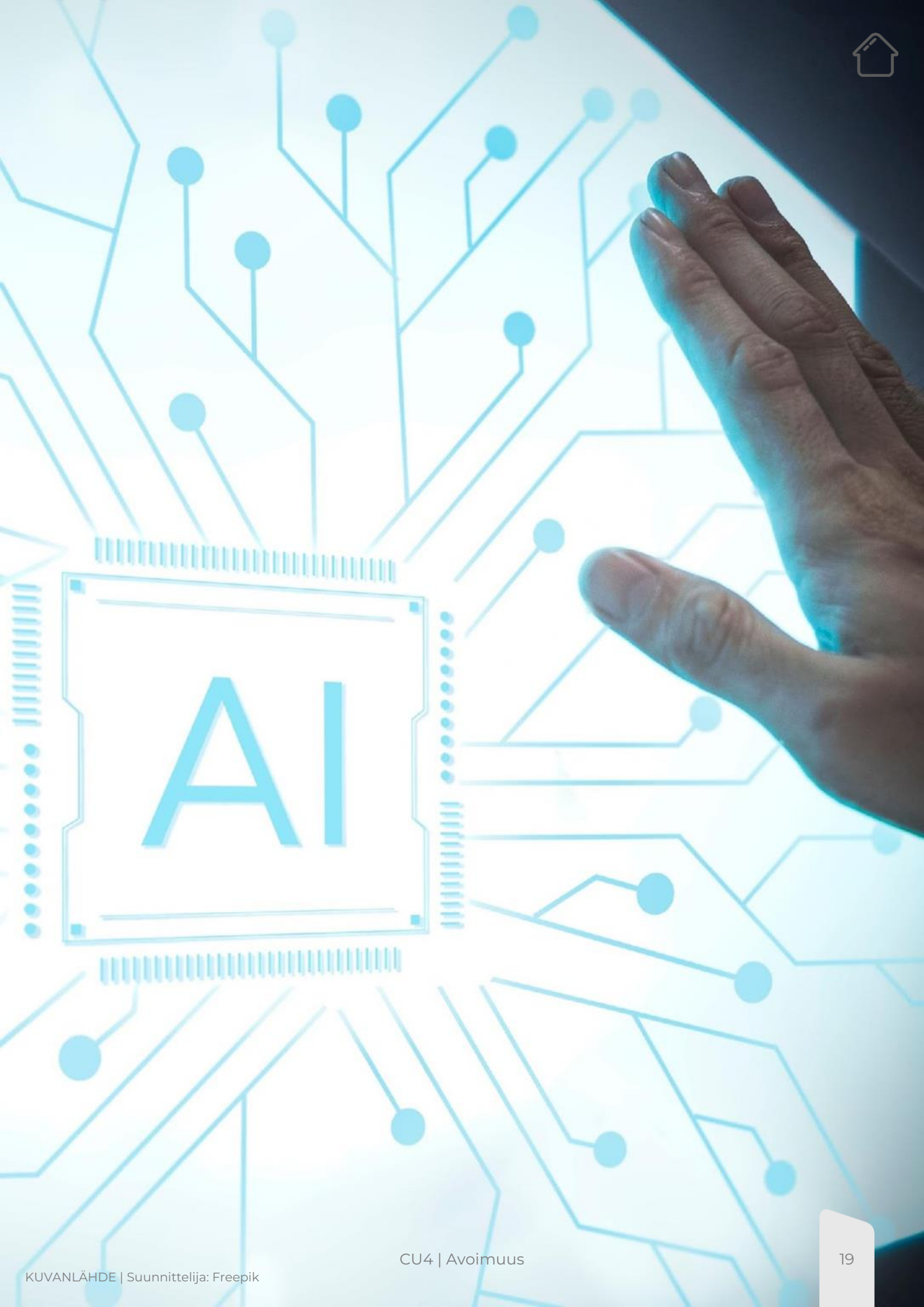
Yksi tärkeimmistä eduista, joita avoimuus tuo algoritmiseen vinoumaan puuttumisessa, on kyky havaita ja lieventää vinoutuneita tuloksia. Kun tekoälyalgoritmit ovat läpinäkyviä, sidosryhmät voivat tutkia päätöksentekoprosessia ja tunnistaa tapaukset, joissa voi esiintyä vinoumaa. Esimerkiksi tekoälyä palkkaavan tekoälyjärjestelmän yhteydessä läpinäkyvyys antaa sidosryhmille mahdollisuuden arvioida, syrjikkö järjestelmä epäoikeudenmukaisesti tiettyjä väestöryhmiä valintaprosessissa. Tunnistamalla vinoutuneet tulokset sidosryhmät voivat ryhtyä korjaaviin toimiin algoritmisen vinouman aiheuttamien haittojen lieventämiseksi ja oikeudenmukaisuuden ja tasapuolisuuden edistämiseksi.

