iLETİŞİM ve DİPLOMASİ
Communication and Diplomacy
ISSN: 2147-6772 e-ISSN: 2791-7649 Yıl/Year: Aralık/December 2023 Sayı/Issue: 11

Yapay Zekânın Siyasi, Etik ve Toplumsal Açıdan Dezenformasyon Tehdidi

The Threat of Disinformation from The Political, Ethical and Social Perspective of Artificial Intelligence

Derleme Makalesi / Review Article

Sorumlu yazar/ Corresponding author:
Kılıç Köçeri

ORCID:
0000-0003-1687-3001

Geliş tarihi/Received:
11.09.2023

Son revizyon teslimi/Last revision received:
16.10.2023

Kabul tarihi/Accepted:
02.11.2023

Yayın tarihi/Published:
16.12.2023

Atıf/Citation:
Köçeri, K. (2023). Yapay zekânın siyasi, etik ve toplumsal açıdan dezenformasyon tehdidi. *İletişim ve Diplomasi, 11*, 247-266.

doi: 10.54722/iletisimvediplomasi.1358267

Kılıç KÖÇERİ[1]

Öz

Makine öğreniminin bilgi işlemde kullanılması, kamusal alanı manipüle eden yapay zekâ yeteneğiyle oluşturulmuş dezenformasyon içeriklerinde hızlı bir artışa neden olmuştur. Yapay zekâ tekniklerinin kullanıldığı dezenformasyon içeriklerinin siyasi, etik ve toplumsal sonuçları, sosyal medya sunucularının kullanıcılarını, devletlerin ise toplumlarını dezenformasyondan koruma zorunluluğunu ortaya çıkarmıştır. Mevcut dezenformasyon sorununa çevrimiçi taciz, basın özgürlüğü, insan hakları ve etik problemler gibi sorunlar eklenmiştir. Bireysel ve devlet destekli dezenformasyon çabaları, toplumsal sistemde giderek yaygınlaşmıştır. Bu çabalar, gerçek haberleri saptırma, gayri meşru hâle getirme, eleştirmenleri susturma ve kamuoyunu manipüle etmek için yapay zekâ sistemlerinden yararlanmaktadır. Bu bağlamda araştırma, dezenformasyonun dinamiklerini ve yapay zekânın dezenformasyondaki rolünü analiz etmeye odaklanmıştır. Araştırmada literatür taraması yöntemine başvurulmuştur. Dezenformasyon ve yapay zekâ kavramları hakkında kapsamlı bir literatür taraması yapılmıştır. Yapay zekâ destekli dezenformasyonun mevcut etkilerinden yola çıkılarak genel bir değerlendirme yapılmış ve yapay zekâ tekniklerinin kullanıldığı dezenformasyon içeriklerinin siyasi, etik ve toplumsal sonuçlarının belirlenmesi amaçlanmıştır.

Anahtar Kelimeler: Yapay zekâ, deepfake, sistem, dezenformasyon, tehdit

[1] Doktora Öğrencisi, Ağrı İbrahim Çeçen Üniversitesi, Türkçe Eğitimi Anabilim Dalı, kılıç.koceri@adalet.gov.tr

ABSTRACT

Using machine learning in computing has led to a rapid increase in AI-capable disinformation content that manipulates the public sphere. The political, ethical, and social consequences of disinformation content using artificial intelligence techniques have created an obligation for social media providers to protect their users and states to protect their societies from disinformation. Online harassment, freedom of the press, human rights, and ethical problems have been added to the existing disinformation problem. Individual and state-sponsored disinformation efforts have become increasingly prevalent in the social system. These efforts use artificial intelligence systems to distort and delegitimize real news, silence critics and manipulate public opinion. In this context, the research focuses on analysing the dynamics of disinformation and the role of artificial intelligence in disinformation. The literature review method was used in the research. A comprehensive literature review was conducted on disinformation and artificial intelligence concepts. Based on the current effects of artificial intelligence-supported disinformation, a general evaluation was made and it was aimed to determine the political, ethical, and social consequences of disinformation content using artificial intelligence techniques.

Keywords: Artificial intelligence, deep fake, system, disinformation, threat

EXTENDED ABSTRACT

Artificial intelligence algorithms are used in the infrastructures of technology-based systems. Algorithms are developed by artificial intelligence systems. These algorithms form the basis of content management that enables the spread of disinformation. Disinformation producers aim to bring the content to the target audience and silence people of different opinions by taking advantage of this feature of artificial intelligence. Disinformation content usually includes comments from real users, copies or bots. Content emitted from fake accounts is detected by real users. Artificial intelligence algorithms make disinformation content more realistic. This has accelerated artificial intelligence research aimed at detecting disinformation content. Big tech companies have allocated important resources to combat disinformation. Given the amount of disinformation content, anti-disinformation tools appear to be insufficient. The use of some even caused legal and social controversy.

The main failure of algorithms used to combat disinformation is their inability to distinguish between true and false content. On the other hand, algorithms have trouble recognizing new disinformation content when focused on predefined data. Regardless of the amount of data defined, disinformation contents are too complex and dynamic to be detected by the algorithm. For this reason, artificial intelligence algorithms to be used in combating disinformation should be active, continuous, and

autonomous. Apart from that, the algorithms' results may also depend on who uses them and the purposes for which they are produced. Therefore, human intervention is still required in the audit of artificial intelligence algorithms used to detect disinformation. Because social network hosts acknowledge that autonomous AI algorithms alone are not successful.

