Joonatan Tikka Tuukkanen

TEKOÄLY, TEKIJÄNOIKEUS, FAIR LEARNING JA TDM

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2023

TIIVISTELMÄ

Tikka Tuukkanen, Joonatan
Tekoäly, tekijänoikeus, fair learning ja TDM
Jyväskylä: Jyväskylän yliopisto, 2023, 27 s.
Tietojärjestelmätiede, kirjallisuuskatsaus
Ohjaaja(t): Vuorinen, Jukka

Tekoälyn (AI) leviäminen moniin eri elämän osa-alueisiin nostaa esiin kysymyksiä, jotka aiheuttavat vahvaa diskurssia tietotekniikka- ja lakipiireissä. Tässä kirjallisuuskatsauksessa tutkittiin sitä, saako ja tulisiko tulevaisuudessakin saada käyttää tekijänoikeudella suojeltuja teoksia tekoälyn opettamiseen ja kenelle tekoälyllä generoitujen teosten tekijänoikeudet kuuluvat. Koska tekoäly on edistynyt harppauksin viime aikoina, ovat nämä kysymykset tärkeää selvittää, sillä maat ovat jo alkaneet luomaan lakeja, joilla voi olla suuret vaikutukset tekoälyn tulevaisuudelle. Siksi on tärkeää lähestyä näitä kysymyksiä laskelmoivalla, varovaisella, reilulla ja rationaalisella lähestymistavalla. Tämä kirjallisuuskatsaus tähtäsi avaamaan tekoälyn, TDM:n, tekijänoikeuksien ja fair learningin risteymää. Fair learning on Lemleyn ja Caseyn (2021) luoma konsepti, joka takaa oikeuden oppia jokaiselle, jopa roboteille. Tämä kirjallisuuskatsaus keskittyi tarkemmin siihen, päteekö fair use tekoälyyn, joka käyttää tekijänoikeudella suojeltuja teoksia. Myös tekoälyllä generoitujen teosten tekijänoikeuksiin ja fair usen liittyviin USA:n oikeuskäytänteisiin ja lakipäätöksiin paneuduttiin tarkemmin. Tutkimuksen pääkysymyksiä oli kaksi, joista ensimmäinen oli: Tulisiko tekoälyä opettaa tekijänoikeudella suojattujen teosten avulla ja tulisiko tämänkaltainen TDM:n (text and data mining) käyttö sallia kaikille, eikä pelkästään tietyille tahoille? Toinen kysymys oli: Tulisiko tekoälyllä generoidulle teokselle suoda tekijänoikeus ja kenelle tämä tekijänoikeus suotaisiin? Tässä tutkielmassa kävi ilmi, että tutkijoiden mielestä TDM:ää tulisi saada käyttää vapaasti tutkimuksiin. Useiden tutkijoiden mielestä kaikkien tulisi saada käyttää tekijänoikeuksilla suojattua materiaalia tekoälyn opettamiseen ja TDM:ään. Tekoälyllä generoitujen teosten tekijänoikeusongelma oli enemmän tutkijoita jakava kysymys. Osa tutkijoista oli sitä mieltä, että tekoälyn luomille teoksille ei koskaan tulisi saada tekijänoikeutta, kun taas toiset sitä mieltä, että tekijänoikeus kuuluisi tekoälyn käyttäjälle. Tämä vastakkainasettelu osoitti, että aiheesta tulisi tehdä jatkotutkimuksia. Tämä tutkielma keskittyi pääasiassa USA:han ja EU:hun, mutta myös vähäisesti Iso-Britanniaan ja muihin maihin.

Asiasanat: AI, copyright, fair use, fair learning, tekoäly, tekijänoikeus, TDM.

ABSTRACT

Tikka Tuukkanen, Joonatan
Artificial intelligence, copyright, fair learning, and TDM
Jyväskylä: University of Jyväskylä, 2023, 27 pp.
Information Systems, literature review
Supervisor(s): Vuorinen, Jukka

As artificial intelligence (AI) becomes more prevalent in various parts of modern life, it raises concerns about the training of AI with copyrighted material. Another question that perplexes IT and copyright law scholars is: who owns the works generated by AI? As AI has progressed in massive leaps in the past few years, these questions have become important topics to tackle, as countries have already started to make laws that could have major implications on the future of AI. It is therefore important to find a balanced way to react to these recent technologies, not with haste, but with a calculated, fair, and rational approach. This systematic literature review and case study analysis aimed to explore the intersection of AI, copyright, and fair learning, which is a principle coined by Lemley and Casey (2021) that ensures the right to learn for everyone, even for AI. Specifically, this thesis investigated the extent to how fair use applies to use of generative AI that is trained on copyrighted works, the jurisprudence of US Courts on uses of copyrighted works, and the potential of fair use and learning to address these copyright concerns. The main questions that this thesis presented were: should artificial intelligence be taught using copyrighted works, and should this kind of text and data mining (TDM) usage be allowed for everyone, not just certain parties, and should a copyright be granted to a work generated by artificial intelligence, and to whom would this copyright be granted? This study found out that researchers believe TDM should be freely used for research purposes. According to several studies, everyone should have the right to use copyrighted material for TDM and teaching artificial intelligence. The issue of copyright for works generated by artificial intelligence was more divisive among researchers. Some researchers believed that AI-generated works should never be granted copyright, while others believed that copyright should belong to the user of the artificial intelligence. This discord highlighted the need for further research on the subject. This thesis focused majorly on the USA and EU, and minorly on the UK, as each of them offered a unique look into the ways in which countries can react to emerging technologies.

Keywords: AI, artificial intelligence, copyright, fair learning, fair use, TDM.

TAULUKOT

SISÄLLYS

1 JOHDANTO

Tekoäly on pitkän, hitaan ja epävarman kehityksen jälkeen edennyt harppauksin ja leviää nyt äärimmäisen voimakkaasti moneen eri yhteiskunnan osa-alueeseen. Tekoälyllä voidaan saavuttaa asioita, jotka ovat aikaisemmin olleet pelkkää tieteisfiktiota. Generoivilla tekoälymalleilla voidaan luoda edellä mainittuja tieteisfiktiota, vaikka jonkun tietyn kuuluisan kirjailijan tyylillä. Generoivan tekoälyn avulla voidaan luoda scifi-seikkailu esimerkiksi Edgar Allan Poen tyylillä (Slater, 2023). Tehdylle novellille voitaisiin luoda kuvituskuvat kuvia luovalla tekoälyllä ja myös äänikirja pelkästään syöttämällä teksti tekoälymallille (Guinness, 2023; Rangwala, 2023). Kirjan markkinointistrategian voisi kehittää myös tekoäly (Jackson, 2023). Tekoälyä voidaan käyttää myös muuhunkin kuin teosten tekemiseen. Tekoälyllä pystyttiin löytämään uusi vastetta luova lääke maksasyöpään pelkästään 30 päivässä (Ren ym., 2023). Sen avulla pystytään myös huomattavasti parantamaan arviointia yrityksen rahallisesta menestyksestä pelkästään kokousten osallistujien äänensävyn perusteella (Yang ym., 2023) ja vähentämään epätosien uutisten määrää (Li ym., 2022). Tekoälyllä on myös paljon muita ihmiskunnalle hyödyllisiä käyttötapoja.

Tätä kaikkea vaikeuttaa tällä hetkellä ympäri maailmaa leviävä tekijänoikeuslakien kiristäminen, väärinkäyttö ja tulkinta. World Intellectual Property Organizationin ja Netanelien mukaan tekijänoikeudella ei ole tarkoitus estää tiedon leviämistä yhteiskunnassa, vaan sen tarkoitus on juuri päinvastainen (Netanel, 2008, s. 11; WIPO, 2016).

Tässä tutkielmassa termiä tekoäly käytetään kuvaamaan kaikkia tekoälyn alla olevia malleja ja teknologioita, eikä tämä tutkielma keskity itse tekoälyn tai sen alla olevien teknologioiden matemaattisiin toimintaperiaatteisiin. Tutkimuksessa käytetään myös selkeyden vuoksi termiä copyright, kun puhutaan USA:sta, ja tekijänoikeus, kun puhutaan muiden maiden tekijänoikeuksista.

