Tekoälyjärjestelmien haavoittuvuuksien testausalusta

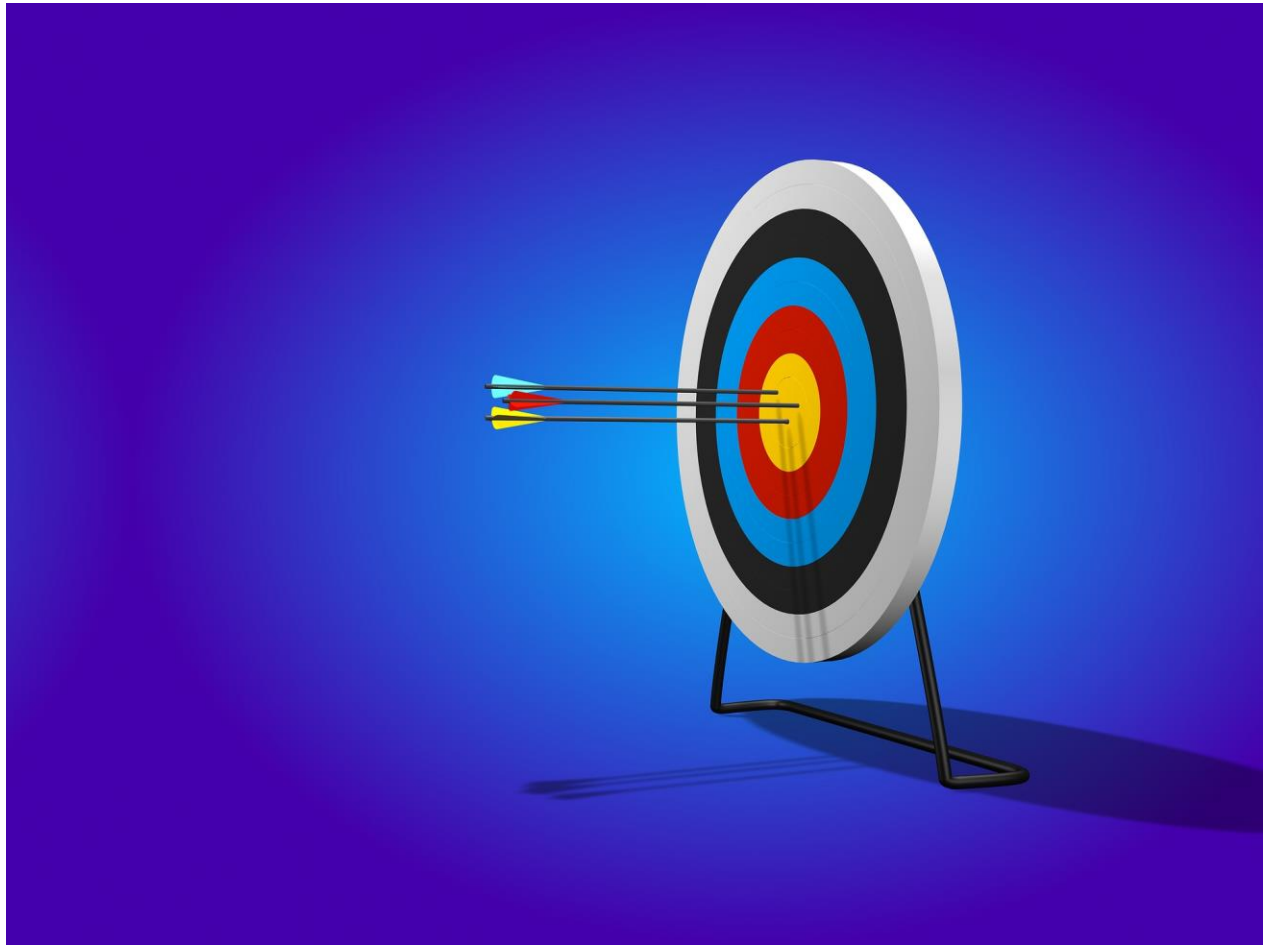
MATINEn rahoitus 158 960 euroa vuosille 2023-2024
2. vuoden tulokset

