Käyttöliittymien pimeät mallit sekä niiden tunnistusmenetelmät

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Huhtikuu 2025
Kari-Pekka Turtiainen

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -järjestelmällä.

TURUN YLIOPISTO
Tietotekniikan laitos

Kari-Pekka Turtiainen: Käyttöliittymien pimeät mallit sekä niiden tunnistusmenetelmät

TkK-tutkielma, 26 s.
Tietotekniikka
Huhtikuu 2025

Pimeät mallit ovat käyttöliittymissä käytettyjä keinoja, joiden suunnittelijat tietoisesti koettavat hämätä käyttäjiä, tehdä käyttäjien varsinaisten halujen toteuttamisesta haastavaa, tai manipuloida käyttäjiä tekemään heille epäsuotuisia toimia. Tutkielma tarjoaa syvällisen katsauksen pimeisiin malleihin sekä niiden luokitteluun kolmitasoisen ontologian avulla. Tämän lisäksi pimeiden mallien vaikutuksia tarkastellaan neljän normatiivisen linssin kautta, jotka osoittavat, että pimeät mallit voivat heikentää käyttäjien päätöksentekokykyä, vähentää yksityisyyttä ja luottamusta digitaalisiin palveluihin sekä johtaa taloudellisiin ja oikeudellisiin haittoihin.
Tutkielmassa myös perehdytään pimeiden mallien tunnistamiseen kehitettyihin manuaalisiin ja automaattisiin menetelmiin sekä niiden arvioimiseen kehitettyihin mittareihin. Tulokset osoittavat, että automaattiset algoritmiset tunnistusmenetelmät kuten logistinen regressio, SVM ja satunnaismetsä saavuttavat korkean tunnistustarkkuuden ja tasapainoisen suorituskyvyn eri mittareilla. Manuaaliset menetelmät soveltuvat kuitenkin automaattisiin menetelmiin nähden paremmin tunnistusluokitusten laatimiseen ja referenssilistojen kokoamiseen.

Asiasanat: Pimeät mallit, harhaanjohtavat suunnittelumallit, käyttäjämanipulaatio, käyttöliittymäsuunnittelu, automaattinen tunnistus, manuaalinen analyysi, käyttöliittymäpsykologia

Sisällys

1 Johdanto **1**

2 Pimeät mallit **5**

2.1 Suunnittelumallien historia 5

2.2 Termin määrittely 6

2.3 Pimeiden mallien luokittelu 8

2.4 Pimeiden mallien vaikutukset 13

2.4.1 Yksilön hyvinvointi 13

2.4.2 Kollektiivinen hyvinvointi 14

2.4.3 Sääntelylliset tavoitteet 15

2.4.4 Yksilön itsemääräämisoikeudet 16

3 Pimeiden mallien tunnistusmenetelmät **17**

3.1 Manuaaliset menetelmät 18

3.2 Automaattiset menetelmät 20

3.3 Haasteet ja ongelmat 23

4 Yhteenveto ja pohdintaa **24**

Lähdeluettelo **27**

Kuvat

3.1 Ontologian luomisprosessi [8] 19

3.2 Pimeiden mallien luokittelu tyyppien ja ominaisuuksien perusteella [12] 20

Taulukot

3.1 Pimeiden/ei-pimeiden mallien luokittelutulokset [12] 22

Termistö

DETOUR (Deceptive Experiences To Online Users Reduction) Yhdysvaltalainen lakiehdotus pimeitä malleja vastaan

Evästeet Tekstitiedostoja, jotka tallennetaan tietokoneen tai mobiililaitteen selaimen muistiin ja jonka avulla voidaan tallentaa käyttäjän tietoja

GDPR (General Data Protection Regulation) Euroopan unionin asetus, jolla pyritään yhtenäistämään tietosuojaa koskeva lainsäädäntö EU:n jäsenvaltioiden kesken

HCI (Human-Computer Interaction) tarkoittaa ihmisen ja koneen välistä vuorovaikutusta ja sen tutkimista

LLM (Large Language Models) tekoälymalleja, jotka ymmärtävät ja tuottavat tekstiä hyödyntäen suuria tekstiaineistoja ja neuroverkkoja

NLP (Natural Language Processing) luonnollisen kielen käsittelyä jolla pyrkitään erityisesti kehittämään tietokoneohjelmia jotka pystyvät käsittelemään ja analysoimaan suurta määrää kielidataa

Ontologia Tarkoittaa jäsenneltyä tapaa kuvata ja luokitella tietoa jossa määritellään käsitteet ja niiden väliset suhteet

Pimeät mallit Käyttöliittymäratkaisut, joiden tarkoituksena on manipuloida käyttäjää tekemään ei-toivottuja tai epäsuotuisia valintoja

1 Johdanto

Digitalisaation kasvaessa lisääntyvät myös tuotteiden sekä verkkosivujen käyttäjämäärät, jolloin yritysten kilpaileminen käyttäjistä on yhä kovempaa. Tästä johtuen, organisaatiot kokeilevat yhä radikaalimpia keinoja käyttäjämäärien ja liiketulojen kasvattamisessa. Sovellusten ja verkkosivujen käyttöliittymät suunnitellaan useasti olemaan miellyttäviä käyttäjille. Yritysten optimoidessa liiketoimintaa käyttöliittymiin on kuitenkin alettu lisätä käyttäjien toimintaa ohjaavia malleja. Kun käytetyn mallin tarkoituksena on harhaan johtaa käyttäjää toimimaan omien etujensa vastaisesti, voidaan puhua pimeistä malleista (engl. dark pattern).

Kyseiset mallit ovat suhteellisen uusi ilmiö. Vuonna 2010 Harry Brignull huomasi, että käyttöliittymät alkoivat käyttää hyväksi yhä enemmän manipuloivia, hämääviä ja pakottavanomaisia suunnittelumalleja uudessa sekä kasvavassa digitaalisessa maailmassa. Hän halusi tehdä asian eteen jotain, joten hän loi verkkosivuston darkpatterns.org [1] sekä nimesi ”dark patterns” -termin ensimmäistä kertaa [1]. Tämä verkkosivusto loi ensimmäisen varsinaisen pohjan ilmiön tutkimiselle, sekä paremman käsityksen siitä, kuinka tätä uutta käyttöliittymäsuunnittelua tulisi lähestyä.

Tämän tutkielman kirjoittaja huomasi pimeiden mallien ilmiön ensimmäistä kertaa vuonna 2018 kun EU:n GDPR-lainsäädäntö (General Data Protection Regulation) astui voimaan. Verkkosivustot olivat jo pitkään käyttäneet evästeitä käyttäjätietojen keräämiseen, mutta ennen GDPR:ää suostumusta ei yleensä pyydetty. Orga-

[1] nykyään: `https://www.deceptive.design`

nisaatiot huomasivat, että ne menettivät suuren määrän arvokkaita käyttäjätietoja, kun käyttäjät alkoivat kieltäytyä suostumuksesta. Organisaatiot löysivät kuitenkin keinoja tämän kiertämiseksi. Hylkäämisvaihtoehtoja piilotettiin eri valikoiden taakse, ja evästeiden hylkäämisestä tehtiin mahdollisimman haastavaa. Kieltäytyminen oli silti mahdollista, mutta hankaloittamisesta johtuen, tietojen luovuttamisesta tuli monille vaivattomin vaihtoehto. Palvelua tarjoavat organisaatiot pystyivät toimimaan laillisuuden nimissä, mutta saada alkuperäisen lakeja edeltävän tahdon läpi joka tapauksessa.

Tämän kandidaattitutkielman tarkoituksena on selvittää, mitä menetelmiä pimeiden mallien tunnistamiseen ja luokitteluun on kehitetty. Tutkielma tulee perehtymään tunnistusmenetelmissä käytettyihin pimeiden mallien luokitteluihin, jonka lisäksi pimeiden mallien vaikutuksia tarkastellaan neljän normatiivisen linssin läpi. Tutkielman aihe on tärkeä, koska yksityisyys, itsemääräämisoikeus ja digitaalinen hyvinvointi ovat tärkeitä, mutta pimeät mallit pyrkivät käyttäjäpsykologisin keinoin näitä hajottamaan. Tämä vaikuttaa negatiivisesti erityisesti palveluiden käyttäjiin usein heidän tietämättään. Tunnistusmenetelmät ovat tärkeä tutkimuskohde, koska niiden avulla voidaan paikantaa pimeitä malleja, lisätä tietoisuutta niistä sekä vähentää käyttäjien manipulointia. Näiden teemojen ympärille rakentuvat myös tutkielman keskeisessä osassa olevat tutkimuskysymykset.