Tekoälyn opettaminen tekijänoikeuksien alla olevien teosten avulla on ollut keskustelua herättävä aihe tiede- ja populaariyhteisöissä. Toisin kuin generoivien tekoälyjen teosten tekijänoikeuksista tästä ongelmasta ollaan tiedeyhteisössä enemmän samaa mieltä. Osa tutkijoista (Gervais, 2020, 2021 & 2023; Bridy, 2019) on sitä mieltä, että tekoälyn opettaminen tekijänoikeuksien alla olevilla

teoksilla tulisi olla sallittua tutkimuksia varten, ja osa on sitä mieltä, että sen tulisi olla sallittua kaikkeen käyttöön, kunhan fair usea ei rikota (Levendowski, 2018; Lemley & Casey 2021; Franceschelli & Musolesi, 2022; Koščík, 2020; Kop, 2019 & 2020, March). Fair use on USA:n tekijänoikeuslainsäädännön käsite, jonka avulla tekijänoikeuksien suojaamia teoksia voidaan käyttää eri tavoin tietyissä tilanteissa ja tämä kirjallisuuskatsaus käsittelee fair usea luvussa 3.2. Euroopan unionissa ei samankaltaista kaikenkattavaa käsitettä ole, vaan EU-mailla on omat tekijänoikeuspoikkeuksia koskevat lait, esimerkiksi Suomessa sitaattioikeudet, yksittäiskäyttöoikeudet ja opetukseen liittyvät oikeudet antavat tietynlaisen oikeuden uudelleen käyttää tekijänoikeudella suojattua materiaalia. EU on kuitenkin aikaisemmin rajoittanut edellä mainittua opettamista erittäin laajoilla ja ristiriitaisilla direktiiveillä (Margoni & Kretschmer, 2022; Kop, 2020; Directive 96/9/EC; Directive 2019/790). TDM-rajoitukset (Text and Data Mining) ovat myös suoraan vaikuttaneet EU:n sisällä tehtyjen tutkimusten määrään (Handke, Guibault, Vallbé, 2015**)** ja myös nämä rajoitukset ovat estäneet EU:n sisällä toimivien tekoäly-yritysten toimintaa (Margoni & Kretschmer, 2022).

Datan suuri määrä on vitaalia koneoppiville malleille. Huono tai vajaa määräinen data voi aiheuttaa vinoumia lopputulemissa (Kordzadeh & Ghasemaghaei, 2021). Nämä vinoutumat voivat olla joissain tilanteissa harmittomia, kun taas joissain ne voivat olla ihmishengille vaaraksi. Tällaisia tilanteita voivat esimerkiksi olla: koira, joka luokitellaan kissaksi, vanki, jota ei päästetä rasististen tekoälymallin takia ehdolliseen, autoon, joka ajaa ihmisen päältä, koska se ei tunnistanut moottoripyöräilevää henkilöä. (Ganguli ym., 2022; Lemley & Casey 2021; CBS News, 2022). On siis tärkeää, että koneoppivat mallit saavat kaiken tarvittavan datan. On myös tärkeää, että dataa voidaan tutkia ihmissilmin ja muokata sitä tarvittavassa määrin (esim. oversampling vähemmistöistä, sovinistisien ja rasististen lopputulemien poistaminen), jotta nämä vinotumat saadaan poistettua. EU:n TDM-direktiivien (Text and Data Mining) hankaloittavat datan verifiointia, tutkimusten toistettavuutta ja tutkimustulosten verifiointia (Margoni & Kretschmer, 2022).

Generoivalla tekoälyllä tuotettujen teosten tekijänoikeudet ovat myös puhuttaneet taide-, laki- ja teknologiayhteisöjä. Tällä tutkimuksella yritetään käydä läpi tärkeitä kohtia kyseisestä aiheesta, ja varsinkin sitä kenelle tekoälyllä generoitujen teosten tekijänoikeudet kuuluisivat.

Tällä tutkielmalla avataan siis tekoälyn, copyright/tekijänoikeuksien, TDM:n ja fair usen risteymää. Tutkielma painottuu TDM:n käyttöön, tekoälyn opettamiseen tekijänoikeudella suojatuilla teoksilla ja tekoälyllä generoitujen teosten tekijänoikeuksiin. Tutkielma tutkii myös sitä, miten copyright lainsäädännön tiukentaminen voi vaikuttaa negatiivisesti innovaatioon, tieteen edistämiseen ja miten tämä säännösten tiukentaminen edistäisi pääasiassa isojen yritysten etuja. Tutkimuksessa myös vertaillaan joidenkin maiden etuasemaa tekoälyn kehittämisessä, minkä kyseiset maat ovat saaneet vapaimmilla tekijänoikeuslailla ja miten tiukemmat tekijänoikeuslait johtavat yleensä heikompaan ja mahdollisesti vinoutuneisiin tekoälyn lopputulemiin.

Tämän kirjallisuuskatsauksen pääkysymykset ovat siis: tulisiko tekoälyä opettaa tekijänoikeudella suojattujen teosten avulla ja tulisiko TDM:n käyttö sallia kaikille, eikä pelkästään tietyille osapuolille, ja tulisiko tekoälyllä generoidulle teokselle suoda tekijänoikeus ja kenelle tämä tekijänoikeus suotaisiin? Tämä tutkielma keskittyy pääasiassa USA:han ja EU:hun, ja vähäisesti Iso-Britanniaan, sillä jokainen näistä hallintoalueista (jurisdiction) tarjoaa uniikin tavan reagoida uusiin teknologioihin.

Tämä tutkielma suoritetaan systemaattisena kirjallisuuskatsauksena. Tutkielma käyttää materiaalina pääosin artikkeleita, jotka ovat julkaistu korkeasti JUFO- luokitelluissa tieteellisissä aikakausilehdissä. JUFO on suomalainen tieteellisten lehtien ja konferenssien laadunarviointia suorittava luokittelualusta suomalaisesta näkökulmasta. Osa artikkeleista on myös julkaistu tietojärjestelmätieteen arvostetuimmissa tieteellisissä aikakausilehdissä, Basket of Eightissä. Artikkeleita on etsitty eri tietokannoista, kuten JYKDOK, Google Scholar ja Oxford Academic. Hakusanoina on käytetty termejä: AI, artificial intelligence, copyright, fair use, fair learning, TDM, text and data mining, machine learning, deep learning, DL, ML ja EU. Osa lähteistä on myös löytynyt aikaisemmin löydettyjen artikkelien lähteistä.

Tämän aiheen tutkimus on tärkeä, sillä seuraavien vuosien tekoälyyn vaikuttavien lakien ja direktiivien vaikutukset tulevat muuttamaan koko tekoälyn kehitystä ja tulevaisuutta. On siten tärkeää, että päätökset laeista ovat tehty tieteellisen tutkimuksen pohjalta ja objektiivisesti.

Luvussa 2 käsitellään tekoälyä, sen alakäsitteitä ja TDM:ää. Luvussa 3 käydään läpi tekijänoikeuksia ja siihen liitännäisiä käsitteitä. Luvussa 4 esitetään fair learning -käsite implikaatioineen ja käydään läpi tekoälyllä generoitujen teosten tekijänoikeuksia.

2 TEKOÄLY JA KONEOPPIMINEN

Tekoälylle (AI) ei ole tiettyä yhtä tarkkaa määritelmää, mutta Andreas Kaplan ja Michael Haenlein kuvailevat tekoälyn olevan systeemin kyvykkyys tulkita ulkoista dataa, oppia datasta ja käyttää tätä oppia saavuttaakseen spesifejä tavoitteita joustavasti (Kaplan & Haenlein, 2019). Tekoälyä pystytään hyödyntämään monenlaisissa erilaisissa käyttökohteissa. Se on kuitenkin paras sellaisissa tehtävissä, jossa käsitellään suuria määriä dataa.

2.1 Koneoppiminen

Koneoppiminen (machine learning, ML) on tekoälyn osa-alue, jolla pyritään luomaan algoritmeja, jotka tunnistavat annetusta datasta yhdistävät ominaisuudet. Näitä malleja hyödyntäen pystytään havaitsemaan uudesta datasta näitä samoja ominaisuuksia (Alpaydin & Bach, 2014, s. 1—4). Koneoppimisen kouluttamismateriaalin keräämiseen käytetään usein TDM:ää, joka kerää dataa esimerkiksi internetistä.

2.2 Syväoppiminen

Koneoppimisen alla olevaan syväoppimisen (Deep Learning, DL) opettamiseen vaaditaan suuria määriä dataa, jotta se pystyy toimimaan halutulla tavalla. Yleensä käytetty data on tekijänoikeuksilla suojattua. On kuitenkin tapoja, joilla datasettien puutteellisuudesta tai epätasapainoisuudesta johtuvaa heikkoutta pystytään korjaamaan otanta algoritmilla (Nasir ym., 2022) tai muilla tavoin (Kordzadeh & Ghasemaghaei, 2021). Tämä ei poista sitä faktaa, että datan määrän ja laadun lisääminen lisää koneoppivien mallien ulostulojen tarkkuutta ja vähentää niiden tuottamia systemaattisia virheitä tai vinoumia (Levendowski, 2018). Jos oppimismateriaalia ei ole tarpeeksi tai sen käyttöä rajoitetaan esimerkiksi siten, että opetusmateriaalia ei saa käydä lävitse, voi se aiheuttaa erinäisiä

negatiivisia vaikutuksia lopputulokseen. Esimerkiksi tämä voi johtaa vähemmistöjen väärin tunnistamiseen (tekoäly yhdisti 28 USA:n kongressin edustajaa rikollisiin), rotuennakkoluuloja vahvistaviin lopputuloksiin ja luottotietoalgoritmin potentiaalisesti vinoutuneihin tuloksiin. (Lemley & Casey, 2021). Generative Deep Learning (GDL) on syväoppimisen alalaji, jossa opetettu GDL-malli luo jotain uutta oppiensa mukaisesti (Franceschelli & Musolesi, 2022).