Lisäksi avoimuus helpottaa vastuullisuutta ja luottamusta tekoälyjärjestelmiin. Kun sidosryhmät saavat tietoa tekoälyalgoritmeista, ne voivat vaatia kehittäjiä ja ammattilaisia vastuuseen tekoälyteknologian eettisestä ja oikeudenmukaisesta käytöstä. Läpinäkyvät tekoälyjärjestelmät luovat luottamusta käyttäjien, sääntelyviranomaisten ja suuren yleisön keskuudessa, mikä edistää luottamusta algoritmien tulosten luotettavuuteen ja oikeudenmukaisuuteen. Kun tekoälyjärjestelmiä otetaan käyttöön esimerkiksi rikosoikeudessa tai terveydenhuollossa, avoimuus antaa sidosryhmille mahdollisuuden ymmärtää, miten päätökset tehdään, ja varmistaa, että päätökset ovat eettisten periaatteiden ja oikeudellisten normien mukaisia.

Läpinäkyvyydellä on keskeinen rooli tekoälyjärjestelmien algoritmisen vinouman käsittelyssä ja lieventämisessä.

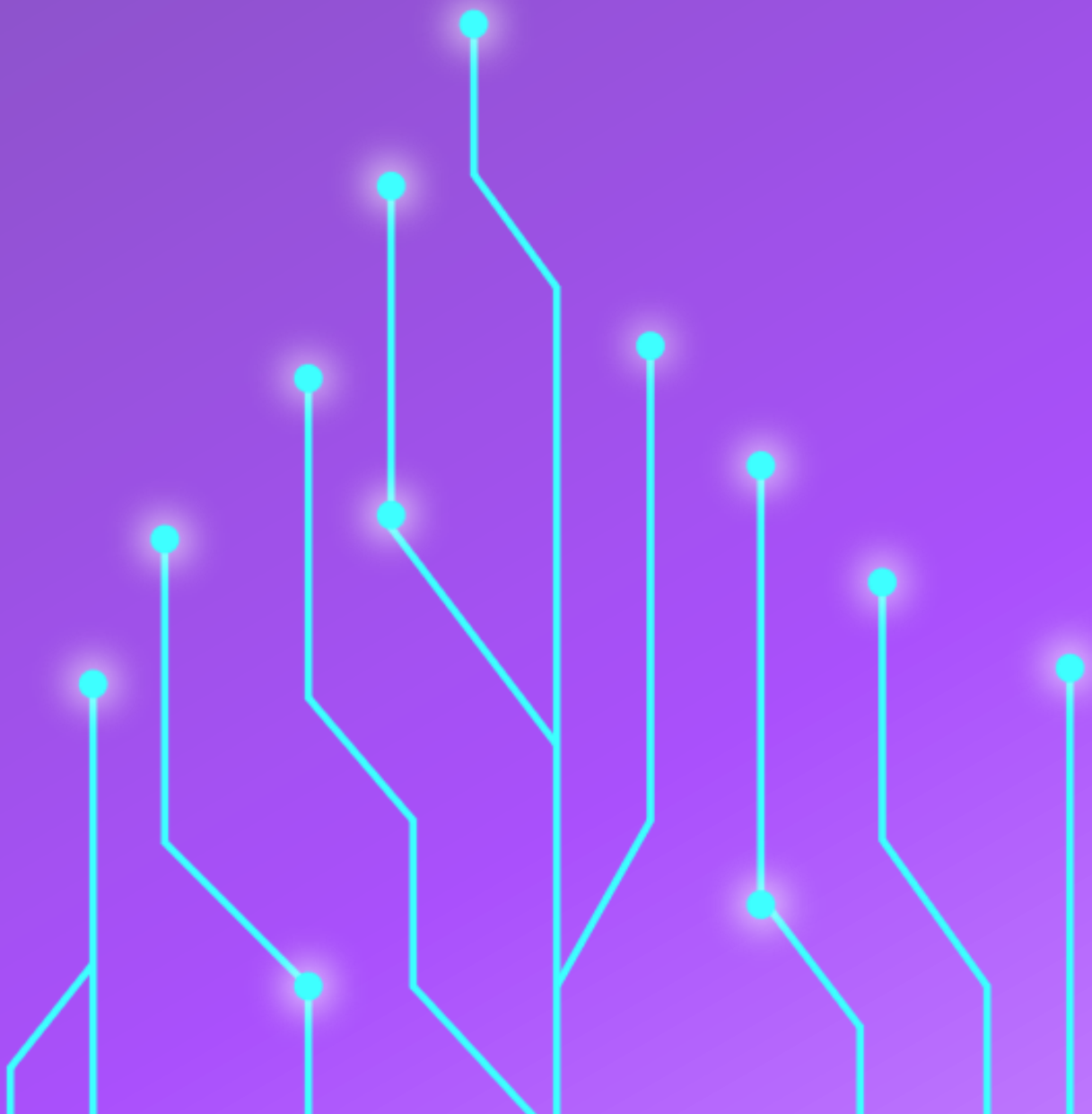
Läpinäkyvyyttä lisäämällä sidosryhmät voivat havaita ja lieventää vinoutuneita tuloksia, edistää vastuullisuutta ja rakentaa luottamusta tekoälyteknologioihin. Kun tekoäly kehittyy ja integroituu yhä enemmän yhteiskunnan eri osa-alueisiin, avoimuus on edelleen olennaisen tärkeää sen varmistamiseksi, että tekoälyjärjestelmiä kehitetään ja otetaan käyttöön eettisiä normeja noudattaen ja oikeudenmukaisuutta ja tasapuolisuutta edistäen. Läpinäkyvyyden ja algoritmisen vinouman välisen suhteen kattavan ymmärtämisen kautta oppijat voivat edistää tekoälyteknologioiden vastuullista ja eettistä kehittämistä ja luoda siten oikeudenmukaisemman ja osallistavamman tulevaisuuden.