TK1: Kuinka pimeitä malleja voidaan luokitella?

Pimeät mallit ovat olleet nopeasti kasvava ilmiö, joten niitä on lukuisia erilaisia. Jotta niiden tunnistamiseen voidaan kehittää tehokkaita sekä tarkkoja menetelmiä, on tärkeää ymmärtää, kuinka niitä voidaan luokitella, ja miten eri pimeiden mallien luokat eroavat toisistaan. Lisäksi tarkastellaan, kuinka pimeät mallit eroavat muusta käyttöliittymäsuunnittelusta.

TK2: Mitä vaikutuksia pimeillä malleilla on?

Pimeiden mallien vaikutusten ymmärtäminen on keskeistä niiden tunnistusmenetelmien kehittämisen kannalta. Tämän kysymyksen pohjalta perehdytään siihen, millaisia vaikutuksia pimeillä malleilla on erityisesti palveluiden käyttäjiin. Näiden vaikutusten kartoittaminen auttaa hahmottamaan miksi tunnistusmenetelmien kehittäminen on tärkeää, sekä millaisia piirteitä tunnistusmenetelmien tulisi havaita.

TK3: Mitä menetelmiä pimeiden mallien tunnistamiseen on kehitetty?

Pimeiden mallien luokittelun ja vaikutusten tarkastelun pohjalta siirrytään niiden tunnistamiseen. Tämä kysymys keskittyy selvittämään, mitä menetelmiä pimeiden mallien tunnistamiseen ja luokitteluun on kehitetty, kuinka ne toimivat, mitä ne mittaavat sekä kuinka tehokkaita ja luotettavia nämä mallit tällä hetkellä ovat. Tutkielmassa perehdytään manuaalisiin ja automaattisiin menetelmiin.

Pimeiden mallien ilmiön tuoreuden sekä siihen liittyvän kirjallisuuden nopean lisääntymisen takia tutkimuksessa on pyritty keskittymään mahdollisimman uusiin tutkimuspapereihin. Aiheeseen liittyvä kirjallisuus on vielä puutteellista ja tutkimusmateriaalia on saatavilla rajallisesti, joten myös joitain vanhempia tutkimuksia on käytetty. Lähteet tutkimukseen on etsitty pääosin Web Of Science:n sekä Google Scholar:in avulla. Tutkimuspapereiden löytämiseksi kyseisistä tietolähteistä haettua tietoa on etsitty siihen liittyvillä hakutermeillä. Käytetyt hakutermit olivat esimerkiksi ”dark pattern*” AND (”automated detection” OR ”algorithmic classification”) -tyyppisiä. Kyseisellä hakutermillä löydettyjen lähteiden määrä oli 132 kappaletta Google Scholar:ssa ja 5 Web Of Science:ssa. Yleisluontaisempia hakutermejä käyttäessä lähteiden rajaamiseksi käytettiin hakukoneiden rajaamistoiminnallisuuksia kuten aihealueen rajaamista tietotekniikaksi. Lähteet seulottiin sen perusteella, kuinka tarkasti ne liittyivät aiheeseen otsikon, abstraktin ja päätösosan perusteella. Tutkimuksessa käytettiin ainoastaan englanninkielisiä lähteitä. Lisäksi tietoa

haettiin käytettyjen tutkimuksien siteeraamista lähteistä sekä aiheen vakiintuneiden asiantuntijoiden muista tutkimuksista.

Tutkielman rakenne perustuu kahden pääkäsittelyluvun ympärille joista ”2. Pimeät mallit” syventyy pimeisiin malleihin tarkemmin. Luvun alussa kerrotaan lyhyesti suunnittelumallien historiasta. Luvun toinen osa selittää, mitä termi tarkoittaa, kuinka se määritellään, sekä kuinka pimeät mallit eroavat muista samankaltaisista käyttöliittymäsuunnittelun malleista. Tämän jälkeen luku siirtyy käsittelemään kuinka eri pimeitä malleja voidaan luokitella. Tämän luvun viimeisessä osiossa selvitetään, mitä vaikutuksia pimeillä malleilla on neljän esitetyn normatiivisen linssin kautta. Tutkielman kolmas luku ”3. Pimeiden mallien tunnistusmenetelmät” siirtyy käsittelemään pimeiden mallien tunnistamiseen kehitettyjä menetelmiä. Luku tutkii sekä manuaalisia, että automaattisia menetelmiä, menetelmissä käytettyjä mittareita ja menetelmien tehokkuutta. Tutkielman viimeinen luku tulee sisältämään yhteenvedon sekä yleistä pohdintaa aiheeseen nähden. Tässä viimeisessä luvussa vastataan myös esitettyihin tutkimuskysymyksiin.

2 Pimeät mallit

Tämä kappale tulee syventymään pimeisiin malleihin tarkemmin. Ensimmäisessä luvussa avataan suunnittelumallien historiaa, jonka jälkeen siirrytään termin määrittelyn kautta pimeiden mallien luokitteluun. Viimeisessä luvussa käsitellään pimeiden mallien vaikutuksia neljän normatiivisen linssin läpi.

2.1 Suunnittelumallien historia

Suunnittelumallien historia juontaa juurensa arkkitehti Christopher Alexanderin vuonna 1977 julkaisemaan kirjaan *A Pattern Language: Towns, Buildings and Construction*, jossa hän esittelee rakennus- ja kaupunkiarkkitehtuurissa käytettyjä suunnittelumalleja. Kirjassa esitetty lähestymistapa tarkastelee rakentamisessa toistuvia, tietyssä kontekstissa esiintyviä suunnitteluongelmia ja tarjoaa niihin tyydyttäviä ratkaisuja. Tämä suunnittelumallien lähestymistapa on myöhemmin omaksuttu myös muilla alueilla, kuten ohjelmistotuotannossa- ja suunnittelussa. Esimerkiksi Gamma ym. julkaisivat 1995 kokoelman ohjelmistosuunnittelun malleja, jotka ovat laajalti käytössä tänäkin päivänä. Suunnittelumallien tarkoitus on kasata ja tarjota vakiintuneita ratkaisuja toistuviin ja yleisiin suunnitteluongelmiin. [2], [3] Samalla tavalla kuin suunnittelumallit tarjoavat hyviä käytäntöjä, myös pimeät mallit voidaan nähdä eräänlaisina negatiivisina suunnittelumalleina, jotka tarjoavat vakiintuneita käytäntöjä huonoihin tarkoituksiin käyttäjien näkökulmasta.

2.2 Termin määrittely

”Dark pattern” -termi on suhteellisen uusi, joten termin tarkka määritelmä on osoittautunut toistaiseksi vielä haastavaksi. Termi on ollut jatkuvassa muutoksessa ja siihen on jatkuvasti lisätty sekä siitä on poistettu määritelmiä. Kun Brignull nimesi termin vuonna 2010, hän kuvaili pimeitä malleja ”verkkosivustoissa ja sovelluksissa käytettyinä temppuina, jotka saavat sinut tekemään asioita joita et ole tarkoittanut, kuten ostamaan tai rekisteröitymään johonkin” [4]. Yksinkertaisesti tiivistäen voidaan siis sanoa, että pimeät mallit ovat käyttöliittymissä käytettyjä keinoja, joiden suunnittelijat tietoisesti koettavat hämätä käyttäjiä, tehdä käyttäjien varsinaisten halujen toteuttamisesta haastavaa tai manipuloida käyttäjiä tekemään heille epäsuotuisia toimia.

Pimeistä malleista voidaan käyttää myös nimitystä ”harhaanjohtavat suunnittelumallit”. Euroopan tietosuojaneuvoston ohjedokumentti [1] selittää termin vastaavasti: ”Harhaanjohtavilla suunnittelumalleilla tarkoitetaan sosiaalisen median alustoilla toteutettuja käyttöliittymiä ja käyttötapoja, joilla yritetään vaikuttaa käyttäjiin, jotta he tekisivät tahattomia, vastentahtoisia ja mahdollisesti haitallisia päätöksiä, jotka usein ovat käyttäjien etujen vastaisia ja sosiaalisen median alustan etujen mukaisia, ja jotka koskevat heidän henkilötietojensa käsittelyä. Harhaanjohtavilla suunnittelumalleilla pyritään vaikuttamaan käyttäjien käyttäytymiseen, ja ne voivat estää käyttäjiä suojaamasta henkilötietojaan tehokkaasti ja tekemästä tietoisia valintoja.” Pimeiden mallien erottamiseksi muista samankaltaisista käytännöistä on hyödyllistä ymmärtää seuraavaksi käsiteltävät termit.