Kimmo Halunen
Kimmo.halunen@oulu.fi
Oulun yliopisto ja Maanpuolustuskorkeakoulu

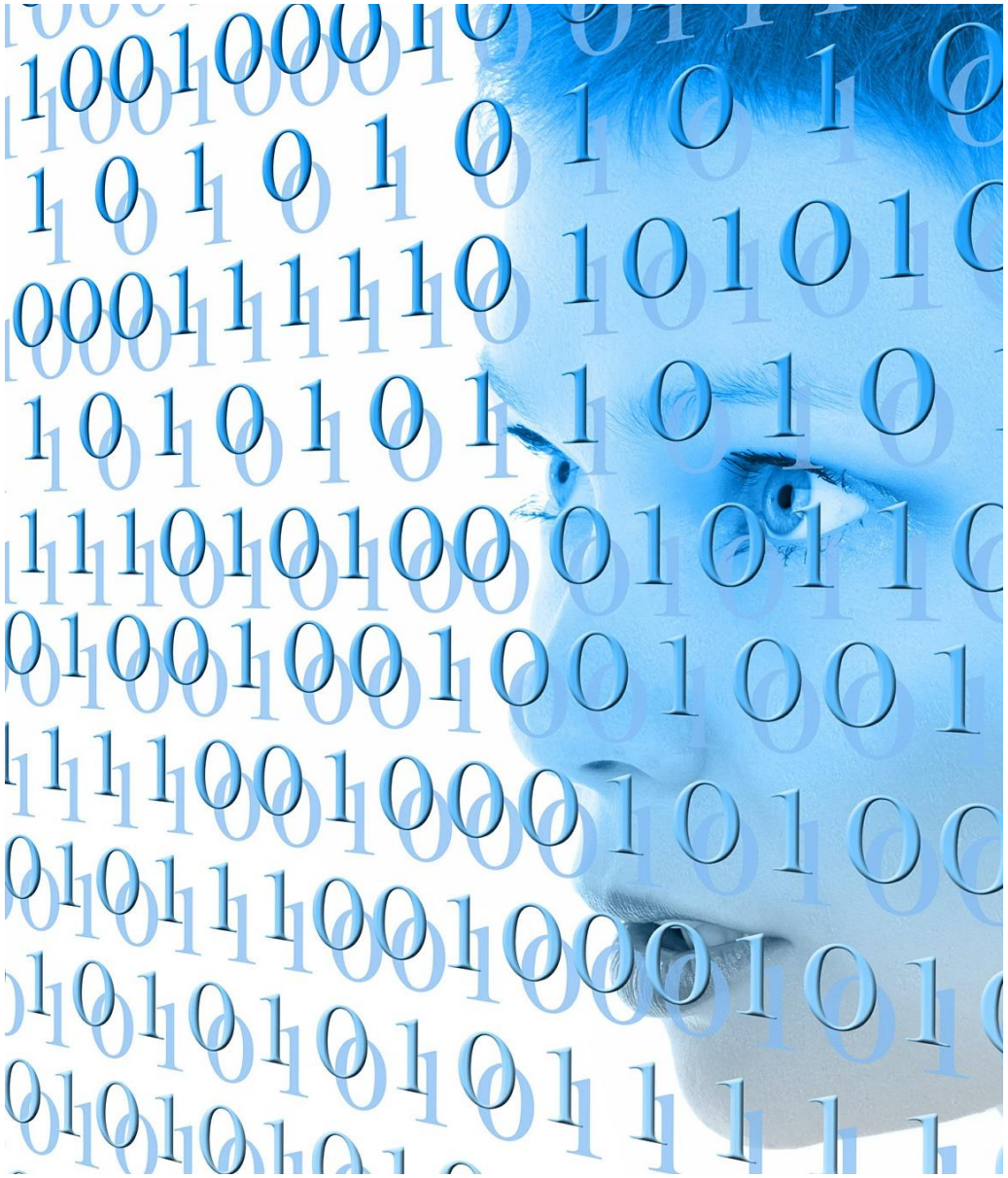
Mistä puhutaan?

- Tutkimuksen tavoite
- Tutkimuksen tausta
- Tämän hetkiset tulokset
- Tulevat toimenpiteet
- Yhteenveto
- Kysymyksiä?