2.3 Generative Adversarial Network

Generative Adversial Network (GAN) on generatiivinen syväoppiva malli, jonka erityiset ominaisuudet tekevät siitä mielenkiintoisen subjektin tekijänoikeusväittelyissä. GAN:ssa on kaksi erillistä mallia, jotka tekevät eri asioita. Ensimmäiselle mallille, diskriminantille, opetetaan esimerkiksi datasetti kuvia. Toiselle mallille, generatiiviselle, ei opeteta mitään. Tämän jälkeen nämä kaksi mallia laitetaan toisiaan vastaan. Generoivan mallin on tarkoitus huijata diskriminanttia mallia luomalla datasettiä muistuttavia kuvia, kun taas diskriminantin mallin on tarkoitus arvata, kumpi kahdesta kuvasta (generoitu ja datasetin kuva) on todennäköisemmin datasetistä otettu kuva. Generoiva ”voittaa”, silloin kun diskriminantti valitsee generoidun kuvan (ts. generoiva malli onnistuu huijaamaan), ja diskriminantti ”voittaa” valittuaan datasetin kuvan. Kun generoiva voittaa, on diskriminantin opittava häviöstään ja vice versa. Diskriminantti oppii häviön kautta sitä, kuinka se voi erottaa paremmin generoituja kuvia datasetin kuvista. Kun taas generoiva häviää, sen pitää muuttaa ulostuloaan, jolla se voisi huijata diskriminanttia paremmin. Generoiva malli yleensä aloittaa kuvalla, joka on satunnainen (Goodfellow ym., 2014; Gillotte, 2020).

Tämä tarkoittaa siis sitä, että generoiva malli ei koskaan tule yhteyteen tekijänoikeuksien alla olevien teosten kanssa. Tämän asian voi nähdä siten, että mallin generoiva teos ei voi kopioida datasetin teosten tekijänoikeudella suojattuja ilmauksia, koska se ei ole koskaan nähnyt niitä. Vaikka asiaa ei näkisi tällä tavalla, on vaikeaa kieltää, etteikö tämä lisäisi teoksen transformaationaalista luonnetta. Väitettä tukee myös se, että väliaikaisia keskitason malleja ei luoda GAN:ia käyttäessä, jotka saattaisivat vanhemmissa tai erilaisissa generoivissa malleissa aiheuttaa joidenkin tutkijoiden (Sobel, 2017) mielestä mahdollisesti tekijänoikeusrikkomuksia.

2.4 Text and Data Mining

Text and data miningilla (TDM) tarkoitetaan prosessia, jolla kerätään esimerkiksi internetistä suuria määrää dataa (Carroll, 2019). Automatisoitu konelukija (tekoäly) kaapii, lukee ja tallentaa dataa, jota se löytää esimerkiksi internetistä. Tämän jälkeen kerätty data prosessoidaan sellaiseen muotoon, jossa sitä voidaan käyttää statistiseen analyysiin ja kuvioanalyysiin (Carroll, 2019). TDM:llä tuotettua dataa

tai datasettejä voidaan myös käyttää koneoppimisen opettamiseen. Koska koneoppiminen käyttää niin suuria määriä dataa, olisi manuaalisesti datan kerääminen erittäin vaivalloista ja epätehokasta.

Koska koneiden on pakko ladata kopio luettavasta kohteista, jotta se pystyy lukemaan kyseisen kohteen, on aihe tuottanut paljon tietoteknillisiä- ja lakikeskusteluita aiheesta. Carrol (2019) kuvailee 4 eri vaihetta, jossa tutkijat tekevät kopioita alkuperäisestä datasta. Nämä vaiheet ovat:

1. keräys ja datan kääntäminen
2. datan formatointi tietokoneelle käsiteltävään muotoon
3. datan prosessointi tietokoneen aktiivisessa muistissa
4. säilytys tai arkistointi, jotta uudelleenanalysointi, validointi ja toistettavuus olisi mahdollista.

Jotkut hallintoalueet, kuten USA ovat sallineet TDM käytön fair usen perustein, kun taas jotkin ovat rajoittaneet sitä suorasti tai epäsuorasti, kuten EU. Tämän jälkeen EU tosin on tehnyt poikkeuksen TDM:lle, joissa tietyin ehdoin se on sallittua (Gervais, 2021; Margoni & Kretschmer, 2022). USA:n tapauksessa, TDM on sallittua muun muassa neljästä syystä. Section 106(A) rajoittaa sitä, mitkä kopiot lasketaan copyrightin suojeluksen alle (17 U.S. Code § 106A). Toisena syynä on fair usen sallima transformaationaalinen käyttö (Carroll, 2019). Kolmantena syynä on se, että TDM ei itsessään pysty kuluttamaan copyrightilla suojatun teoksen ilmausta (Sobel, 2018). Viimeisenä syynä on se, että TDM:n käyttö ei vaikuta samoihin markkinoihin kuin TDM:llä kerätty teos (Sobel, 2018). Kolmas syy voi Sobelin mukaan vaikuttaa oikeuspäätöksiin, kun puhutaan kehittyneemmästä tekoälystä (Sobel, 2018). Neljäs syy voi myös aiheuttaa generoivan tekoälyn tilanteessa ongelmia (Sobel, 2018; Henderson ym., 2023; Lemley & Casey, 2021) ja tätä käydään läpi 4.1 luvussa.

EU:n tapauksessa ainoastaan tutkimusorganisaatiot ja kulttuuriperintöinstituutit saavat käyttää TDM:ää kaikkea tekijänoikeudella ja tietokantaoikeudella suojattua dataa, mutta vain ainoastaan tutkimuskäyttöön. Tietokantaoikeus eli sui generis database right (SGDR), luo EU:ssa tietokannoille ylimääräisiä immateriaalioikeuksia. Vaikka tietokanta pitäisi sisällään pelkästään faktuaalista tietoa, on sen kokonaisuus suojattu samoin tavoin kuin tekijänoikeudella suojattu teos (Directive 96/9/EC). EU:ssa olevilla mailla voi näiden säännösten lisäksi olla omia lakejaan, jotka tiukentavat edellä mainittuja sääntöjä.

Taulukossa (1) luetellaan eri maiden säätelyitä ja poikkeuksia TDM:n käytöstä tutkimuskäyttöön. Tutkimuksessaan Flynn, Schirru, Palmedo ja Izquierdo (2022) loivat listauksen lähes kaikkien maiden tutkimukseen liittyvistä TDM-poikkeuksista. Vaikkakin Latviassa on TDM-poikkeuksia, estävät muut tekijänoikeuslait TDM:n käytön. Suurin osa EU:hun kuuluvista maista ovat Suomen kaltaisessa tilassa, eli rajoituksia edelleen löytyy, mutta käyttöä ei ole kokonaan kielletty (Flynn ym., 2022).

TAULUKKO 1 Onko TDM:n käyttö sallittua tutkimuskäytössä eri maissa

TDM sallittua tutkimuskäytössä	**On**	**On, jakaminen on rajoitettua**	**On, ainoastaan tietyt tahot saavat käyttää**	**On, tiettyjä teoksia tai dataa ei saa käyttää**	**Rajoitettu niin, että melkein mahdotonta käyttää**
EU*			X	X	
USA	X				
Japani	X				
Iso-Britannia	X				
Suomi**		X		X	
Latvia**			X	X	X
Viro**	X				
Saksa**	X				
Argentiina				X	X

* EU:hun kuuluvat maat joutuvat silti tottelemaan EU:n direktiivejä, jotka antavat tekijänoikeuksien omistajille oikeuden kieltää se, että heidän teoksiaan käytetäisiin TDM:ssä. Myös tutkimus- ja kulttuuriperintöinstituuteilla on oikeus käyttää rajoittamatonta TDM:ää tutkimuskäyttöön.
** EU:hun kuuluvien maiden tulee siltikin noudattaa edellä mainittuja asioita.