Kehittäjät voivat tehdä virheitä käyttöliittymien suunnittelussa esimerkiksi huonojen suunnittelukäytäntöjen vuoksi. Usein näitä virheitä pyritään korjaamaan tai piilottamaan nopeilla ratkaisuilla jolloin nämä nopeat ratkaisut johtavat usein uusiin ongelmiin. Näitä kyseisiä käytäntöjä kutsutaan **vastamalleiksi** (engl. anti-

[1] *Deceptive design patterns in social media platform interfaces: how to recognise and avoid them*

pattern) [5]. Toisin kuin pimeät mallit, niin vastamallien tarkoituksena ei kuitenkaan ole kehittäjän näkökulmasta tarkoituksellisesti manipuloida tai vaikuttaa käyttäjiin negatiivisesti, vaikka niiden esiintyminen voi tähän kuitenkin joskus tahattomasti johtaa.

Valoisat mallit (engl. light patterns) ovat nimensä mukaisesti pimeiden mallien vastakohtia. Siinä missä pimeät mallit manipuloivat käyttäjiä ja toimivat heidän etujaan vastaan, valoisat mallit tukevat käyttäjän autonomiaa ja päätöksentekoa [6]. Ne ovat suunniteltu edistämään käyttäjän etua esimerkiksi lisäämällä läpinäkyvyyttä, helpottamalla informoituja valintoja ja suojelemalla käyttäjän yksityisyyttä. Esimerkiksi nykyaikaiset älypuhelinjärjestelmät, kuten Applen iOS ja Android, pyytävät käyttäjältä lupaa sovellusten tarvitsemille oikeuksille, kuten kamerarullan tai sijaintipalvelun käyttöön. Jos käyttäjä aluksi kieltäytyy antamasta näitä oikeuksia, mutta muuttaa myöhemmin mielensä, hän voi aktivoida ne uudelleen muutaman valikon kautta. Tässä tapauksessa kieltäytyminen on tehty helpoksi, kun taas päätöksen muuttaminen käyttäjän tietoturvaa mahdollisesti heikentävään suuntaan vaatii enemmän vaivaa. Tämä mekanismi suojaa käyttäjää harkitsemattomilta päätöksiltä ja tukee hänen yksityisyyttään.

Harmaat mallit (engl. gray patterns) ovat pimeiden- ja valoisien mallien välimuotoja, eli harmaalla alueella olevia käyttöliittymien toimintamalleja. Niillä voi olla joko positiivisia tai negatiivisia vaikutuksia käyttäjästä sekä tilanteesta riippuen. [7] Sanotaan esimerkiksi että käyttäjä katsoo sarjaa suoratoistopalvelun kautta. Kun jakso loppuu, niin suoratoistopalvelu vaihtaa automaattisesti jakson seuraavaan. Henkilön ei tarvitse itse siirtyä seuraavaan jaksoon, ja hän voi keskittyä sarjasta nauttimiseen. Toisaalta jos henkilöllä on muita tärkeämpiä velvollisuuksia hoidettavana, ja uuteen jaksoon automaattinen siirtyminen vaikeuttaa hänen palvelun käytöstä irtautumista, niin vaikutukset voivat olla negatiivisia. Yleensä har-

maat mallit yrittävät parantaa käyttäjäkokemusta ja edistävät käyttäjän toiminnan jatkumista, mutta niiden vaikutukset riippuvat kuitenkin esiintymiskontekstista.

2.3 Pimeiden mallien luokittelu

Pimeiden mallien luokittelu on jatkuvasti kehittyvä prosessi, ja eri tutkijat sekä asiantuntijat käyttävät edelleen hieman eri lähestymistapoja niiden luokittelemiseen. Vaikka tarkkaa yhteisymmärrystä luokittelun tarkasta rakenteesta ei vielä ole saavutettu, niin tutkimusta tämän yhteisen yhteisymmärryksen saavuttamiseksi jatketaan aktiivisesti.

Grey ym. [8] yrittivät tutkimuksessaan *An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action* yhdentää useiden eri lähteiden pimeiden mallien tyyppiluokituksia sekä luoda näiden pohjalta universaalin luokittelumallin. He paikansivat vuoteen 2022 mennessä julkaistusta kirjallisuudesta yhteensä 262 eri luokitusta. Tutkimuksessaan he halusivat kuitenkin tarkentaa luokittelujen yhtenäistä määritelmää, joten he kehittivät mallin, joka koostuu kolmetasoisesta ontologiasta. Tämä ontologia sijoittelee pimeät mallit kolmeen eri kategoriaan: ylä-, keski- ja matalan tason luokitukseen. Tällä hetkellä kyseinen tutkimus sisältää edistyneimmän version pimeiden mallien luokittelusta, koska se pyrkii yhdistämään muita aiemmin luotuja luokittelumalleja. Seuraavat kappaleet tulevat avaamaan eri tasojen luokitusmalleja tämän tutkimuksen perusteella.

Ylätason pimeät mallit on kaikkein abstraktein luokitustaso, joka sisältää viisi yleistä strategiaa manipuloivien, pakottavien tai harhaanjohtavien elementtien sisällyttämiseksi. Nämä strategiat rajoittavat käyttäjän autonomiaa ja päätöksentekoa esimerkiksi toimintamahdollisuuksien löytämisen vaikeuttamisella, epäolennaisten tehtävien korostamisella tai käyttäjää hämmentämällä. Toteutustapoihin kuuluvat muun muassa toimintojen piilottaminen, viivyttäminen, uudelleenohjaaminen, esteiden luominen ja valintojen painottaminen käyttöliittymässä. Useimmat näistä

strategioista toimivat huomaamattomasti ja johtavat käyttäjän harhaan, mutta esimerkiksi pakotettu toiminta (engl. forced action) -malli perustuu käyttäjän pakottamiseen tiettyihin toimiin [6]. Ylätason pimeiden mallien luokitukset ovat kontekstista riippumattomia, ja niitä voidaan hyödyntää monilla alustoilla ja teknologioilla (esim. työpöytä- ja mobiilikäyttöliittymät) sekä erilaisissa sovellustyypeissä (esim. verkkokauppa, pelit, sosiaalinen media) [8], [9].

Keskitason luokitus toimii siltana ylätason ja alatason luokitusten välillä ja se muodostettiin käyttäjien odotusten pohjalta. Sen sisältämät pimeät mallit perustuvat käyttäjän odotuksien manipuloimiseen tavalla, joka johtaa odottamattomiin tai haitallisiin lopputuloksiin. Keskitason pimeät mallit hyödyntävät hyökkäyskulmia tai tiettyjä lähestymistapoja, joilla käyttäjän mahdollisuuksia tehdä itsenäisiä ja informoituja päätöksiä rajoitetaan, heikennetään tai sabotoidaan. Käyttäjän odotuksiin kuuluvat esimerkiksi merkityksellinen ja oikea-aikainen tiedonsaatavuus, tavoitteen ja toiminnan vastaavuus, annettujen tietojen täydellisyys ja totuudenmukaisuus sekä mahdollisuus muuttaa mieltään tai peruuttaa tehty päätös. Keskitason pimeät mallit voivat vaikuttaa käyttäjään monin tavoin, kuten aiheuttamalla odottamattomia tai epätoivottuja lopputuloksia, lisäämällä hämmennystä tai painetta, estämällä käyttäjää löytämästä olennaista tietoa tai ohjaamalla häntä tekemään valinnan jota hän ei muuten tekisi. Keskitason pimeät mallit voivat olla joko staattisia, jolloin ne ilmenevät tietyssä hetkessä käyttäjäpolulla, tai ajallisia, jolloin niiden vaikutukset rakentuvat pidemmän aikavälin vuorovaikutuksen myötä. Näitä keskitason malleja on tutkimuksessa tunnistettu 24 kappaletta, ja niiden tulkinta voi vaihdella kontekstin ja käyttösovelluksen mukaan. [8]

Alatason luokitustaso sisältää 35 tilannesidonnaista ja kontekstiriippuvaista mallia, jotka kuvaavat konkreettisia toteutustapoja käyttäjän autonomian ja päätöksenteon heikentämiseksi. Näitä toteutuksia ovat esimerkiksi ristiriitaisen tai harhaanjohtavan tiedon tarjoaminen, tiettyjen vuorovaikutusten estäminen, käyttäjän

huomiota harhauttavat elementit sekä sosiaalisen paineen hyödyntäminen päätöksenteossa. Näiden pimeiden mallien toteutustavat voivat ilmetä visuaalisessa ja/tai ajallisessa muodossa ja ne vaikuttavat suoraan käyttäjäkokemukseen, mutta sen eri osa-alueilla. Tämä toteutuu esimerkiksi estämällä käyttäjää tekemästä oikeaan tietoon perustuvia päätöksiä tai huomaamatta ohjaamalla häntä ei-toivottuihin valintoihin.