Tutkimuksen tavoite



- Hankkeessa kehitetään tekoälyjärjestelmien haavoittuvuuksien testausalusta.
- Tavoitteena on tutkia alustan mahdollisuuksia erityisesti sensorifuusioon perustuvien tekoälymenetelmien harhauttamiseen ja haavoittuvuuksien löytämiseen.
- Hankkeessa tutkitaan myös mahdollisuuksia suojata järjestelmiä vaikuttamiselta.
- Testausjärjestelmää testataan yhdessä MPKK:n kanssa LAYKKA –alustan tekoälyjärjestelmien kanssa.
- Järjestelmä toteutetaan avoimen lähdekoodin alustana, mikä mahdollistaa järjestelmän jatkokehittämisen projektin jälkeen

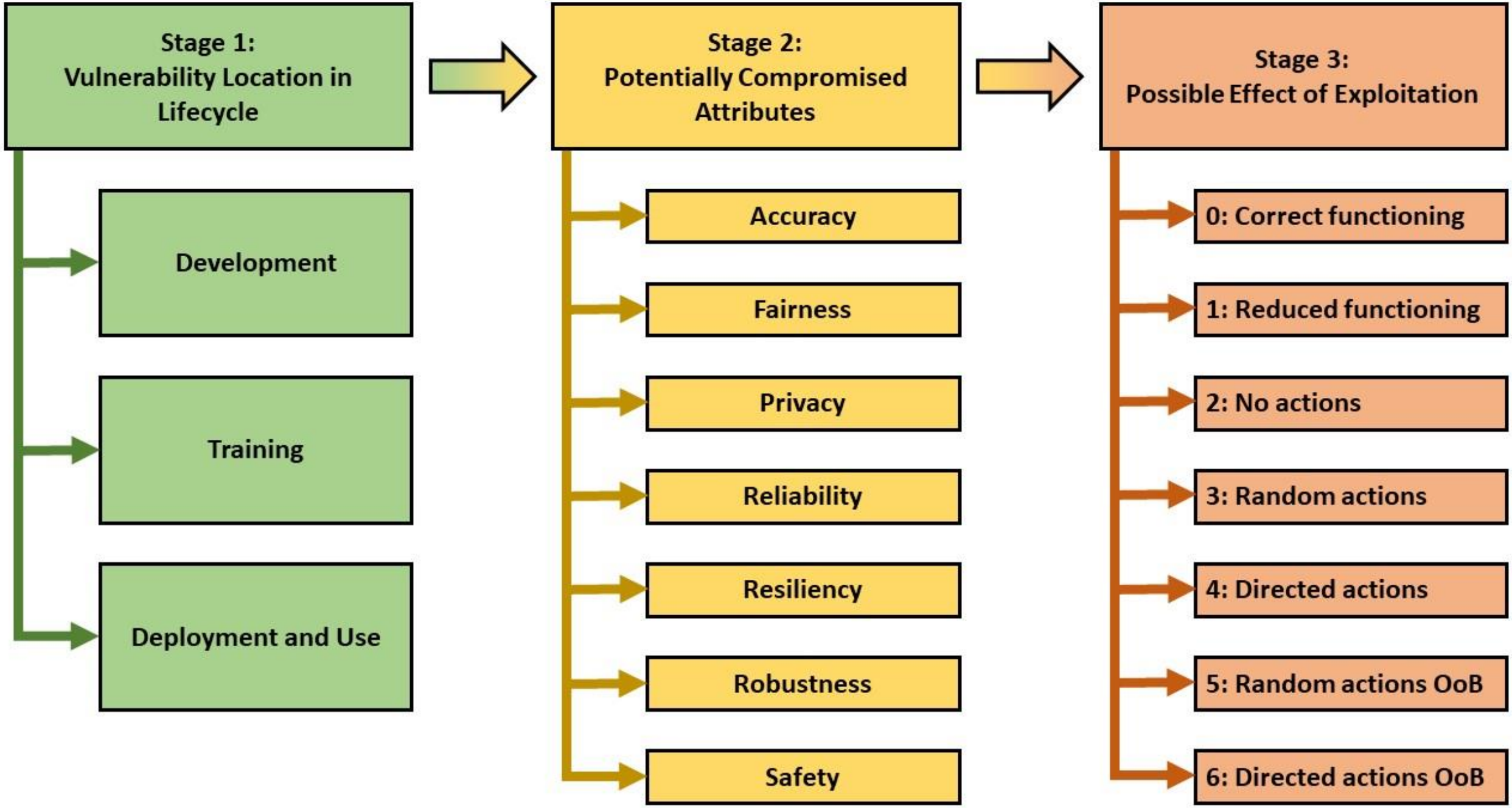


Tutkimuksen tausta

- Tekoälyjärjestelmät ovat yleistyneet yhteiskunnassamme valtavasti
- Monet Puolustusvoimien kannalta merkitykselliset järjestelmät käyttävät nyt tekoälyä ja koneoppimista itseohjautuvuuteen ja päätöksenteon tueksi