Maailmanlaajuinen harmonisointi tekoälyn säätelyssä estäisi tiettyjä maita saamasta etulyönti asemaa tekoälyn suhteen, sekä se estäisi näitä maita tuottamasta vinoutuneita ja vaarallisia malleja tai ulostuloja (Gervais, 2023). Tämän toteuttaminen on tietenkin hankalaa ja mahdollisesti jopa mahdotonta. Gervais (2023) on kuitenkin ehdottanut kolmea eri tapaa, joilla tällaiseen maailmanlaajuiseen säätelyyn päästäisiin. Ensimmäisessä tavassa samaa mieltä olevat maat säätelisivät käytänteitä, joita he vaatisivat kauppakumppaneiltansa. Toisessa tavassa WTO (World Trade Organization) ottaisi ohjat tekoälyn säätelyssä. Viimeisessä tavassa UN:ään (United Nations) kuuluvat maat tekisivät kansainvälisen sopimuksen tekoälyn säännöistä (Gervais, 2023). Nämä tavat eriävät vaikutukseltaan ja kompleksisuudeltaan, mutta kaikki tavat ovat kuitenkin paljon työtä vaativia edesottamuksia.

TDM:n avulla pystytään avaamaan uusia väyliä tutkimuksessa ja sen avulla voidaan löytää uusia korrelaatioita tutkimuksesta, joita olisi melkein mahdotonta löytää ilman TDM:ää (Carroll, 2019). Suurin osa alan tutkijoista onkin sitä mieltä, että TDM:ää saisi käyttää myös tekijänoikeuksien alla olevien teosten keräämiseen (Levendowski, 2018; Lemley & Casey 2021; Franceschelli & Musolesi, 2022; Koščík, 2020; Kop, 2019 & 2020; Carroll, 2019; Bridy, 2019). Ainoastaan 28 % tieteellisistä artikkeleista (2018) ovat vapaan pääsyn ja käytön takana, vaikkakin

tämä määrä näyttäisi nousevan koko ajan (Piwowar ym., 2018). Osa julkaisijoista, jotka pitävät tieteelliset artikkelinsa maksumuurin takana ovat tehneet lisenssejä juurikin TDM varten, mutta palveluita, jotka yhdistäisivät useamman julkaisijan julkaisemia artikkeleja ei ole joko tehty, tai ne ovat epäsopivia TDM:n kanssa käytettäväksi. Täten on tutkijan joskus vain helpompi (tai jotta tutkimus pystytään suorittamaan rahallisesti) hankkia materiaaleja laittomia reittejä, kuten esimerkiksi Sci-hubista. USA:ssa tällainen käyttö on laillista ja Carroll esittää, että tämän vuoksi USA:lla on kilpailullinen etuasema innovaatiossa EU:hun verrattuna. (Carroll, 2019). On myös hyvä huomioida, että fair use -käsitellyssä ei aina oteta huomioon, sitä onko epäilty tekijänoikeusrikkomus tehty hyvässä vaiko pahassa tarkoituksessa. Vaikka oikeussalit ottaisivat tämän huomioon, ei sillä pitäisi olla merkitystä, koska tutkimus tuottaa transformatiivisiä etuja vahingoittamatta markkinoita (Carroll, 2019). Euroopassa tällainen toiminta ei ole sallittua, vaan poikkeuksissa määritellään, että TDM:ää käyttävällä tulee olla myös laillinen pääsy materiaaliin. Carrol myös tutkimuksessaan väittää, että tällainen laki on haitallista innovaatiolle (Carroll, 2019).

3 COPYRIGHT, TEKIJÄNOIKEUS JA FAIR USE

3.1 Copyright ja Tekijänoikeus

Copyright ja tekijänoikeus kuuluvat immateriaalioikeuksiin, ja ne käsittelevät tekijän oikeuksista tuotoksiinsa. Näihin tuotoksiin kuuluvat muun muassa kirjalliset ja artistiset teokset kuten: kirjat, elokuvat, kappaleet, maalaukset, tietokoneohjelmat, sähköiset tietokannat ja kuvat. Copyright-lainsäädännössä tekijä halutessaan pidättää teoksen kopiointioikeudet. Tekijänoikeuslainsäädännössä, joka on tiukempi ja laajempi, tekijä myös edellisen lisäksi omaa tiettyjä henkilökohtaisia oikeuksia, kuten oikeus estää vääristyneitä reproduktioita töistään (WIPO, 2016). Tekijänoikeudella ei kuitenkaan ole tarkoitus suojella teosten ideoita, teosten omistusta eikä faktuaalista informaatiota (Netaniel, 2008, s. 11), vaan tekijänoikeudella turvataan ilmausutapa (WIPO, 2016). Suurimmassa osassa maita copyright ja tekijänoikeus annetaan teokselle heti sen luonnin yhteydessä. Tämä tarkoittaa sitä, että lähes kaikki luonti, jota tehdään, on tekijänoikeuksien alla, kuten paria lausetta pidemmät viestit, selfiet ja ym. Tekijänoikeus kestää eri pituuksia riippuen hallintoalueista, yleisin on 70 vuotta tekijän kuolemasta, mutta poikkeuksia on olemassa.

Tekijänoikeuden päätarkoitus, riippuen maan laista, on edistää innovaatiota ja mahdollistaa sosiaalista ja ekonomista edistymistä. Toisena tärkeänä syynä on määritellä lakisääteinen oikeus teosten tekijöille, mutta nämä oikeudet pitävät kuitenkin olla tasapainossa yhteisen edun kanssa (WIPO, 2016), jotta tekijänoikeus edistäisi uuden tuottamista, eikä rajoittaisi sitä (Kop, 2019). Tekijänoikeuden tarkoitus on siis edistää yhteistä hyvää, antamalla uusille sukupolville, tai koneille, oikeus rakentaa taiteensa ja tieteensä edellispolvien tiedon ja taidon avulla.

Korkeimmillaan yhdestä tuomitusta copyrightin rikkomisesta, voi USA:ssa vaatia vähintään 200 ja enintään 150 000 dollarin hyvitykset. Koska opetusmateriaalit saattavat yltää helposti miljardeihin kuviin tai teoksiin, voivat korvausten

tasot nousta tähtitieteellisille tasoille. Tämä rajoittaa, tai ainakin pelottelee, pienempiä toimijoita edistämästä tekoälyn, koneoppimisen ja syvän oppimisen tiedettä. Getty Images haastoi Stability AI:n oikeuteen vaatien korvausta 12 miljoonasta kuvasta. Korvausten määrä on täten 2 400 000 000–1 800 000 000 000 € (2,4 miljoonaa–1,8 biljoonaa) jos Getty Images voittaa oikeudenkäynnin. Korkeimmin arvioidun yrityksen, Applen, valuaatio on 2.6 biljoonaa, ja edes silläkään ei olisi realistisesti mahdollisuutta maksaa korkeimpaa mahdollista korvausta. Pienempien, keskisuurien ja jopa suurien yritysten on mahdotonta maksaa korvauksia, jos ne häviäisivät vastaavalla skaalalla olevan oikeudenkäynnin.

Copyright ei vaadi USA:ssa sitä, että tekijä olisi ihminen (Bridy, 2012). Useammassa oikeudenkäynnissä on päätetty, että teos ei vaadi copyrightin saamiseen sitä, että tekijä olisi ihminen (Cummins v. Bond; Penguin Books U.S.A., Inc. v. New Christian Church of Full Endeavor, Ltd; Urantia Foundation v. Maaherra). Näissä tapauksissa on kyseessä ollut spirituaalisen ”ohjauksen” alla olevia henkilöitä, jotka olivat tuottaneet teoksia. Näille teoksille annettiin copyright sen takia, koska copyrightin saamiseen USA:ssa ei vaadita ihmistekijää. Copyrightin omistajaksi näissä tapauksissa osoitettiin ”ohjattu” henkilö.

Feist-oikeuspäätöksessä, copyright hylättiin puhelinluettelolta, sillä puhelinnumeroiden listaamista aakkosjärjestyksessä ei hyväksytty copyrightin suojeluun (Feist Publications, Inc., v. Rural Telephone Service Co.). Niin sanottu ”sweat-of-the-brow” -argumentti hylättiin, koska vaikka työnmäärä olisi ollut suuri, ei teos sisältänyt mitään ilmausta (expression). Tekoälyn ediystyksen kannalta tämä päätös on hyvä asia, sillä se tarkoittaa, että kattavat faktuaalista informaatiota sisältävät tietokannat eivät ole copyrightin alaisia. Mutta pienikin ”luovuutta” vaativa muutos listaan voi tehdä listasta copyrightillä suojellun.

EU:ssa asia on vain hieman eri tavalla, mutta lopputulos on eri, koska USA:ssa saa käyttää TDM:ää copyrightin alaisiin teoksiin. Tietokantaoikeus eli SGDR suojaa tietokannat lähes samalla tavalla kuin minkä tahansa tekijänoikeudella suojatun teoksen. SGDR kestää tosin vain 15 vuotta, mutta se annetaan aina uudestaan, kunhan tietokantaa päivitetään (Directive 96/9/EC).