Ylä- eli korkeimman tason luokan malleihin kuuluvat estäminen, piilottaminen, käyttöliittymähäirintä, pakotettu toiminta ja sosiaalinen manipulointi. Tutkimuksessa kehitetty luokittelumalli etenee kolmivaiheisena hierarkiana: ensin tarkastellaan laajempaa strategiaa, sitten hyökkäyskulmaa, ja lopulta täytäntöönpanokeinoa eli konkreettista toteutustapaa. Seuraavat kappaleet tulevat käsittelemään tarkemmin luokiteltuja pimeitä malleja siten, että ensin määritellään ylätason strategia, kuten estäminen. Seuraavaksi siirrytään keskitason hyökkäyskulmaan, joka kuvaa tapaa, jolla strategia toteutetaan – esimerkiksi ylimääräisten vaiheiden luomista. Lopuksi analysoidaan alatason toteutustapa, kuten yksityisyyslabyrintti. Koska luokituksia on suuri määrä, tässä käsitellään tarkemmin vain yleisimpiä pimeiden mallien strategioita ja niiden ilmenemismuotoja.

Estäminen on strategia, jolla tarkoitetaan käyttäjän toiminnan vaikeuttamista esimerkiksi lisäämällä johonkin toimintaan turhia askeleita tai muuten yrittämällä estää käyttäjää tekemästä jotain hänelle hyödyllistä toimintoa [8], [10]. Estämisen keskitason luokan malleihin kuuluvat ”torakkahotelli”, ”esteiden luominen” sekä ”ylimääräisten askelten luominen”. Estämisen ”torakkahotelli” -pimeällä mallilla tarkoitettaan keskitason hyökkäyskulmaa, joka kuvaa palvelun toimintaa, jossa kumotaan käyttäjän odotus siitä, että jokin toiminto on yhtä helppoa peruuttaa kuin se on tehdä. Tämä luo tilanteen, johon on helppo mennä, mutta josta on vaikea päästä pois. ”Torakkahotellilla” on myös kaksi alatason strategiaa ”kuolemattomat tilit” ja ”umpikuja”. ”Kuolemattomat tilit” luovat ”torakkahotellin” käyttämällä estämis-

tä joka tekee luodun tilin poistamisesta mahdotonta tai haastavaa. Tällöin käyttäjät luovat tilin tai jakavat tietoja virheellisellä olettamuksella siitä, että he voisivat myöhemmin helposti poistaa tilin tai jaetun tiedon.

Piilottaminen on strategia, jossa piilotetaan, naamioidaan tai viivytetään sellaisten tärkeiden tietojen paljastamista, jotka, ilman että ne saatettaisiin käyttäjän näkyville, saisi käyttäjän tekemään tahattomia toimia, joita hän todennäköisesti muuten vastustaisi [8], [10]. Tämä voi käytännössä tarkoittaa esimerkiksi lisäkulujen, ylimääräisten palveluiden tai tuotteiden lisäämistä ilman käyttäjän suostumusta tai hänen huomaamattaan. Salaamisen keskitason luokkiin kuuluvat ”syötti ja vaihto”, ”tiedon piilottaminen” ja ”(de)kontekstualisoidut vihjeet”. Näistä esimerkiksi ”tiedon piilottamisen” yksi alaluokka on ”ostoskoriin piilottaminen”, jossa asiakkaan huomaamatta ostoskoriin lisätään jotain mitä hän ei itse ole lisännyt, ja joka voi johtaa haluamattomaan ostokseen.

Käyttöliittymähäirintä on strategia, jossa käyttöliittymää manipuloimalla annetaan tietyille toiminnoille etusija muihin nähden, mikä hämmentää käyttäjää tai rajoittaa merkityksellisten toimintamahdollisuuksien löytämistä [8], [10]. Käyttöliittymähäirinnän keskitason malleja ovat ”valinta-arkkitehtuurin manipulaatio”, ”huonot oletusasetukset”, ”emotionaalinen tai sensorillinen manipulaatio”, ”kompakysymys”, ”valinnallinen ylikuormitus”, ”piilotettu informaatio”, ”kielellinen esteellisyys” ja ”eteenpäin suuntautuva epäselvyys”. Näistä esimerkiksi ”huonoilla oletusasetuksilla” tarkoitetaan toteutustapaa tai hyökkäyskulmaa, jossa jo valmiiksi asetetut valinnat kumoavat käyttäjän odotuksen siitä, että ne ovat hänen etujensa mukaisia. Sen sijaan ne vaativat käyttäjää muuttamaan erikseen asetuksia jotka voivat aiheuttaa hänelle vahinkoa, kuten tietojen tahatonta jakamista. Tämä toteutustapa ilmenee erityisesti esimerkiksi evästebannereissa joiden helposti hyväksyttävät oletusasetukset ohjaavat käyttäjää jakamaan kaikki käyttäjätiedot kyseiselle verkkosivustolle tai palvelulle.

Pakotettu toiminta on strategia, joka edellyttää käyttäjää suorittamaan tietoisesti tai tiedostamattaan ylimääräisen ja/tai sivuavan toiminnon suorittaakseen toiminnallisuuden (tai jatkaakseen sitä) ja estää käyttäjää jatkamasta vuorovaikutusta järjestelmän kanssa ilman kyseisen toiminnan suorittamista [8], [10]. Tämän strategian keskitason hyökkäyskulmiin kuuluvat ”nalkuttaminen”, ”pakotettu jatkaminen”, ”pakotettu rekisteröityminen”, ”pakkoviestintä tai tiedon jakaminen”, ”pakotettu pelillistäminen” ja ”huomion kiinnittäminen”. Huomion kiinnittämisellä tarkoitetaan käyttäjän odotuksen kumoamista siitä, että hänellä olisi rationaalinen kontrolli siihen aikaan, jonka hän käyttää palveluun. Tämän yksi alaluokka on ”automaattinen toisto”. Aiemman selvityksen perusteella automaattinen toisto voitaisiin luokitella harmaaksi malliksi, mutta tässä tutkimuksessa pimeitä malleja sivuavia malleja ei ole käytetty.

Sosiaalinen manipulointi on strategia, jossa esitetään vaihtoehtoja tai tietoja, jotka saavat käyttäjän todennäköisemmin suorittamaan tietyn toimenpiteen yksilöllisten ja/tai sosiaalisten kognitiivisten ennakkoluulojen perusteella [8], [10]. Tämä strategia hyödyntää käyttäjän halua noudattaa odotettuja tai määrättyjä sosiaalisia normeja. Sosiaalisen manipuloinnin keskitason luokituksiin kuuluvat ”niukkuus ja suosion väite”, ”sosiaalinen todiste”, ”kiireellisyys”, ”häpeällistäminen” ja ”personointi”. Näistä ”sosiaalinen todiste” selitetään toimivan vastaavasti: ”Sosiaalinen todiste kumoaa käyttäjän odotukset siitä, että muiden osoittama käyttäytyminen tietyssä tilanteessa on oikein tai toivottavaa, ja sen sijaan nopeuttaa käyttäjän päätöksentekoa ja rohkaisee käyttäjää luottamaan annettujen tietojen virheellisiin vaikutuksiin.” Tämän alaluokan malli ”parasosiaalinen paine” on strategia, jossa käytetään sosiaalista näyttöä eräänlaisena sosiaalisen manipuloinnin muotona, jotta voidaan osoittaa, että julkkis, vaikuttaja tai muu taho, johon käyttäjä luottaa, on suositellut tuotetta tai palvelua, vaikka suosittelun lähde olisi puolueellinen tai harhaanjohtava. Tämän seurauksena käyttäjä voi olettaa, että suositus on tarkka ja puolueeton.

Hän saattaa ostaa tuotteen tai palvelun, vaikka kyseisen tahon mielipide tuotteesta olisi valheellinen tai täysin esitetystä eroava.

2.4 Pimeiden mallien vaikutukset

Iso osa pimeiden mallien kirjallisuudesta on keskittynyt pitkälti niitä kuvaileviin näkökohtiin. Näiden taustalla on kuitenkin joukko normatiivisia näkökohtia, jotka pyrkivät selittämään juuri miksi pimeiden mallien vaikutusten tulisi huolestuttaa meitä. Organisaatiot käyttävät pimeitä malleja, koska niillä pyritään saavuttamaan jotain vaikuttamalla vahingollisesti käyttäjiin. Se, mikä tekee pimeistä malleista ”pimeitä”, ovat niiden epäeettiset vaikutustavat sekä negatiiviset vaikutukset. Pimeiden mallien vaikutukset vaihtelevat niiden tyypin ja strategian mukaan, ja käyttäjien yksilöllisillä eroilla on myös huomattava vaikutus siihen, millaisina pimeät mallit koetaan. Mathur ym. [4] avasivat tutkimuksessaan *What Makes a Dark Pattern... Dark?* pimeiden mallien vaikutuksia neljän normatiivisen linssin kautta: yksilön hyvinvointi, kollektiivinen hyvinvointi, sääntelylliset tavoitteet ja yksilön itsemääräämisoikeudet.