04. Strategiat tekoälyjärjestelmien läpinäkyvyyden edistämiseksi

Osaamiskokonaisuus 4 | Läpinäkyvyys





04. Strategiat tekoälyjärjestelmien läpinäkyvyyden edistämiseksi

Tekoälyjärjestelmien läpinäkyvyyden edistämiseksi on olemassa erilaisia strategioita, kuten selitettävien mallien käyttö, selkeä dokumentointi ja päätöksentekoprosesseista viestiminen.

> Selitettävät mallit

Selitettävät mallit ovat keskeinen strategia tekoälyjärjestelmien avoimuuden edistämiseksi. Ne ovat koneoppimismalleja, jotka tuottavat tuloksia, joita ihmisten on helppo ymmärtää ja tulkita. Seuraavassa on muutamia esimerkkejä:

- **Lineaarinen regressio:** Lineaarinen regressio on yksinkertainen ja selitettävissä oleva malli, jota käytetään yleisesti numeeristen tulosten ennustamiseen. Se toimii sovittamalla datapisteisiin suora viiva, jolloin syötemuuttujien ja tuloksen välinen suhde on helppo tulkita.
- **Päätöspuut:** Päätöspuut ovat hierarkisia malleja, jotka tekevät päätöksiä sarjan "jos-jos" -lausekkeiden perusteella. Jokainen puun oksa edustaa datan ominaisuuteen perustuvaa päätöstä, joten mallin ennusteiden taustalla olevaa logiikkaa on helppo seurata.