Tekoälyn tuottamat tuotokset ovat luovia (Lemley & Casey, 2020), mutta parempi kysymys olisi: ovatko ne uusia (novel)? On kuitenkin tärkeää ottaa huomioon se, että käyttävätkö ihmisetkin luodessaan teoksia ainoastaan tuntemiaan, näkemiään ja oppimiaan asioita? Ovatko kaikki artistiset uudet (novel) tuotokset pelkästään asioita, joita henkilö on kokenut elämässään, jota hän sitten yhdistelee mielessään? Nämä ovat enemmänkin filosofisia kysymyksiä, joihin tämä tutkielma ei paneudu syvemmin

Koska aikaisemmin mainituissa oikeuspäätöksissä on annettu niin sanotulle ”epäsuoralle tekijälle” copyright oikeudet, olisi hyvin luontevaa, että tekoälyllä generoidun teoksen tilanteessa USA:n oikeusjärjestelmät päätyisivät samaan ratkaisuun. Tosin tässä tilanteessa teoksen tekijöitä on monia: mallin ohjelmoija, koulutusdatan omistaja ja loppukäyttäjä. Luvussa 4 jatketaan tämän asian tutkimista.

3.2 Fair Use

Fair Use on doktriini, joka takaa tekijänoikeuksellisten teosten kohtuullisen uudelleenkäytön ja kopioimisen. Fair use on USA:ssa yksittäistapauksellisesti selvitettävä asia. Tällaisia tapauksia ovat esimerkiksi kritiikki, kommentti, uutisraportointi, opetus, opinnäyte ja tutkimus, Artikla 107 mukaisesti (Copyright Act of 1976, 17 U.S.C. § 107, USA). EU:ssa tällaista käytäntöä ei ole, vaan jokaisella maalla on omat tekijänoikeuspoikkeuksia koskevat lait. Se, että onko käyttö fair usen alaista, tarkkaillaan neljän eri kohdan avulla:

1. käytön tarkoitus ja luonne, onko käytöllä tarkoitus tehdä rahaa vaiko ei;
2. tekijänoikeuden alla olevan teoksen luonne;
3. kopiotavan teoksen käytettävän osuuden määrä verrattuna koko teokseen;
4. potentiaalinen vaikutus alkuperäisteoksen kokonais- tai markettiarvoon.

Kohdat 1. ja 4. ovat yleisimmin tärkeimmät kohdat, joita tarkastellaan silloin, kun oikeus päättää, onko teos rikkonut tekijänoikeutta vai onko se soveltanut fair usea oikein. Teoksen transformatiivisuus (yksi asia, jota tarkastellaan kohdassa 1.), vaikkakaan ei pakollista, on yksi tärkeä osa teoksen fair use -määrittelyssä (Asay ym., 2020). Asay määrittelee transformatiivisuuden olevan sitä, että jos teos lisää jotain uutta, eikä ole alkuperäisen teoksen korvike, on se silloin tarpeeksi transformatiivinen fair usen kannalta (Asay ym., 2020).

4 TEKOÄLY, TEKIJÄNOIKEUS JA FAIR LEARNING

4.1 Fair Learning

Fair learning on periaate, jolla suojellaan jokaisen etuoikeutta oppia uutta, riippumatta siitä, onko oppija kone vai ihminen (Lemley & Casey, 2021). Artikkelissaan Lemley ja Casey toteavat, että pääsääntöisesti, tekijänoikeuksien alla olevien teosten käyttö tekoälyn opettamisessa tulisi olla sallittua (Lemley & Casey, 2021). Tätä fair learningiä tukee tällä hetkellä (USA:n) Fair Use -doktriini.

Fair learningiä perustellaan myös juridisesta näkökulmasta, mutta myös teknisestä ja oppimista edistävästä näkökulmasta. Teknillisesti, koneoppiminen vaatii valtavan määrän teoksia toimiakseen. Tämä voi tarkoittaa sadoilla tuhansilla tai miljoonilla teoksilla opettamisen. Koska lisensointi tai oikeuteen vieminen maksaisi niin paljon, ainoastaan isoimmilla yrityksillä ja instituutioilla olisi varaa kehittää ja käyttää tätä teknologiaa.

Fair learningiä ei tule sekoittaa fair machine learningiin, joka käsittelee koneoppimismallien reiluutta niiden lopputulemissa. Tämä on kuitenkin relevanttia fair learning -konseptiin, sillä jos yritykset ja instituutiot joutuisivat käyttämään halvempaa tai ilmaista, mutta heikompaa dataa, voi tämä johtaa reiluuden laskuun (Margoni & Kretschmer, 2022). Margoni ja Kretschmer varoittavat artikkelissaan EU:n CDSM-direktiivin (Directive on Copyright in the Digital Single Market) mahdollisista negatiivisista vaikutuksista luotaviin tekoälymalleihin (Margoni & Kretschmer, 2022; Directive 2019/790). Artikla 4 (2019) sallii tällä hetkellä oikeudenhaltijan kieltää datansa käytön muissa kuin tutkimus- ja kulttuuriperintöorganisaatioiden tutkimuspohjaisissa käyttötarkoituksissa, kun taas aikaisemmin direktiivi määräsi opt-in periaatteen. Margonin ja Kretschmerin mielestä tämä direktiivi (2019) on edistystä, mutta asia ei ole vielä optimaalisessa tilanteessa.

Datan lisensointi on yksi tapa, jota on ehdotettu tavaksi, jotta vältyttäisiin tekijänoikeusrikkomuksilta. Kuitenkin osa tutkijoista ovat sitä mieltä, että datan lisensointi ei olisi hyvä ratkaisu. Jos lisensointi tehdään pakolliseksi, ainoastaan isot, vakiintuneet yritykset ja organisaatiot tulisivat hyötymään siitä, sillä niillä ainoastaan olisi tarpeeksi rahaa maksaa miljoonien teosten käyttölisenssit tai he ovat itse keränneet alustoillaan ihmisistä tarpeeksi dataa ja teoksia (Levendowski, 2018; Lemley & Casey, 2021; Tang, 2021). EU:n tapauksessa poikkeuksen lisensointiin saisivat myös vakiintuneet tutkimus- ja kulttuuriperintöorganisaatiot (Margoni & Kretschmer, 2022.). Lisensointi poissulkisi mahdollisuuden tekoälyllä tutkimiseen yksittäisiltä henkilöiltä, tutkijoilta, journalisteilta ja kilpailijoilta (Levendowski, 2018; Lemley & Casey 2021; Tang, 2021).

Välttääkseen lisensoinnin tuomat lisäkulut, direktiivi voi ”kannustaa” EU:ssa toimivia yrityksiä opettamaan tekoälymallejaan datalla, joka on vanhempaa tai vinoutunutta, tai yritykset voivat hankkia valmiita tekoälymalleja, joita on opetettu epäverifioidulla datalla (Margoni & Kretschmer, 2022). Tämä voi aiheuttaa sen, että EU:ssa toimivat tekoälyn kehittäjät jäävät tekoälyn kehityksessä muualla toimivia kilpailijoita jälkeen. Vinoutunutta tai muuten vain huonolaatuista dataa käyttäessä on se vaara, että tekoälyn output on myös vinoutunutta ja huonoa. ”Garbage in – garbage out” pätee myös tekoälymallien käytössä. Jos dataan pääsy estetään, voi se johtaa vinoutuneisiin lopputulemiin, kuten esimerkiksi rotuennakkoluuloisiin, vähemmistöjä syrjiviin tai muuten vain vääriin ja mahdollisesti vaarallisiin lopputulemiin (Lemley & Casey, 2021; Ganguli ym., 2022). Tämän takia olisi tärkeää, että myös copyright-materiaalia voi saada tutkia ja arvioida, jolloin tällaisiin vinoutumiin voitaisiin puuttua. EU ovat rajoittaneet tekijänoikeuksien alaisten teosten tarkkailua ja ainakin vaikeuttaneet vapaan vertaisarvioinnin toimintaa (Gervais, 2021; Margoni & Kretschmer, 2022).

Parantamalla tekoälyn lopputulemien laatua, pystytään myös poistamaan tekoälyn saamaa stigmaa, jonka tekoäly on saanut joissain piireissä. Lisäämällä dataa ja parantamalla malleja (Nasir ym., 2022; Kordzadeh & Ghasemaghaei, 2021), voidaan korjata kohtia, joista tekoälyn kriitikot ovat huomauttaneet (Lemley & Casey 2021).