2.4.1 Yksilön hyvinvointi

Yksilön hyvinvointi on normatiivinen linssi, joka tarkastelee pimeitä malleja sen perusteella, heikentävätkö ne käyttäjän hyvinvointia. Tämän näkökulman mukaan pimeät mallit voivat heikentää käyttäjän hyvinvointia muokkaamalla hänen valintarakennetta tavalla, joka suosii palveluntarjoajaa käyttäjän edun kustannuksella. Tutkimuksessa yksilön hyvinvoinnin heikkeneminen nähdään keskeisimpänä eettisenä ongelmana pimeissä malleissa. Ne voivat haitata käyttäjää esimerkiksi vahingoittamalla hänen etujaan, aiheuttamalla negatiivisia kokemuksia tai heikentämällä päätöksentekoa.

Yksinkertaisin esimerkki yksilön hyvinvoinnin huonontamisesta on rahallinen menetys. Verkkokaupat voivat aiheuttaa tämän kuluttajalle esimerkiksi lisäämällä ostoskoriin jotain ilman käyttäjän suostumusta, mikä voi johtaa käyttäjän kuluttamaan vahingossa enemmän rahaa kuin hän olisi alun perin tarkoittanut. Toinen yksilön hyvinvointiin vaikuttava asia on yksityisyyden loukkaaminen. Huonot ennakkoasetukset, kuten oletuksena valitut tietojenkeruuvaihtoehdot, tai suostumukseen häpeällistäminen (esimerkiksi hylkäysvaihtoehdon muotoileminen syyllistäväksi) voivat johtaa kuluttajaa antamaan enemmän informaatiota kuin hän oikeasti haluaisi jakaa. Jotkut näkemykset kuvaavat yksityisyyttä perusoikeudeksi ja yksilöllisen autonomian kulmakiveksi, joka tekee yksityisyyden häiritsemisestä hyvin huolestuttavaa. Kolmas yksilön hyvinvoinnin aspekti on kognitiivinen taakka. Tämän mukaan pimeät mallit johtavat kuluttajaa käyttämään tarpeetonta aikaa, energiaa ja huomiota. Tästä johtuva kognitiivinen rasitus voi johtaa käyttäjiä tekemään heille epäsuotuisia, helpompia valintoja jotka hyödyntävät palvelun tarkoitusperiä. [4]

2.4.2 Kollektiivinen hyvinvointi

Kollektiivinen hyvinvointi on pimeiden mallien kirjallisuudessa ollut vähemmän huomioitu kuin yksilön hyvinvointi. Tutkimuksessa tätä käsitellään yhteiskunnallisten ja markkinallisten vaikutusten kautta. Tämän mukaan pimeät mallit voivat heikentää kilpailua markkinoilla, erityisesti silloin, kun suuret toimijat väärinkäyttävät asemaansa manipuloimalla kuluttajan valintoja. Kun pimeät mallit heikentävät käyttäjien kykyä toimia rationaalisesti lisäämällä riippuvuutta tai manipuloimalla käyttäjää, voi tämä antaa alustoille mahdollisuuden kerätä varallisuutta ja vahvistaa markkina-asemaa ilman todellista ansiota. Hinnan läpinäkyvyys mahdollistaa kuluttajia tekemään informoituja päätöksiä ja luomaan toimivat markkinat. Hinnan läpinäkyvyyden hankaloittaminen esimerkiksi piilotetuilla kuluilla tai vertailun estämisellä voi johtaa rahalliseen menetykseen tai kilpailullisuuden vähenemiseen

markkinoilla. Luotto markkinoihin voi myös heikentyä pimeiden mallien takia, mistä erityisesti rehellisesti kilpailevat organisaatiot kärsivät. Käyttäjät voivat myös kärsiä luoton menetyksestä jos he esimerkiksi alkavat virheellisesti käsittää todellisia tarjouksia pimeiksi malleiksi. Pimeät mallit voivat lisäksi aiheuttaa odottamattomia yhteiskunnallisia vaikutuksia. Tästä esimerkkinä voidaan pitää Facebook:in tietojen keräämistä, jota Cambridge Analytica käytti hyväksi vaikuttaakseen USA:n presidentinvaaleihin vuonna 2016 [4].

2.4.3 Sääntelylliset tavoitteet

Sääntelyllisten tavoitteiden näkökulmasta pimeitä malleja tarkastellaan niiden vaikutuksen perusteella lainsäädäntöön. Tämän linssin mukaan pimeä malli on mikä tahansa käyttöliittymäratkaisu, joka muokkaa valinta-arkkitehtuuria siten, että se häiritsee tai heikentää laillisen sääntelyn tavoitteita. Sääntelyviranomaiset, kuten EU ja Yhdysvaltain Federal Trade Commission (FTC), soveltavat erilaisia lähestymistapoja pimeiden mallien torjumiseksi.

Yhdysvalloissa FTC:n valtuudet nojautuvat kuluttajansuojalakeihin, kuten Federal Trade Commission Actin 5. pykälään, jolla voidaan puuttua harhaanjohtaviin ja epäreiluihin käytäntöihin. EU:ssa puolestaan GDPR on keskeinen sääntelykeino, joka asettaa vaatimuksia käyttäjän suostumuksen läpinäkyvyydelle ja helppoudelle. Eri sääntelyjärjestelmien välillä on eroja: EU painottaa tarkkarajaisia lakeja, kun taas Yhdysvallat nojaa tapauskohtaiseen ja periaatteisiin perustuvaan arviointiin.

Viime vuosina pimeiden mallien sääntelyä on pyritty vahvistamaan uusilla aloitteilla, kuten Yhdysvaltain DETOUR Act -lakiehdotuksella[2], joka kieltäisi käyttöliittymät, joiden tarkoituksena on manipuloida käyttäjän päätöksentekoa. Lisäksi Kalifornian CPRA-laki kieltää tiettyjen yksityisyyteen liittyvien suostumusten hankkimisen pimeiden mallien avulla. [4] Tutkimuskirjallisuus on alkanut hiljattain tar-

[2]DETOUR: `https://www.scribd.com/document/405606873/Detour-Act-Final`

kastella pimeiden mallien laillisuutta suhteessa sääntelyyn, ja esimerkiksi GDPR:n mukaisten suostumushallintajärjestelmien lainmukaisuus on ollut usean tutkimuksen kohteena.

2.4.4 Yksilön itsemääräämisoikeudet

Hyvinvointi- ja sääntelynäkökulmiin katsottuna yksilön itsemääräämisoikeudet ovat yksilön oikeuksiin perustuva linssi. Tämän linssin perusteella pimeät mallit ovat käyttöliittymiä, jotka heikentävät käyttäjän yksilöllistä päätöksentekoa. Heikentyneen päätöksenteon johdosta käyttäjät voivat tehdä valintoja, joita he eivät olisi tehneet ilman muutettua päätösarkkitehtuuria. Yksilön itsemääräämisoikeutta loukkaavat pimeät mallit voivat estää käyttäjää tekemästä tiettyjä päätöksiä, piilottaa mahdollisia valintoja tai vaikeuttaa päätöksentekoa. Pimeät mallit, jotka vaikeuttavat itsemääräämisoikeutta muokkaamalla päätösarkkitehtuuria siten, että se aiheuttaa riippuvuutta, voivat saada aikaan negatiivisia vaikutuksia kuten psykologisia (huono keskittymiskyky) tai fyysisiä (unen laadun huononeminen) [4]. Kyseinen linssi on tärkeä, koska suurin osa pimeistä malleista vaikuttaa yksilön itsemääräämisoikeuteen jollain tavoin. Tämän normatiivisen linssin ongelmana on kuitenkin kuinka erottaa pimeät mallit, jotka vain rasittavat yksilön itsemääräämisoikeutta (esimerkiksi mainokset, jotka muuttavat kuluttajan käyttäytymistä) ja jotka vakavammin loukkaavat yksilön itsemääräämisoikeutta.