- **Logistinen regressio:** Logistinen regressio on tilastollinen malli, jota käytetään binääriluokittelutehtävissä. Se laskee todennäköisyyden, jolla instanssi kuuluu tiettyyn luokkaan, sen syöttöominaisuuksien perusteella, mikä tekee siitä selitettävän ja helposti ymmärrettävän.
- **Sääntöpohjaiset mallit:** Sääntöpohjaiset mallit, kuten luokitus- ja regressiopuut (CART) tai päätöksentekosäännöt, muuttavat syötetiedot suoraan päätöksentekosäännöiksi. Näitä sääntöjä on helppo tulkita, ja ne voivat antaa tietoa siitä, miten malli tekee ennusteita.
- **Yleistetyt additiiviset mallit (GAM):** GAM-mallit ovat joustavia malleja, joilla voidaan kuvata monimutkaisia suhteita syötetietojen ja tulosten välillä säilyttäen samalla selitettävyys. Niissä käytetään tasaisia funktioita kuvaamaan kunkin syötemuuttujan ja tuloksen välistä suhdetta, mikä mahdollistaa mallin ennusteiden helpon tulkinnan.

Toisin kuin monimutkaiset mustan laatikon mallit, selitettävät mallit antavat sidosryhmille mahdollisuuden ymmärtää, miten tekoälyalgoritmit tekevät päätöksiä ja mitkä tekijät vaikuttavat niiden tuloksiin. Käyttämällä tulkittavia malleja tekoälyn kehittäjät voivat lisätä läpinäkyvyyttä ja vastuullisuutta, jolloin sidosryhmät voivat validoida algoritmien tulokset ja tunnistaa mahdolliset vinoumat tai virheet.

Esimerkiksi luottopisteytysjärjestelmän tekoälyssä selitettävien mallien käyttö antaa sidosryhmille mahdollisuuden ymmärtää luottopäätöksiin vaikuttavat tekijät, kuten tulot, luottohistoria ja velkaantuneisuus, ja edistää siten läpinäkyvyyttä ja oikeudenmukaisuutta luotonantokäytännöissä.

> **Selkeä dokumentaatio**

Selkeä dokumentointi on toinen keskeinen strategia tekoälyjärjestelmien avoimuuden edistämiseksi. Dokumentointi tarjoaa sidosryhmille tietoa tekoälyalgoritmien suunnittelusta, kehittämisestä ja käyttöönotosta, mukaan lukien tietolähteet, esikäsittelytekniikat, malliarkkitehtuurit ja arviointimittarit.

Dokumentoimalla tekoälyjärjestelmät kattavasti kehittäjät voivat lisätä läpinäkyvyyttä ja vastuullisuutta, jolloin sidosryhmät voivat ymmärtää tekoälyteknologioiden taustalla olevat prosessit ja oletukset. Esimerkiksi teollisuuslaitteiden ennakoivan kunnossapidon tekoälyjärjestelmää kehitettäessä sidosryhmät voivat selkeän dokumentoinnin avulla arvioida ennakoivien mallien luotettavuutta ja tarkkuutta, ymmärtää kunnossapitosuosituksia ja tarkistaa turvallisuusstandardien noudattamisen.