Kuten aikaisemmin tutkimuksessa todettiin, faktuaalinen tieto ei ole tekijänoikeudella tai copyrightilla suojattua (WIPO, 2016; Netaniel, 2008, s. 11), pois lukien SGDR:llä ja copyrigthilla suojatut valikoidut kokoelmat näistä faktuaalisista tiedoista. EU:n tilanteessa tämä estää tekoälyn opettamisen näillä suojelluilla teoksilla, jos teosten tekijät tai tietokantojen omistajat niin haluavat. USA:ssa sen sijaan fair use suojelee opettamista, koska suurin osa tekoälymalleista eivät ole kiinnostuneita teosten ilmauksista, eli siitä osasta mitä copyright, ja tietyissä maissa tekijänoikeus, suojelee (Lemley & Casey, 2021). Suurin osa tekoälymalleista ei välitä teosten luovista osista, mutta niiden on pakko kopioida myös ne osat, kun ne oppivat teoksista (Lemley & Casey, 2021). Esimerkiksi large language model -tekoälymalli (LLM) ei välitä kirjoitetun sähköpostin sisällöstä ja luovasta osuudesta, se haluaa vain tietää, kuinka sanoja käytetään yhteyksissä toisiinsa (Lemley & Casey, 2021).

Fair learningillä halutaan varmistaa myös ihmisten oikeus oppia faktuaalista tietoa teoksista, vaikka ne olisivatkin tekijänoikeuksien alaisia (Lemley & Casey, 2021).

> Most ML systems copy works not to consume the expression copyright law protects, but to get access to the facts or structures copyright law dedicates to the public. – – But the idea of fair learning doesn't just matter for robots. – – it reminds us that fair use is about more than just transforming copyrighted works into new works. It's about preserving our ability to create, share, and build upon new ideas. In other words, it's about preserving the ability to learn—whether the entity doing the learning is a person or a robot. (Lemley & Casey, 2021)

Muun muassa näiden perustelujen avulla Lemley ja Casey perustelevat fair learning -konseptia tärkeänä ja oleellisena yhteiskunnan etuja edistävänä asiana, joka tulisi ottaa huomioon varsinkin tekoälyä säädellessä.

4.2 Tekoäly ja Tekijänoikeus

Tällä hetkellä, vaikkakaan asia ei ole pakosti muuttumassakaan, tekoäly nauttii osittain siitä, että sillä ei ole oikeuksia. Koska tekoäly ei ole legaalinen entiteetti, niin ei siihen myöskään kohdistu ihmisiä koskevat lait. Tämän voi huomata siitä, että Text and Data Mining (TDM) on sallittua "roboteilta", mutta ei ihmiseltä. (Grimmelmann, 2017). Tämä myös tarkoittaa sitä, että ainakaan tällä hetkellä tekoäly ei voi omistaa tekemiään teoksia ja sitä ei voi asettaa vastuuseen.

Joidenkin tutkijoiden mielestä (Gervais, 2020; Sobel, 2018) tekoälyllä tuettujen teosten ei koskaan tulisi saada tekijänoikeutta, riippumatta siitä, onko tekoäly opetettu copyrightatulla tai tekijänoikeuksilla suojellulla materiaalilla. Toiset tutkijat (Lemley & Casey, 2021; Franceschelli & Musolesi, 2022; Assinen, 2018, s. 68; Škiljić, 2021) taas ovat sitä mieltä, että tekijänoikeus tulisi asettaa joko tekoälyn luoneelle ohjelmoijalle, datan omistajalle, tekoälyn käyttäjälle tai näiden yhdistelmälle. Tästä on erinäisiä väitteitä ja mielipiteitä, riippuen näkökulmista ja mielipiteistä. Jälkimmäistä väitettä tukee tutkimukset, joissa esitetään yhteiskunnan hyötyvän siitä, että tekijänoikeus jaetaan jollekin ihmiselle, eikä teosta vain todeta tekijänoikeudettomaksi (Franceschelli & Musolesi, 2022; Lemley & Casey, 2021; Bridy, 2012). On tietenkin tilanteita, joissa harmia saattaa aiheutua pienille artisteille, jos tekijänoikeuksilla suojeltujen teosten avulla luodaan uusia teoksia (Lemley & Casey 2021). Tässä asiassa tulee kuitenkin huomata Lysyakovin ja Viswanathanin empiirin tutkimus, jossa seurattiin tekoälyn vaikutusta logon suunnittelukilpailuissa (Lysyakov & Viswanathan, 2022). Tutkimuksessa verrattiin kuinka kilpailuihin osallistuvat reagoivat siihen, että logokilpailussa olisi mukana tekoälyllä generoituja teoksia. Tutkimus pystyttiin suorittamaan, koska dataa löytyi vuodesta ennen tekoälyn mukaan tulemista kilpailuihin ja vuodesta sen jälkeen. Osa kilpailijoista, jotka osallistuivat pääasiassa pienemmän rahasumman ja yksinkertaisempien logojen kilpailuihin, lähtivät alustalta. Taitavammat kilpailijat sen sijaan pääosin siirtyivät enemmän komplekseihin ja isomman

rahasumman kilpailuihin. Tutkimuksessa selvisi myös, että kilpailuissa aikaisemmin huonosti suorittaneet artistit tekivät tekoälyn lisäämisen jälkeen lisää tuotoksia, mutta parantamatta tuotosten laatua. Sen sijaan kilpailuissa aikaisemmin hyvin suorittaneet lisäsivät huomattavasti tuotoksiensa määrää (30—60 %), ja he myös lisäsivät logojensa laatua tekoälyn lisäämisen jälkeen (Lysyakov & Viswanathan, 2022). Tästä voisi päätellä, että vaikka tekoäly vähentäisi ”alemman” laadun teoksia, nostaisi se ainakin joissakin tapauksissa teosten laatua ja määrää.

Tulee ottaa myös huomioon, että tekijänoikeuksien kiristäminen auttaisi pääasiassa vain isoja, monopolistisia yrityksiä. (Tang, 2021; Lemley & Casey 2021). Myöskään vakiintuneet artistit eivät oletettavasti kärsisi siitä, että tekoälyn luomat teokset olisivat suojattu tekijänoikeuslailla, sillä populaarikulttuurissa ihmisten mielenkiinto kohdistuu pääasiassa rajattuun, suosittuun osioon teoksista (Gillotte, 2020; Lemley & Casey 2021).

Helmikuussa 2023 USA:n Copyright Office päätti, että tekoälyllä generoidut teokset eivät saisi nauttia tekijänoikeuksista (Novak, 2023). Vaikkakaan toimiston päätös ei päätä oikeudenkäynneistä tai laista, voi tällä olla suuri riski aiheuttaa lumipalloefekti tiukempaan tekijänoikeuslakiin. Tällaiset päätökset ja lait hidastavat tai jopa estävät yhteiskunnan edistymistä tekoälyn saralla. Copyright Office kuitenkin esitti aloitteen, jossa se kävisi läpi tekoälyn vaikutuksia tekijänoikeuksiin asianomaisten, IT-alan-, lakiammattilaisten kanssa.

EU:ssa asiat ovat vielä avoinna, mutta esimerkiksi Trapova (2023) väittää, että tällä hetkellä ei olisi mahdollista saada tekijänoikeuksia tekoälyllä generoidulle teokselle. Hän myöskin suosittelee, että tekoälyllä generoiduille teoksille ei suotaisi tulevaisuudessakaan tekijänoikeuksia (Trapova, 2023).

Tällä hetkellä on kuitenkin maita, jotka nauttivat vapaammasta tekoälylaista kuten: Iso-Britannia, Japani, Israel ja Irlanti. Isossa-Britanniassa esimerkiksi henkilö, joka on tehnyt tarvittavat järjestelyt teosten luontiin, saa tekijänoikeuden itselleen, mutta on vaikeaa sanoa, kuka on generoidun teoksen ”päätekijä” (Franceschelli & Musolesi, 2022). Tämä herättää tutkimuksessa aikaisemmin kysyttyä kysymystä, kenelle tekoälyllä generoidun teoksen tekijänoikeus kuuluu. Yksi vahvimmin perustelluista ja suosituimmista generoitujen teosten tekijänoikeuden jakamisen tavoista on USA:n work-made-for-hire-konsepti. Tämä tarkoittaa siis sitä, että tässä tapauksessa tekoälyn käyttäjä niin sanotusti ”palkkaa” tekoälyn tuottamaan teoksen, jonka oikeudet käyttäjä siten omistaa, aivan samalla tavalla kuin yritys, joka palkkaa työntekijän tekemään esimerkiksi logon yritykselle. Tämä on erittäin yleistä yritysten toiminnassa, ja se myös vaatisi vähiten muutoksia eri maiden lainsäädännöissä (Škiljić, 2021; Assinen, 2018).

Jotkin tutkijat huomioivat (Sobel, 2017; Lamley & Casey, 2021), että jos tekoälyllä generoitu teos vaikuttaa liian samankaltaiselta kuin jokin sen opetusmateriaalin kuva, voi olla, että fair use ei suojele uutta teosta. Aikaisemmin mainittu GAN-malli voi auttaa tähän ongelmaan. Koska GAN-mallin generatiivinen puoli ei koskaan käytä opetusmateriaalia, voidaan GAN-mallilla vähentää samankaltaisuuksien riskiä (Franceschelli & Musolesi, 2022). Muihinkin generatiivisiin

malleihin pystytään luomaan turvaverkkoja, joilla pystytään estämään samankaltaisuuksia.