- Tekoälyjärjestelmien haavoittuvuuksien tuntemus on vähäistä
- Tekoälyjärjestelmien haavoittuvuuksien tutkimiseen ei ole samankaltaisia työkaluja kuin perinteisiin järjestelmiin
- Tällaisten menetelmien ja järjestelmien kehittäminen palvelee sekä Puolustusvoimia että laajemmin yhteiskuntaa

Tuloksia

- Tekoälyhaavoittuvuuksien taksonomia
- Julkaistu ECCWS –konferenssissa Jyväskylässä kesällä 2024
- Pispa, Arttu, and Kimmo Halunen. "A Comprehensive Artificial Intelligence Vulnerability Taxonomy." European Conference on Cyber Warfare and Security. Vol. 23. No. 1. 2024.


Stage 1:
Vulnerability Location in Lifecycle
Development
Training
Deployment and Use
Stage 2:
Potentially Compromised Attributes
Accuracy
Fairness
Privacy
Reliability
Resiliency
Robustness
Safety
Stage 3:
Possible Effect of Exploitation
0: Correct functioning
1: Reduced functioning
2: No actions
3: Random actions
4: Directed actions
5: Random actions OoB
6: Directed actions OoB

Tuloksia

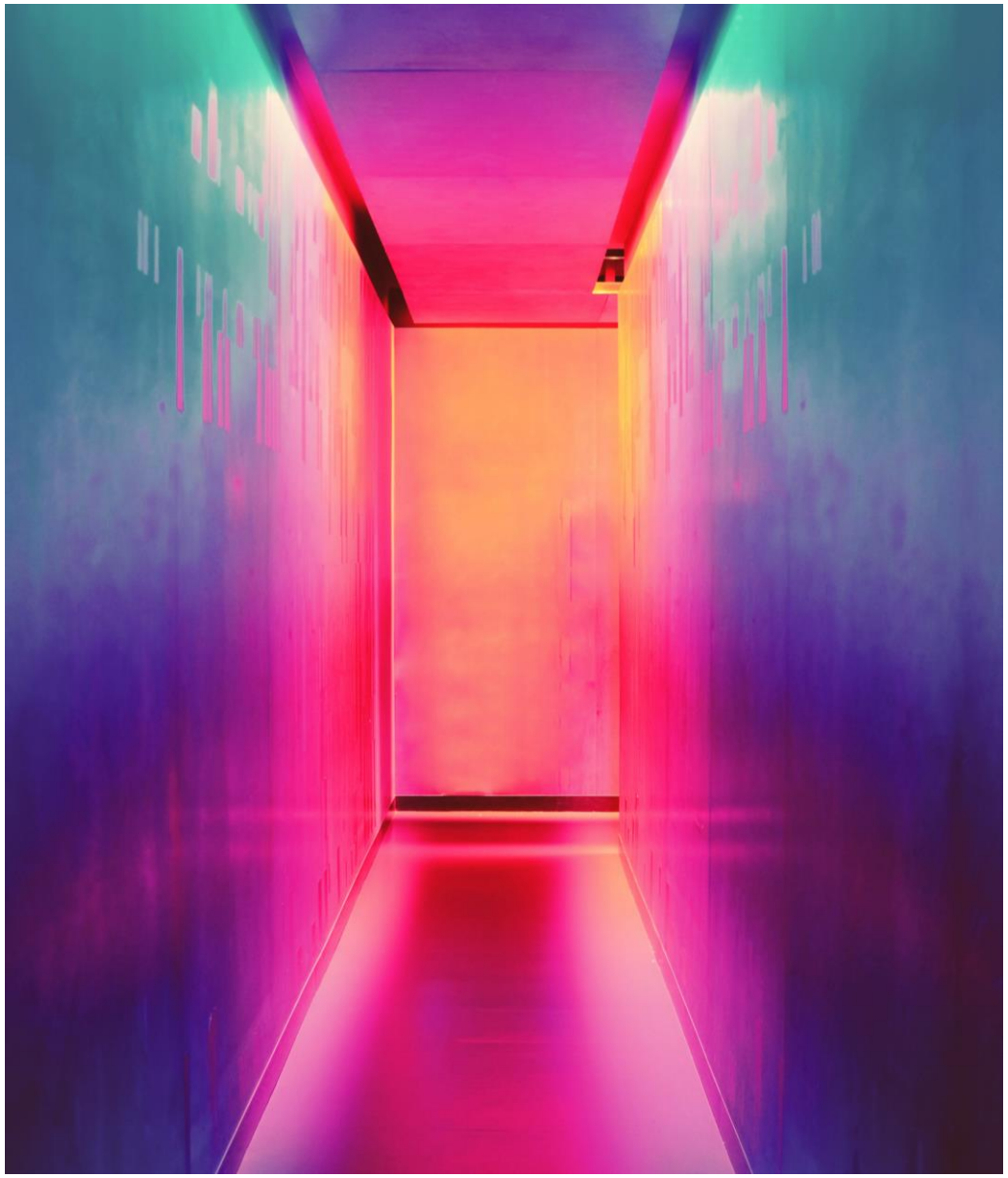
- Tekoälyjärjestelmien haavoittuvuuksien testausalustan kehitys
  - Aluksi vaikutti, että alustoja on useita
  - Alustavat tulokset osoittivat, että vain yksi (ART, Adversarial Robustness Toolbox) on käyttökelpoinen projektin tarpeisiin (Anssi Antilan DI-työ)
  - Tätä testattiin ja aloitettiin jatkokehitys eli uusien testien toteuttaminen ART:lle
  - Uudet testit ovat osoittaneet, että ART ei sovellu sellaisenaan uusien menetelmien lisäämiseksi ja näiden testaamiseksi (Jakob Colesin diplomityö, kesken)