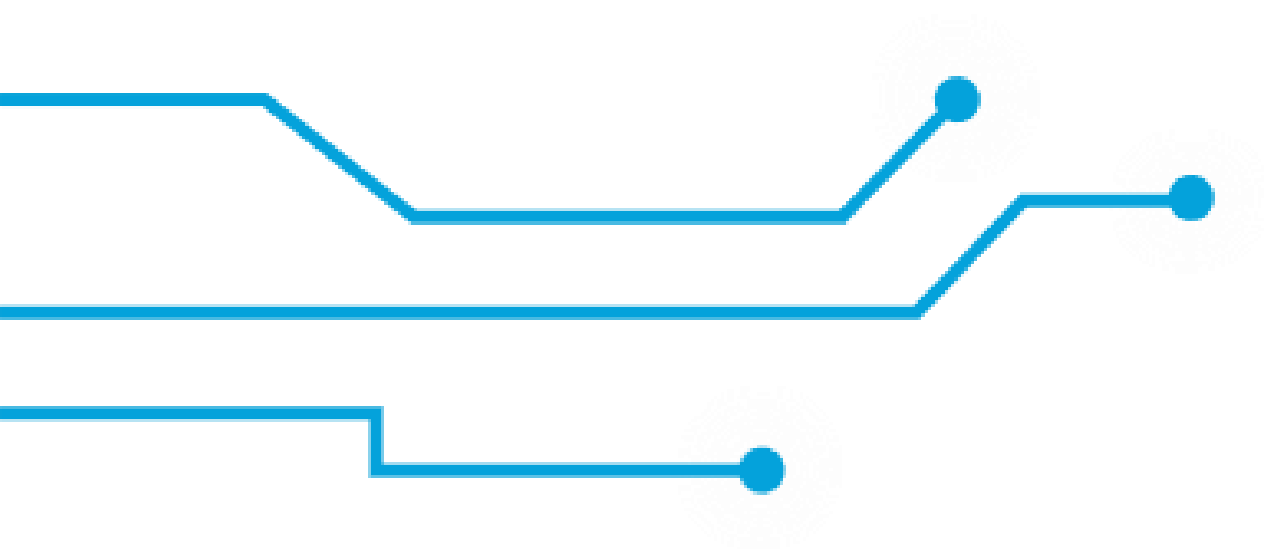




> Tehokas päätöksentekoviestintä

Päätöksentekoprosesseista viestiminen on olennaista tekoälyjärjestelmien avoimuuden edistämiseksi. Viestinnällä varmistetaan, että sidosryhmät saavat tietoa tekoälyalgoritmien päätösten perusteluista, logiikasta ja vaikutuksista.

Ilmoittamalla päätöksentekoprosesseista selkeästi ja avoimesti tekoälyn kehittäjät voivat rakentaa luottamusta käyttäjien, sääntelyviranomaisten ja suuren yleisön keskuudessa. Esimerkiksi terveydenhuollon diagnosointiin käytettävien tekoälyjärjestelmien käyttöönotossa tehokkaalla viestinnällä varmistetaan, että terveydenhuollon ammattilaiset ja potilaat ymmärtävät, miten diagnoosipäätökset tehdään, jolloin he voivat luottaa tekoälyn tuottamien diagnoosien tarkkuuteen ja tarkistaa sen.



➤ **Tekoälyn sidosryhmien sitouttaminen ja vaikutusmahdollisuuksien lisääminen**

Tekoälyn sidosryhmien osallistuminen ja vaikutusmahdollisuuksien lisääminen koko tekoälyn elinkaaren ajan on ratkaisevan tärkeää avoimuuden ja vastuullisuuden varmistamiseksi. Sidosryhmien sitouttaminen koskee käyttäjiä, asiakkaita, työntekijöitä, johtajia, sääntelyviranomaisia ja yhteiskuntaa tekoälyjärjestelmien suunnittelussa, kehittämisessä, käyttöönotossa ja arvioinnissa.