Generatiivisella tekoälyillä tehtyjä parodioita on otettu alas internetistä (Canales, 2020). Jos parodiat sallitaan ihmiseltä, miksei myös tekoälyltä? Tämä kysymys menee tämän tutkimuksen laajuuden ulkopuolelle, mutta tästä voi huomata, että tekoälyn lainkuulumattomuus on samalla kirous, että siunaus.

Tämä tutkielma luo tietääkseen uuden (novel) huomion siitä, että tekoälyn käyttäjä valitsee luovia ja valikoituja termejä syötteeseen, luodessaan generatiivisen tekoälyn avulla, mikä täyttäisi mahdollisesti SGDR:n ja ainakin copyrightin vaatimukset. Tämä aihe vaatisi jatkotutkimuksia, sillä aihe menee tämän tutkimuksen laajuuden ulkopuolelle ja siksi, koska tämä tutkielma ei tee suoraa *de lege ferenda* -kannanottoa.

Koko tähän aiheeseen on vaikeaa sanoa lopullista päätelmää, sillä näitä asioita käsitellään oikeudenkäynneissä USA:ssa ja EU:n selvityksissä tälläkin hetkellä. Nämä päätökset tulevat muuttamaan generoivien tekoälyjen tulevaisuutta merkittävästi. Ongelmaa myös vaikeuttaa se, että tekijänoikeuslait ovat pääosin tehty silloin, kun digitaalisesta tekijästä ei ollut tietoakaan. Myös olemassa olevien lakien monitulkinnaisuus, jota ei helpota edellä mainittu asia, tekee ongelmasta vaikeasti lähestyttävän, sillä jokaisella tutkijalla on omat näkemyksensä siitä, mitä lailla on tarkoitettu.

5 YHTEENVETO

Tässä kirjallisuuskatsauksessa tutkittiin tekoälyn, tekijänoikeuksien ja fair learningin risteyskohtaa. Tutkielman tavoitteena oli vastata seuraaviin kysymyksiin:

1. Tulisiko tekoälyä opettaa tekijänoikeudella suojattujen teosten avulla ja tulisiko tämänkaltainen TDM:n (text and data mining) käyttö sallia kaikille, eikä pelkästään tietyille osapuolille?
2. Tulisiko tekoälyllä generoidulle teokselle suoda tekijänoikeus ja kenelle tämä tekijänoikeus suotaisiin?

TDM:llä tarkoitetaan prosessia, jonka avulla voidaan kerätä ja prosessoida suuria määriä dataa tai teoksia haluttuun käyttötarkoitukseen. TDM:n käyttöä on kuitenkin rajoittanut eri hallintoalueiden tekijänoikeuslait. Jotkin hallintoalueista ovat kuitenkin tehneet TDM-poikkeuksia, jotka ovat helpottaneet tieteellistä tutkimusta. Fair learning –konsepti (Lemley & Casey, 2021) suojaa ihmisen ja myös tekoälyn oikeutta oppia uutta aikaisemmista yhteiskunnan tieteellisistä ja taiteellisista tuotoksista. Tekoälyllä generoidut teokset on luotu syöttämällä tekoälylle luovia ja valikoituja termejä. Näiden termien avulla tekoäly luo teoksen, kuten esimerkiksi novellin, kuvan tai runon. Tekoälyllä luotujen teosten tekijänoikeuksista on esitetty monia mielipiteitä populaarikulttuurissa ja tieteellisessä tutkimuksessa. Tutkielma suoritettiin kirjallisuuskatsauksena ja lähteinä toimivat pääosin informaatioteknologia- ja lakiartikkelit, jotka ovat julkaistu arvostetuissa tieteellisissä aikakausilehdissä.

Kirjallisuuskatsauksen aineiston perustella kävi ilmi, että eri alojen tutkijoiden mielestä TDM:llä tulisi saada vapaasti tutkia aineistoja ja dataa, ainakin tieteellisissä tarkoituksissa. Jos tämänkaltainen TDM kiellettäisiin, voisi sillä olla merkittäviä negatiivisia vaikutuksia tieteellisiin tutkimuksiin. Joidenkin alan tutkijoiden mukaan tämä vapaus tulisi antaa myös niille TDM:n käyttäjille, jotka eivät kuulu vakiintuneisiin tieteellis- ja kulttuuriperintöinstituutteihin. Tämä mahdollistaisi itsenäisten tutkijoiden, yksittäisten henkilöiden ja yritysten suorittaa parempaa tutkimusta.

Tekijänoikeuksia ei ole luotu rajoittamaan yhteiskunnan tieteellistä ja taiteellista edistymistä tai tiedon kerääntymistä, eikä sen tarkoitus ole suojata faktuaalista tietoa. Tutkijoiden mielestä tekoälyä tulisi pystyä opettamaan myös tekijänoikeudella suojatuilla materiaaleilla. Tekijänoikeudella suojatun materiaalin poisjättäminen opetusmateriaalista voi lisätä lopputulemien vinoutumia merkittävästi, millä voi olla konkreettisia ja vakavia seuraamuksia.

Ilman laajempaa harmonisointia tekoälyn säätelyssä on vaikeaa tuottaa tasapuolista kilpailua maiden välillä, mikä voi tarkoittaa EU:n kilpailukyvyn laskemista globaaleilla markkinoilla, tutkimuksen heikentymistä ja heikommasta datasta johtuvia eriarvoisuutta ja vaaraa lisääviä vinoutumia. Tutkielma osoitti, että TDM:n rajoittaminen saattaa johtaa EU:ssa toimivia yrityksiä ostamaan koulutettu tekoäly tai data sellaisista maista, joissa tekoälyä tai TDM:ää ei ole rajoitettu millään tavoin. Rajoitukset voivat johtaa siihen, että yritykset saattavat käyttää ilmaista, mutta heikompaa dataa. Tämä saattaa johtaa edellä mainittuihin vinoumiin.

Generoivalla tekoälyllä tuotettujen teosten tekijänoikeusongelmat ovat kuitenkin monimutkaisempia ja vivahteikkaampia. Hallintoalueiden päätökset tätä asiaa koskevasta laista tulevat olemaan kauaskantoisia tekoälyn tulevaisuudelle. Eri alojen tutkijoilla on eriäviä näkemyksiä tähän kysymykseen ja niihin vaikuttavat esimerkiksi hallintoalueiden lait ja niiden monitulkinnaisuus. Tilannetta vaikeuttaa se, että tekijänoikeuslait ovat suurimmalta osin otettu käyttöön silloin, kun digitaalisesta tekijästä ei ollut tietoakaan. Tämä saattaa vaikeuttaa laintulkintaa huomattavalla tavalla. Uusia direktiivejä sekä lakeja laaditaan ja kartoitetaan eri hallintoalueissa, joten TDM:n käyttö ja tekoälyllä generoitujen teosten tekijänoikeudet ovat epävarmassa tilassa.

Tässä tutkielmassa havaittiin, että generoivan tekoälyn tuotoksien tekijänoikeuksien käsittelyyn ovat useat tutkijat esittäneet parhaimmaksi ratkaisuksi USA:n work-made-for-hire -käytäntöä. Sen käyttöönotto olisi helppoa suurimmassa osassa maita, sillä se on yleisesti yritysten käytössä globaalisti. Tekoälyn käyttäjä niin sanotusti palkkaisi tekoälyn tuottamaan hänelle toiveidensa mukaisen taideteoksen. Tällöin luodun teoksen tekijänoikeus siirtyisi tekoälyn käyttäjälle. Tähän tekijänoikeusongelmaan ei kuitenkaan löytynyt yhtenäistä vastausta, sillä tutkijat ovat eri mieltä tekoälyllä generoitujen teosten tekijänoikeuksista.

Tätä kirjallisuuskatsauksen laatimista rajoitti tutkimuksen keskittyminen pääosin USA:han ja EU:hun. Maat, kuten Japani ja Iso-Britannia, ovat kehittäneet omanlaiset ratkaisunsa osaan tutkielmassa mainituista ongelmista. Varsinkin generatiivisen tekoälyn tuottamien teosten tekijänoikeusongelma vaatisi lisää tutkimusta, koska eri tieteenalojen kesken ei olla päästy konsensukseen. Myös tekoälyn vaikutusta taidemarkkinoihin olisi hyvä tutkia laajemmalla skaalalla kuin mitä tässä tutkielmassa on käyty läpi. Maailmanlaajuinen harmonisointi tekoälyn juridisessa säätelyssä on haastavaa, mutta siihen olisi olennaista pyrkiä. Tätä aihetta tulisi tutkia lisää, jotta tekoäly ei tuottaisi enemmän harmia kuin hyötyä yhteiskunnalle.