3 Pimeiden mallien tunnistusmenetelmät

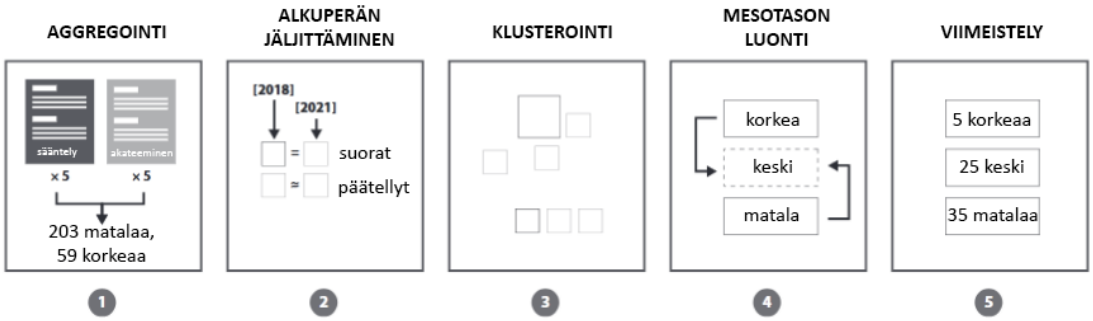
Pimeiden mallien tunnistamiseen ja ennaltaehkäisyyn on kehitetty useita menetelmiä, jotka voidaan jakaa manuaalisiin analyysimenetelmiin ja automaattisiin tunnistusratkaisuihin [11]. Manuaalisia menetelmiä ovat esimerkiksi käyttäjätestaus ja asiantuntija-analyysi, kun taas automaattiset ratkaisut perustuvat usein teksti- ja kuvantunnistukseen.

Viime vuosina tekoälyn ja koneoppimisen hyödyntäminen pimeiden mallien tunnistamisessa on yleistynyt tutkimuksissa. Pimeät mallit voivat ilmetä käyttöliittymissä strukturaalisina, visuaalisina tai tekstipohjaisina elementteinä, ja näiden havaitsemiseen on kehitetty useita näihin komponenttien tarkasteluun kohdistuvia lähestymistapoja. Lähestymistapoihin kuuluvat esimerkiksi LLM -mallien hyödyntämä tekstianalyysi, tekoälyavusteinen visuaalinen kuvantunnistus sekä käyttöliittymän lähdekoodin tarkastelu. Useimmat automaattiset tunnistusratkaisut hyödyntävät usein tekoälyä tai koneoppimista, mutta manuaalisia menetelmiä käytetään edelleen esimerkiksi tunnistusluokitusten laatimiseen ja automaattisten menetelmien validointiin.

3.1 Manuaaliset menetelmät

Manuaaliset menetelmät, kuten käyttäjätestit, heuristinen arviointi ja asiantuntija-analyysit, tarjoavat ihmislähtöisen tavan tunnistaa ja luokitella pimeitä malleja. Näiden menetelmien avulla voidaan tunnistaa sekä havaita monimutkaisia ja vaikeasti havaittavia pimeitä malleja, joita monet automaattiset ratkaisut eivät vielä kykene tehokkaasti tunnistamaan. Toisin kuin automaattisiin menetelmiin perustuvat ratkaisut, manuaaliset menetelmät eivät ole sidottuja tiettyyn alustaan, mikä mahdollistaa pimeiden mallien tunnistamisen riippumatta niiden ilmenemisympäristöstä. Manuaaliset menetelmät ovat kuitenkin aikaa vieviä ja resursseja vaativia, sillä ne edellyttävät asiantuntijoiden panosta eivätkä ole automatisoitavissa. Manuaalisten menetelmien avulla voidaan kuitenkin saavuttaa syvällisempi ymmärrys pimeiden mallien toiminnasta. Lisäksi ne mahdollistavat tunnistusluokitusten ja referenssilistojen luomisen, joita voidaan hyödyntää automaattisten menetelmien kehittämisessä ja validoinnissa.

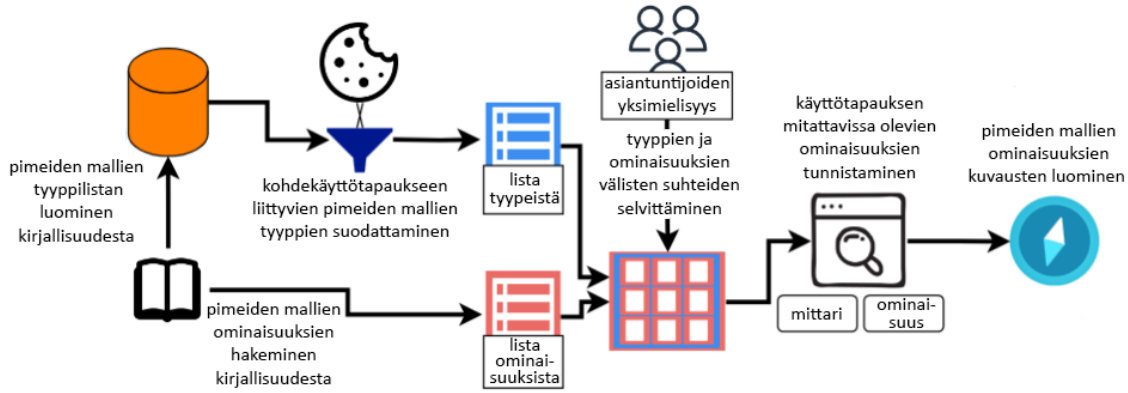
Grey ym. [8] käyttivät ontologiansa luomiseen manuaalisia menetelmiä. Menetelmän prosessi on esitetty kuvassa 3.1. Tutkimusryhmä koostui vakiintuneista pimeiden mallien tutkijoista, joista kaksi keskittyi ihmisen ja tietokoneen vuorovaikutukseen sekä muotoiluun, kolmas tietojenkäsittelytieteeseen, verkkoseurantaan ja sääntelyyn ja neljäs tietojenkäsittelytieteeseen sekä tietosuojalainsäädäntöön. Tutkimuksessa analysoitiin kymmentä aiempaa pimeiden mallien taksonomiaa, joista tunnistettiin 245 erillistä pimeiden mallien luokitusta. Kesällä 2023 päivitetyistä lähteistä lisättiin 17 uutta mallia, jolloin lopullinen analyysi kattoi 262 mallia (kuvan kohta 1). Mallien yhdistämisen jälkeen niiden alkuperä kartoitettiin suorien viittausten ja sisällöllisten yhtäläisyyksien perusteella (kuvan kohta 2). Tämän jälkeen samankaltaiset mallit ryhmiteltiin, minkä pohjalta luotiin kolmitasoinen ontologia (kuvan kohdat 3 ja 4). Lopuksi ontologia viimeisteltiin yhdenmukaistamalla se, sekä validoimalla se monitieteisessä tutkijayhteisössä (kuvan kohta 5).



Kuva 3.1: Ontologian luomisprosessi [8]

Pimeiden mallien tunnistamisprosessia on vaikeuttanut puute niiden tarkasta ja objektiivisesta määritelmästä, joka olisi johdannainen niiden mitattavista ominaisuuksista. Kocyigit ym. [12] pyrkivät tutkimuksessaan kehittämään mitattavia ominaisuuksia manuaalisin menetelmin, joiden avulla pimeiden mallien tunnistaminen voisi olla luotettavampaa ja vähemmän subjektiivista. Heidän menetelmänsä perustuu ihmisen ja tietokoneen vuorovaikutuksen (HCI) analysointiin. Tutkimuksessa he loivat taulukon, jossa he luokittelevat pimeiden mallien tyypit niiden ominaisuuksien perusteella. He lähestyivät ongelmaa tunnistamalla ensin evästebannereissa esiintyvät pimeiden mallien tyypit ja suodattamalla niistä asiaankuuluvat tapaukset asiantuntija-arvioiden perusteella. Tämän jälkeen he analysoivat jokaisen mallin kuuden valitun ominaisuuden näkökulmasta ja pyrkivät saavuttamaan yksimielisyyden niiden luokittelussa. Kyseinen lähestymistapa on esitetty kuvassa 3.2.

Tutkimuksessa määriteltiin mitattavat ominaisuudet entiteetti-metriikka -mallin avulla, ja nämä ominaisuudet jaoteltiin kolmeen luokkaan: visuaaliset (esim. painikkeen koko), kielelliset (esim. tekstin luettavuus) ja prosessuaaliset (esim. käyttäjän tekemien klikkausten määrä). Lopuksi pimeiden mallien ominaisuudet määriteltiin loogisina sääntöinä, jotka yhdistävät nämä mitattavat ominaisuudet ja vähentävät subjektiivisuuden määrää niiden tunnistamisprosesseissa.