Tuloksia

- Tekoälyhaavoittuvuuksien ilmeneminen tosielämässä
- Tutkimus julkaistu TETHICS 2024 –konferenssissa
  - Hyvin vähän todennettuja tosielämän vaikutuksia
  - Erityisesti, mikäli verrataan perinteisiin haavoittuvuuksiin
  - Tietolähteitä on vähän ja ne saattavat antaa hyvin vinoutuneen kuvan tästä aiheesta
- Tekoälyhaavoittuvuuksien tietokanta
  - Alustava tietokanta suunniteltu ja rakennettu
  - Tarkoitus julkaista ja alkaa ylläpitämään vuoden 2025 alussa

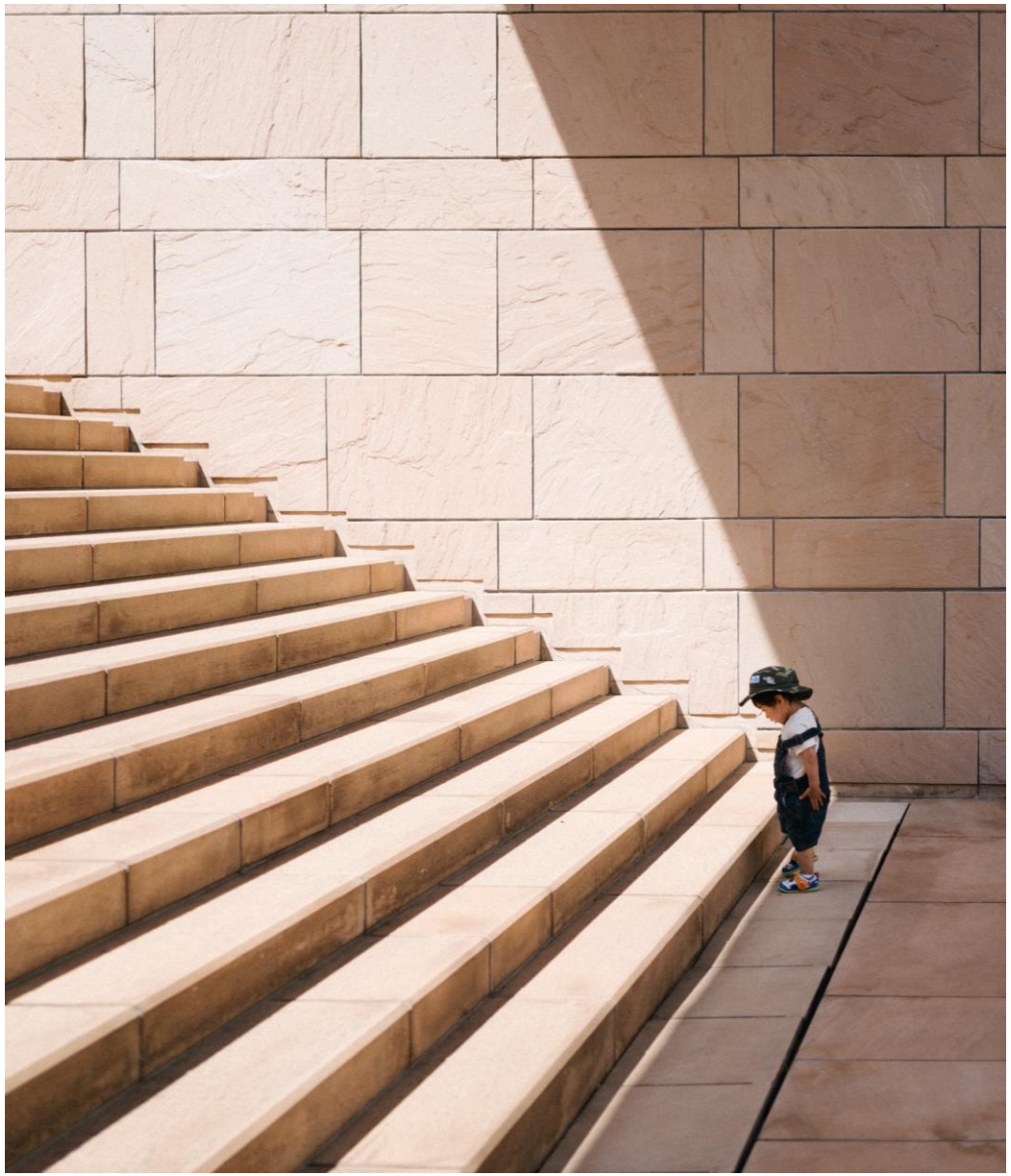
Tuloksia

- Suojautumismenetelmien tutkimus
- Diplomityö valmistuu syksyllä 2024 (Dennis Goyal, arvioitu, hyväksynnässä)
- Tuloksena löydettiin menetelmiä, joilla sensorifuusiota vastaan voidaan hyökätä
- Toisaalta tunnistettiin myös suojautumisen mahdollisuuksia
  - Nämä saattavat vaatia kuitenkin paljon lisää resursseja



Tulevat toimenpiteet

- Projektille haettu jatkoaikaa kesäkuun 2025 loppuun saakka
- Koska ART-alustaan ei voitu lisätä uusia menetelmiä helposti, suunnitellaan uusi alusta ja tehdään siitä ensimmäinen prototyyppi
- Jatkorahoitusta tämän uuden alustan kehittämiselle kesäkuusta 2025 eteenpäin on haettu
- Tällaista työkalua tarvitaan



Yhteenveto

- Tekoälyhaavoittuvuudet ovat sekä samanlaisia että erilaisia kuin perinteiset haavoittuvuudet
  - Tarve taksonomialle ja ehkä uusille tavoille raportoida haavoittuvuuksia
  - Sekä uusia hyökkäyksiä että uusia puolustusmekanismeja julkaistaan jatkuvasti
- Uusi testausalusta pitää ehkä sittenkin tehdä ihan itse ☺
- Vielä tällä hetkellä tekoälyhaavoittuvuuksien vaikutukset ovat huomattavasti vähäisemmät kuin perinteisten haavoittuvuuksien

Kysymyksiä?

Kiitos paljon mielenkiinnosta!


MAKE THIS
WORLD
BETTER