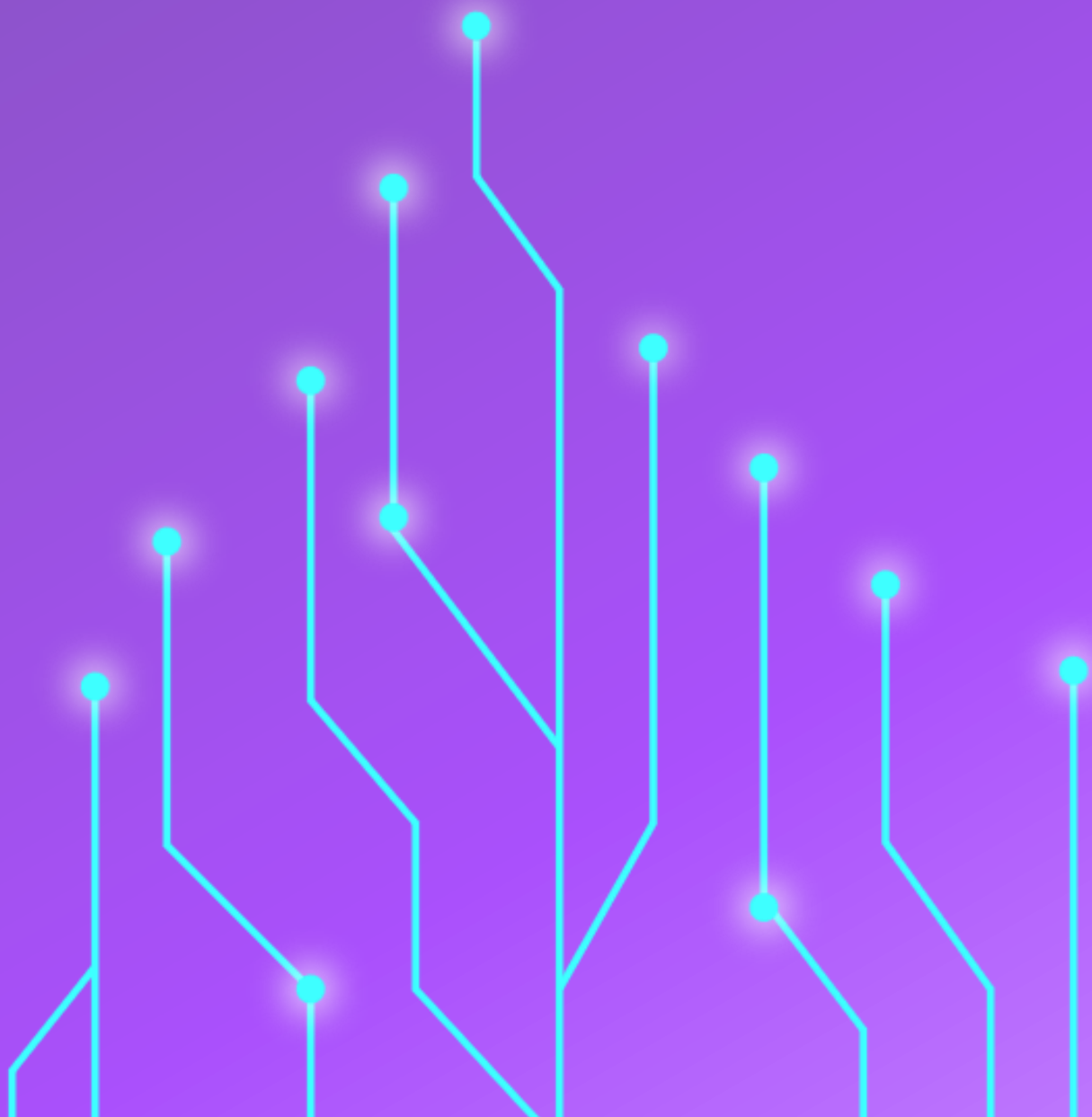
Ottamalla sidosryhmät mukaan toimintaan tekoälyn kehittäjät voivat saada arvokasta tietoa niiden tarpeista, mieltymyksistä ja huolenaiheista, mikä edistää avoimuutta, vastuullisuutta ja eettistä päätöksentekoa. Esimerkiksi tekoälyllä toimivien autonomisten ajoneuvojen kehittämisessä sitouttamalla sääntelyviranomaiset ja yhteiskunta varmistetaan, että turvallisuus-, yksityisyys- ja eettiset näkökohdat otetaan huomioon, mikä lisää avoimuutta ja luottamusta teknologiaan.