Kuva 3.2: Pimeiden mallien luokittelu tyyppien ja ominaisuuksien perusteella [12]

3.2 Automaattiset menetelmät

Automaattiset menetelmät kuten koneoppimiseen, luonnollisen kielen käsittelyyn (NLP) ja kuvantunnistukseen perustuvat ratkaisut ovat automatisoituja, manuaalisiin menetelmiin verrattuna objektiivisempia menetelmiä pimeiden mallien tunnistamiseen. Pimeiden mallien tunnistamisen prosessi sisältää usein kohdennetun tunnistusalustan, listan tunnistettavista pimeiden mallien tyypeistä sekä menetelmät joilla niitä tunnistetaan. Keskeisessä asemassa tunnistuksessa on alustalla käytettyjen pimeiden mallien luokkiin sijoittelu, joten automaattiset menetelmät aloittavat usein esittelemällä mitä luokkia ne pyrkivät tunnistamaan ja millä alustoilla kyseistä ratkaisua on tarkoitus käyttää.

Khanna ym. [13] esittelevät tutkimuksessaan *Detecting Deception: An AI-Driven Approach to Identify Dark Patterns* tunnistusmenetelmissä käytettyjä mittareita. Käytettyjä mittareita ovat Tarkkuus (Accuracy), Osumatarkkuus (Precision), Herkkyys (Recall), F1-arvo (F1-score) sekä ROC/AUC-pisteytys (Receiver Operating Characteristic ja Area Under the Curve). Näiden mittareiden tarkka ymmärtäminen ei ole välttämätöntä käytettyjen automaattisten menetelmien hahmottamisek-

si, mutta niiden selittäminen auttaa ymmärtämään menetelmien ominaisuuksia ja tehokkuutta.

- Tarkkuus (Accuracy) mittaa mallin yleistä oikeellisuutta vertaamalla oikein ennustettujen tapausten määrää esiintyneiden pimeiden mallien kokonaismäärään. Korkea tarkkuusarvo tarkoittaa, että malli ennustaa kaikki pimeiden mallien luokat tehokkaasti.

- Osumatarkkuus (Precision) määrittää positiivisten ennusteiden tarkkuuden vertaamalla todellisten positiivisten tapausten määrää kaikkien positiivisiksi ennustettujen tapausten määrään (todetut ja virheelliset positiiviset). Tämä mittari on erityisen hyödyllinen, kun väärien positiivisten minimointi on tärkeää.

- Herkkys (Recall) mittaa mallin kykyä tunnistaa kaikki positiiviset tapaukset laskemalla todellisten positiivisten tapausten osuus suhteessa kaikkiin oikeisiin positiivisiin ja virheellisiin positiivisiin tapauksiin. Herkkyys on tärkeä silloin, kun kaikki positiiviset tapaukset halutaan tunnistaa mahdollisimman kattavasti.

- F1-arvo (F1-score) on osumatarkkuuden ja herkkyyden harmoninen keskiarvo, joka tasapainottaa mallin suorituskyvyn. F1-pisteytys on hyödyllinen erityisesti silloin, kun positiivisten ja negatiivisten tapausten määrät ovat epätasapainossa ja väärien positiivisten sekä väärien negatiivisten minimointi on tärkeää.

- ROC ja AUC -arvot mittaavat mallin kykyä erotella eri pimeiden mallien luokat toisistaan eri kynnysarvoilla. ROC/AUC -arvo mittaa mallin suorituskykyä eri herkkyys- ja tarkkuustasoilla, ja ne ovat kokonaisvaltainen arvio mallin ennustekyvystä.

Tutkimuksen päätarkoituksena oli kehittää tekoälyyn pohjautuva järjestelmä, joka tunnistaa pimeitä malleja hyödyntäen luonnollisen kielen käsittelyä sekä koneoppimista. Järjestelmä analysoi alustoja tunnistaakseen piirteitä, jotka viittaavat harhaanjohtaviin tai pakottaviin suunnitteluratkaisuihin, kuten manipuloiviin käyttäjäkehotteisiin, harhaanjohtaviin visuaalisiin vihjeisiin tai epäselvään kieleen.

Menetelmän ensimmäisessä vaiheessa suoritettiin laajamittainen datan keruu verkkosivustoilta ja sovelluksista. Keruu toteutettiin hyödyntämällä avoimia tietolähteitä sekä verkkosivujen haravointitekniikoita. Esikäsittelyvaiheessa kerätystä datasta poistettiin kohina, normalisoitiin aineisto ja suoritettiin piirteiden poiminta. Näiden pohjalta tunnistettiin käyttäjäinteraktioita, tekstisisältöjä ja visuaalisia elementtejä, joita pimeät mallit hyödyntävät.

Tunnistusjärjestelmän pääosassa olivat koneoppimismallit, joilla analysoitiin ja tunnistettiin 14:sta eri pimeää mallia. Mallien koulutuksessa hyödynnettiin laajaa aineistoa, ja luokitusten hyperparametrien optimointiin käytettiin ruudukkohaku -menetelmää. Lopuksi järjestelmän suorituskykyä arvioitiin tarkkuuden, osumatarkkuuden, herkkyyden ja F1-arvon perusteella ja nämä tulokset on esitelty taulukossa 3.1. Tutkimuksessa kokeiltiin useita koneoppimismalleja, joista tehokkaimmiksi osoittautuivat tarkkuuden perusteella logistinen regressio, tukivektorikone (SVM) ja satunnaismetsä. Näiden mallien erottelukyky oli erinomainen, mikä näkyi niiden ROC/AUC -arvoissa [13].

Taulukko 3.1: Pimeiden/ei-pimeiden mallien luokittelutulokset [12]

Luokittelija	Tarkkuus	Osumatarkkuus	Herkkyys	F1	ROC	AUC
Multinominaalinen Naive Bayes	0.95	0.94	0.95	0.95	0.997	0.93
Logistinen regressio	0.97	0.97	0.97	0.97	0.997	0.95
SVM	0.96	0.96	0.96	0.96	0.997	0.94
Satunnaismetsä	0.96	0.95	0.96	0.95	0.996	0.94
Gradientin lisääminen	0.93	0.93	0.93	0.93	0.991	0.89
KNN	0.95	0.95	0.95	0.95	0.985	0.92
Päätöspuu	0.94	0.94	0.94	0.94	0.960	0.91

3.3 Haasteet ja ongelmat

Pimeitä malleja on laaja ja jatkuvasti kasvava määrä, ja niitä esiintyy useilla eri alustoilla, kuten verkkosivuilla, mobiilisovelluksissa, IoT-laitteissa ja jopa kotiroboteissa [9], [12]. Tämä monimuotoisuus luo merkittäviä haasteita yksittäisen työkalun kehittämiselle, joka kykenisi tunnistamaan pimeät mallit kaikilla näillä alustoilla. Nykyiset tunnistusmenetelmät rajoittuvat usein vain tiettyihin pimeiden mallien tyyppeihin ja ne ovat käytettävissä vain tietyillä valituilla alustoilla. Nykyisten tutkimusten tunnistusmenetelmät esimerkiksi tarkastelevat lähes jokaisessa tapauksessa pelkästään verkkosivustoja tai mobiilisovelluksia. Lisäksi kehitettyjen tunnistusmenetelmien pohjalta tulisi aloittaa vapaasti käytettävien työkalujen julkaiseminen.

Tällä hetkellä lupaavin alusta pimeiden mallien tunnistamiseen ovat verkkosivut, sillä niiden lähdekoodia voidaan analysoida, ja selainpohjaiset työkalut mahdollistavat tehokkaan havainnoinnin. Pimeät mallit voivat kuitenkin esiintyä eri käyttöliittymän osioissa, kuten rekisteröintiprosesseissa, evästeilmoituksissa tai syvemmin integroituna alustan toiminnallisuuteen, mikä luo lisähaasteita niiden tunnistamiselle ja torjunnalle. On myös esitetty hypoteesi, jonka mukaan tulevaisuudessa tekoäly voisi alkaa luomaan pimeitä malleja käyttöliittymäsuunnittelijoiden sijaan, mikä vaikeuttaisi niiden tunnistamista tulevaisuudessa yhä enemmän [14].

4 Yhteenveto ja pohdintaa

Tässä tutkimuksessa selvitettiin pimeiden mallien luokitusta, niiden vaikutuksia sekä niiden tunnistamiseen kehitettyjä menetelmiä. Luokittelu tapahtui kolmetasoisen ontologian avulla ja vaikutuksia tarkasteltiin neljän normatiivisen linssin läpi. Tunnistusmenetelmistä tarkasteltiin manuaalisia ja automaattisia ratkaisuja. Lopuksi tässä luvussa vastataan vielä tutkimuskysymyksiin ja esitetään yleistä aiheeseen liittyvää pohdintaa.