LÄHTEET

Alpaydin, E. (2020). Introduction to machine learning. *MIT press*.

Asay, C. D., Sloan, A., & Sobczak, D. (2020). Is transformative use eating the world. *BCL Rev.*, *61*, 905.

Assinen, S. (2018). European Union Copyright Protection for AI-Generated Works. [pro gradu -tutkielma, Turun yliopisto]. https://www.semanticscholar.org/paper/European-Union-Copyright-Protection-for-Works-Assinen/9700607b67c1fdcf705d20bc26aaf6b3e33760df?p2df

Bridy, A. (2012). Coding creativity: copyright and the artificially intelligent author. *Stan. Tech. L. Rev.*, 5.

Canales, K. (24.7.2020) A researcher created a 'Weird A.I. Yancovic' algorithm that generates parodies of existing songs, and now the record industry is accusing him of copyright violations. *Business Insider*. https://www.businessinsider.com/weird-ai-yancovic-algorithm-parody-song-fair-use-2020-7?r=US&IR=T

Carroll, M. W. (2019). Copyright and the Progress of Science: Why Text and Data Mining Is Lawful. *53*.

Copyright Act of 1976, 17 U.S.C. § 106(A). Annettu Washington D.C:ssä 19.10.1976. Saatavilla sähköisesti osoitteessa: https://www.law.cornell.edu/uscode/text/17/106A

Copyright Act of 1976, 17 U.S.C. § 107. Annettu Washington D.C:ssä 19.10.1976. Saatavilla sähköisesti osoitteessa: https://www.law.cornell.edu/uscode/text/17/107

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Annettu 6.6.2019. Saatavilla sähköisesti osoitteesta: https://eur-lex.europa.eu/eli/dir/2019/790/oj

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. Annettu 27.3.1996. Saatavilla sähköisesti osoitteesta: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

Flynn, S., Schirru, L., Palmedo, M., & Izquierdo, A. (2022). Research exceptions in comparative copyright.

Franceschelli, G., & Musolesi, M. (2022). Copyright in generative deep learning. *Data & Policy*, *4*, *E17*. https://www.cambridge.org/core/journals/data-and-policy/article/copyright-in-generative-deep-learning/C401539FDF79A6AC6CEE8C5256508B5E

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N., El Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Johnston, S., Jones, A., Joseph, N., Kernian, J., Kravec, S., … Clark, J. (2022). Predictability and Surprise in Large Generative Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764. https://doi.org/10.1145/3531146.3533229

Gervais, D. J. (2020). The machine as author. *Iowa Law Review, 105(5), 2053-2106*. https://www.proquest.com/scholarly-journals/machine-as-author/docview/2436415378/se-2

Gervais, D. J. (2023). Towards an effective transnational regulation of AI. AI & society, 1-20.

Gervais, D.J. (2021). TRIPS Meets Big Data. In M. Burri (Ed.), Big Data and Global Trade Law (pp. 160-176). *Cambridge University Press.* https://www.cambridge.org/core/books/big-data-and-global-trade-law/trips-meets-big-data/5BBC9D440FC583634A6C08B72CB0FA30

Gillotte, J. L. (2020). Copyright Infringement in AI-Generated Artworks. 53.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (arXiv:1406.2661). arXiv. http://arxiv.org/abs/1406.2661

Grimmelmann, J. (2016). Copyright for Literate Robots. *Iowa Law Review*, *101*(2), 657–681.

Guinness, H. (16.03.2023) How to use DALL·E 2 to create AI images. *Zapier*. https://zapier.com/blog/how-to-use-dall-e-2/

Handke, C., Guibault, L., & Vallbé, J.-J. (2015). Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research. https://doi.org/10.2139/ssrn.2608513

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). *Foundation Models and Fair Use* (arXiv:2303.15715). arXiv. http://arxiv.org/abs/2303.15715

Jackosn, S. (28.03.2023) A Wharton professor had AI chatbots work on a business project for 30 minutes — Bing wrote 1,757 words in under 3 minutes, and ChatGPT wrote code to build an entire website. *Business Insider*. https://www.businessinsider.com/wharton-professor-tested-ai-work-in-30-minutes-chatgpt-bing-2023-3?r=US&IR=T

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62(1), 15–25*. https://doi.org/10.1016/j.bushor.2018.08.004

Kop, M. (2019). AI & intellectual property: Towards an articulated public domain. *Tex. Intell. Prop. LJ*, *28*, 297. https://doi.org/10.2139/ssrn.3409715

Kop, M. (2020). The Right To Process Data for Machine Learning Purposes in the EU. *Harvard Law School, Harvard Journal of Law & Technology (JOLT) Volume*, *34*. https://doi.org/10.2139/ssrn.3653537

Kop, M. (2020, March). Machine learning & EU data sharing practices. *Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University*, Issue No. 1/2020.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*, 1–22. https://doi.org/10.1080/0960085X.2021.1927212

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*, 1–22. https://doi.org/10.1080/0960085X.2021.1927212

Lemley, M. A., & Casey, B. (2021). Fair learning. *Texas Law Review, 99(4), 743-785*. https://www.proquest.com/scholarly-journals/fair-learning/docview/2506474838/se-2

Levendowski, A. (2018). How copyright law can fix artificial intelligence's implicit bias problem. *Washington Law Review, 93*(2), 579-630. https://www.proquest.com/scholarly-journals/how-copyright-law-can-fix-artificial/docview/2214888330/se-2

Lysyakov, M., & Viswanathan, S. (2022). Threatened by AI: Analyzing Users' Responses to the Introduction of AI in a Crowd-Sourcing Platform. *Information Systems Research*, isre.2022.1184. https://doi.org/10.1287/isre.2022.1184

Margoni T, & Kretschmer T. (2022) "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology." *GRUR International 71, no.* 8 685–701. https://academic.oup.com/grurint/article/71/8/685/6650009

Nasir, M., Dag, A., Simsek, S., Ivanov, A., & Oztekin, A. (2022). Improving Imbalanced Machine Learning with Neighborhood-Informed Synthetic Sample Placement. *Journal of Management Information Systems*, *39*(4), 1116–1145. https://doi.org/10.1080/07421222.2022.2127453

Netanel, N. W. (2008). *Copyright's Paradox*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195137620.001.0001

Novak, M. (22.2.2023). AI-Created Images Aren't Protected By Copyright Law According To U.S. Copyright Office. *Forbes.* https://www.forbes.com/sites/mattnovak/2023/02/22/ai-created-images-in-new-comic-book-arent-protected-by-copyright-law-according-to-us-copyright-office/

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale

analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

Rangwala, A. (03.05.2023) How to use Eleven Lab AI? *Open AI Master*. https://openaimaster.com/how-to-use-eleven-lab-ai/

Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Mantsyzov, A., Aliper, A., Aladinskiy, V., Cao, Z., Kong, S., Long, X., Liu, B. H. M., Liu, Y., Naumov, V., Shneyderman, A., Ozerov, I. V., Wang, J., Pun, F. W., … Zhavoronkov, A. (2023). AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel CDK20 small molecule inhibitor. *Chemical Science*, *14*(6), 1443–1452. https://doi.org/10.1039/D2SC05709C

Škiljić, A. (2021). When Art Meets Technology or Vice Versa: Key Challenges at the Crossroads of AI-Generated Artworks and Copyright Law. *IIC - International Review of Intellectual Property and Competition Law*, *52*(10), 1338–1369. https://doi.org/10.1007/s40319-021-01119-w

Slater, D. (27.03.2023) How to Use ChatGPT to Write a Short Story. *GripRoom*. https://www.griproom.com/fun/how-to-use-chatgpt-to-write-a-short-story

Sobel, B. L. W. (2018). Artificial Intelligence's Fair Use Crisis. *The Columbia Journal of Law & the Arts*, *41*(1), Article 1. https://doi.org/10.7916/jla.v41i1.2036

Tang, X. (2021). Copyright's techno-pessimist creep. *Fordham Law Review*, *90*.

Trapova, A. (2023) Copyright for AI-generated Works: a Task for the Internal Market? *European Law Review, 48.*

U.S. agency probes Tesla crashes that killed 2 motorcyclists. (5.7.2022). CBS News. https://www.cbsnews.com/news/tesla-crashes-killed-2-motorcyclists-autopilot-nhtsa

WIPO. (2016). *Understanding Copyright and Related Rights*. Saatavilla sähköisesti osoitteessa: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_909_2016.pdf

Yi Yang, Yu Qin, Yangyang Fan, & Zhongju Zhang. (2023). Unlocking the Power of Voice for Financial Risk Prediction: A Theory-Driven Deep Learning Design Approach. *MIS Quarterly, 47(1), 63–96*. https://doi.org/10.25300/MISQ/2022/17062