05. Yhteenveto

Osaamiskokonaisuus 4 | Läpinäkyvyys

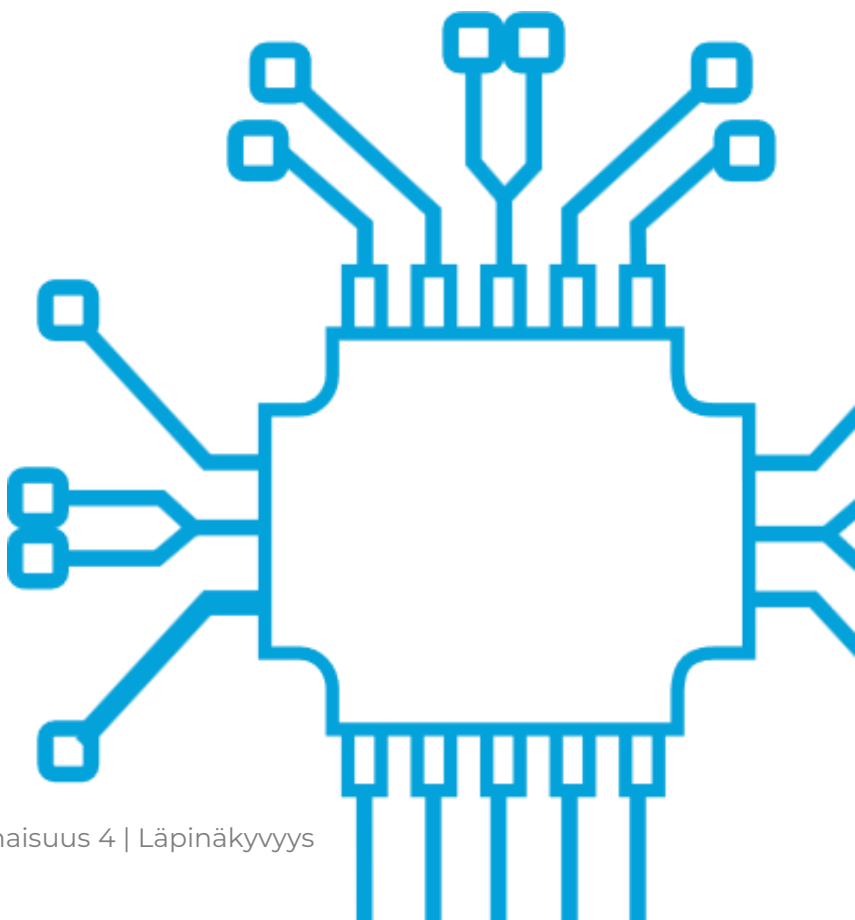


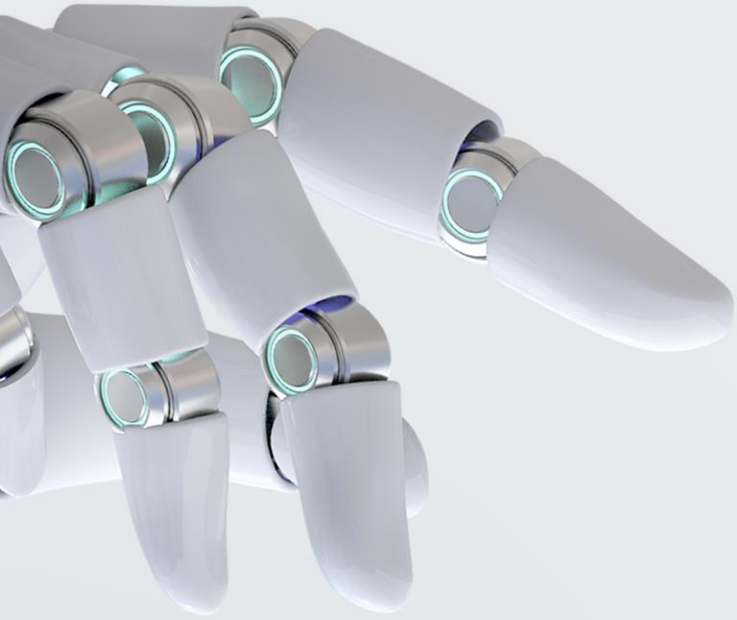


05. Yhteenveto

Yhteenvetona voidaan todeta, että tekoälyjärjestelmien avoimuuden merkitystä ei voi korostaa liikaa, sillä se muodostaa perustan luottamuksen ja vastuuvellvollisuuden rakentamiselle sekä algoritmisen vinouman lieventämiselle. Lisäksi avoimuuden ja algoritmisen vinouman välisen suhteen ymmärtäminen korostaa tarvetta puuttua läpinäkyvyyteen keinona tunnistaa, ehkäistä ja lieventää tekoälyjärjestelmien vinoutuneita tuloksia.

Lisäksi avoimuuden edistämiseen tähtäävien strategioiden tutkiminen antaa opiskelijoille käytännön työkaluja vastuullisuuden ja luottamuksen lisäämiseksi tekoälyteknologiaan. Kun opiskelijat ymmärtävät nämä käsitteet kattavasti, heillä on paremmat valmiudet selviytyä tekoälyn kehittämisen ja käyttöönoton eettisistä haasteista ja edistää vastuullisten ja oikeudenmukaisten tekoälyjärjestelmien kehittämistä.







Charlie



**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.



**Universitat
de les Illes Balears**



helixconnect



2022-1-ES01-KA220-HED-000085257