Ensimmäinen tutkimuskysymys oli: Kuinka pimeitä malleja voidaan luokitella? Pimeiden mallien määrittelyssä ja luokittelussa ei ole vielä saavutettu tarkkaa yhteisymmärrystä, mutta useat tutkimukset ovat pyrkineet kehittämään systemaattisia luokittelumalleja. Erityisesti Mathur ym. [4] tarjoavat vankan pohjan yhtenäiselle kolmitasoiselle ontologiselle luokittelumallille, jossa pimeät mallit jaotellaan eri kategorioihin niiden käyttötarkoituksen ja vaikutusmekanismien perusteella.

Toisen tutkimuskysymyksen, "Mitä vaikutuksia pimeillä malleilla on?", pohjalta voidaan todeta, että pimeät mallit vaikuttavat käyttäjiin lähes poikkeuksetta negatiivisesti. Ne voivat johtaa tahattomiin valintoihin, heikentää yksityisyyttä, aiheuttaa taloudellisia menetyksiä sekä vähentää luottamusta digitaalisiin palveluihin ja markkinoihin. Pimeiden mallien vaikutuksien ennaltaehkäisemiseksi on myös kehitetty joitain lakeja, mutta ne seuraavat usein pimeiden mallien nopean kehityksen sekä tutkimustyön jäljessä.

Kolmanteen tutkimuskysymykseen, "Mitä menetelmiä pimeiden mallien tunnistamiseen on kehitetty?", voidaan tiivistäen vastata, että pimeiden mallien tunnistamiseen on kehitetty sekä manuaalisia että automaattisia menetelmiä. Näiden menetelmien avulla voidaan tunnistaa valittujen pimeiden mallien tyyppejä melko luotettavasti. Manuaaliset menetelmät perustuvat asiantuntijoiden suorittamaan käyttöliittymäanalyysiin, kun taas automaattiset menetelmät hyödyntävät koneoppimista ja luonnollisen kielen käsittelyä pimeiden mallien havaitsemiseen. Tunnistusmenetelmien haasteena ovat kuitenkin pimeiden mallien jatkuva kehitys ja tästä johtuva uusien tyyppien ilmaantuminen. Tunnistusmenetelmät ovat myös usein rajattuja tietylle alustalle eikä vapaasti käytettäviä sekä luotettavia tunnistustyökaluja ole paljoa yksityishenkilöiden saatavilla. Tulevaisuuden uhkana ovat myös tekoälyn kehittämät pimeät mallit joiden tunnistaminen sekä ennaltaehkäiseminen voi olla aiempaa haastavampaa [14].

Pimeät mallit ovat nopeasti kasvava ilmiö ja uhka ihmisten hyvinvoinnille. Niihin liittyvät eettiset kysymykset tulevat nousemaan yhä tärkeämmiksi, kun niiden käyttö laajenee entistä useammille aloille. Yhä useampi yritys huomaa niiden tarjoamat taloudelliset hyödyt, ja on vaarana, että niiden käyttö vakiintuu - kuten on jo tapahtunut esimerkiksi mobiilipelimarkkinoilla [15]. Tämä herättää kuitenkin kysymyksen siitä, kenellä on perimmäinen vastuu pimeiden mallien vastaisessa taistelussa. Pitäisikö kuluttajan tehdä enemmän informoituja päätöksiä, vai tulisiko yritysten sitoutua eettisten suunnittelumallien käyttöön? Vai olisiko vaihtoehtoinen ratkaisu kehittää tiukempaa käyttöliittymäsuunnitteluun liittyvää lainsäädäntöä? Lakien ongelmana on kuitenkin se, että niiden kehitys on usein teknologiaa hitaampaa, ja yritykset pyrkivät silti usein kiertämään sääntelyä. Mahdollisesti merkittävin keino ilmiön torjumiseksi tällä hetkellä on aiheesta tietoisuuden lisääminen - tavoite, jota myös tämä tutkielma pyrkii edistämään.

Tämän kandidaattitutkimuksen merkittävimpinä haasteina olivat aihealueen tarkka rajaaminen sekä keskeisten lähteiden valinta. Tutkielman aikana käytiin läpi kymmeniä eri tutkimuksia, ja aihe vaihtui tämän myötä muutamaan kertaan. Alkuvaiheessa tutkimuksen aihealue oli hyvin laaja ja epämääräinen, minkä vuoksi etsittyjä lähteitä oli suuri määrä ja ne käsittelivät eri näkökulmia. Tämä tutkimus käsitteli myös vain muutamia tunnistusmenetelmiä, joten jatkotutkimusta aiheesta voisi tehdä esimerkiksi vertailemalla useampien eri menetelmien tehokkuutta ja soveltuvuutta keskenään. Toinen merkittävä jatkotutkimusaihe voisi olla pimeiden mallien vaikutusten mittaaminen käyttäjien hyvinvointiin. Tässä yhteydessä voitaisiin tutkia esimerkiksi, kuinka käyttäjän stressitaso, päätöksentekokyky tai tunnetila muuttuvat altistuessa pimeille malleille.

Lähdeluettelo

[1] N. Zingales, ”19 Dark Patterns”, *PLATFORM*, s. 87,

[2] S. L. Pauwels, C. Hübscher, J. A. Bargas-Avila ja K. Opwis, ”Building an interaction design pattern language: A case study”, *Computers in Human Behavior*, vol. 26, nro 3, s. 452–463, 2010, ISSN: 0747-5632.

[3] A. Seffah, ”The evolution of design patterns in HCI: from pattern languages to pattern-oriented design”, sarja PEICS ’10, Berlin, Germany: Association for Computing Machinery, 2010, ISBN: 9781450302463.

[4] A. Mathur, M. Kshirsagar ja J. Mayer, ”What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods”, sarja CHI ’21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966.

[5] L. Alberts, U. Lyngs ja M. Van Kleek, ”Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially”, vol. 8, nro CSCW1, huhtikuu 2024.

[6] J. Luguri ja L. J. Strahilevitz, ”SHINING A LIGHT ON DARK PATTERNS”, *JOURNAL OF LEGAL ANALYSIS*, vol. 13, s. 43–109, maaliskuu 2021, ISSN: 2161-7201.

[7] T. Montefiore ja P. Formosa, ”Dark Patterns Meet the Gamer’s Dilemma: Contrasting Morally Objectionable Content with Systems in Video Games”,

[8] C. M. Gray, C. T. Santos, N. Bielova ja T. Mildner, ”An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building”, teoksessa *PROCEEDINGS OF THE 2024 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYTEMS (CHI 2024)*, ACM CHI Conference on Human Factors in Computing Sytems (CHI), Honolulu, HI, MAY 11-16, 2024, Assoc Comp Machinery; ACM SIGCHI; Apple; Google; NSF; Tianqiao & Chrissy Chen Inst, 2024, ISBN: 979-8-4007-0330-0.

[9] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog ja C. Wilson, ”A Comparative Study of Dark Patterns Across Web and Mobile Modalities”, *Proc. ACM Hum.-Comput. Interact.*, vol. 5, nro CSCW2, lokakuu 2021.

[10] A. Mathur, G. Acar, M. J. Friedman et al., ”Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites”, *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, nro CSCW, s. 1–32, marraskuu 2019, ISSN: 2573-0142.

[11] A. Bajaj, K. Uppal, R. Razdan, Y. Tuteja, A. Bhardwaj ja A. Abraham, ”A Comprehensive Analysis for Dark Pattern Detection Using Structural, Visual and Textual Information”, *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 17, s. 12, tammikuu 2025.

[12] E. Kocyigit, A. Rossi ja G. Lenzini, ”A Systematic Approach for A Reliable Detection of Deceptive Design Patterns Through Measurable HCI Features”, sarja EuroUSEC ’24, Association for Computing Machinery, 2024, s. 290–308, ISBN: 9798400717963.

[13] S. T. Khanna, ”Detecting Deception: An AI-Driven Approach to Identify Dark Patterns”, *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, nro 4, s. 1658–1669, kesäkuu 2024.

[14] H. R. Sangaraju, S. Waris, A. Salina, D. Chandra Jadala ja G. Rao, ”Smart Dark Pattern Detection: Making Aware of Misleading Patterns Through the Intended App”, teoksessa tammikuu 2022, s. 933–947, ISBN: 978-981-16-5156-4.

[15] S. Niknejad, T. Mildner, N. Zargham, S. Putze ja R. Malaka, ”Level Up or Game Over: Exploring How Dark Patterns Shape Mobile Games”, teoksessa *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*, sarja MUM ’24, Association for Computing Machinery, 2024, ISBN: 9798400712838.