Techniques used against disinformation should be transparent. Institutions collecting data should be able to explain and discuss this when individual rights are concerned. The algorithms used by some social media servers may collect more data than necessary, hide and keep it as a trade secret. Therefore, transparency has a broad meaning. The individual should know what data is being collected and shared. Social media networks should inform researchers about how data is collected. At the same time, researchers must have adequate legal protection. Public and private partnerships are also valuable in this regard. Both privacy and individual rights concerns must be taken into account.

Companies that develop information technologies and artificial intelligence tools are also responsible for both inputs and outputs. These companies should work with teams that can prevent disinformation. Data from studies should be documented in the form of datasets and models to use against disinformation. The potential impact and actual consequences of autonomous vehicles used in this way can be understood. Additionally, privacy and security should always be considered when using data. Managers who oversee tools developed to combat disinformation content should consider ethical issues. This can slow viral propagation or limit the effect of disinformation content. Autonomous tools do more than define or prevent disinformation. Therefore, the effectiveness of autonomous tools needs to be further investigated. Such research can be advanced through government and the private sector.

GİRİŞ

Araştırmalar, sosyal medyanın insanlar için değerli olduğunu göstermiştir (Candi, 2018, ss. 731-749). Sosyal medyada yapılan değerlendirme ve yorumlar, insanların bilgi kaynağı hâline gelmiştir. Ayrıca farklı araştırmalar, sosyal medya kullanımının kamusal söylemin yönlendirilmesinde potansiyel zararları olabileceğine yönelik bulgular da elde etmiştir (Miranda et al., 2016, ss. 303-330). Sosyal ağ mecralarında kişisel ya da kurumsal olarak üretilen bilgilerin dezenformasyon, gizlilik ve kamu güvenliğini tehlikeye düşürmek gibi etik olmayan sonuçları mevcuttur. Sosyal medya mecralarında yapay zekâ kullanılarak üretilen algoritma ve yazılımlar, dezenformasyonun otonom bir şekilde yayılmasına neden olan başlıca aktörler arasındadır. Yapay zekâ algoritmalarıyla üretilen botlar, dezenformasyonun yayılmasında önemli bir rol oynamaktadır. Kötü niyetli botlar, zarar vermek amacıyla geliştirilirken, araştırmalardan elde edilen bulgular, onların yalan haberler yaydığına, ekonomi sektörünü aldattığına, toplum-

sal bir kargaşa oluşturduğuna ve sosyal medya diyaloglarını üst düzeyde etkilediğine dair verilere ulaşmıştır (Kudugunta & Ferrara, 2018, ss. 312-322).

Gerçeğin saptırılması tarih boyunca tekrarlanan bir olgudur (Mork et al., 2020). Bu açıdan dezenformasyon, gelişmiş teknolojiyle desteklenen eski bir hikâyedir. Hızla gelişen bilişim teknolojileri, dezenformasyonun giderek yaygınlaşmasını, yapay zekâ teknikleriyle de yalan haberlerin hızlı bir şekilde hedef kitleye ulaşmasını sağlamıştır. 2016 ABD başkanlık seçimleri ve 2016 Birleşik Krallık AB'den ayrılma referandumuyla ilgili olarak vatandaşların görüşleri ve oy verme kararlarını etkilemek için sosyal medya ekosisteminden yararlanılmıştır (Howard & Kollanyi, 2016). Myanmar'da Facebook kullanıcıları, Rohingya Müslümanlarına karşı nefret yaymak isteyenlerin kullanışlı bir aracı olmuştur (Mantelero, 2018, ss. 754-772). 2022 yılında Türk Silahlı Kuvvetleri PKK terör örgütüne karşı kimyasal silah kullandı yalanı, sosyal mecralarda yürütülen dezenformasyon örneklerinden sadece birkaçıdır. Yalan haberler günümüzde de üretilmeye devam etmektedir.

Anlaşıldığı üzere dezenformasyon, alıcıyı kandırmak amacıyla paylaşılan yanlış, hatalı ve yanıltıcı bilgidir (Ireton & Posetti, 2018, s. 7). Ayrıca dezenformasyon, kandırma amacı gütmeden paylaşılan yanlış ve hatalı bilgileri de ifade eder. Gelişmiş dijital teknolojiler ve sosyal ağ mecraları hatalı ve yanlış bilgilerin artmasına neden olmuş ve çoğu devlet bu durumu yasal ve teknik yollarla mücadele edilmesi gereken bir tehdit olarak görmüştür (Avrupa Komisyonu, 2018). AB'de olduğu gibi Türkiye'de de dezenformasyonun manipülatif karakterinin yanı sıra temel hak ve özgürlüklerin, özellikle de ifade ve bilgi edinme özgürlüğünün korunmasına ilişkin yasalar çıkartılmıştır. Bazı kurum, kuruluş veya kişiler tarafından gerçeğin yanlış olarak kabul edilmesi, ifade ve bilgi edinme özgürlüğüne ciddi zararlar vermektedir. Dezenformasyon sorunu öncelikle yayılan bilginin toplumu manipüle etmek için kasıtlı bir şekilde aldatıcı olmasını ve sonrasında bilgiyi yayanın modern tekniklerden yararlanması anlamında özeldir.

Saniyede 6 bin, dakikada 350 bin ve günde 500 milyon tweet atıldığı, atılan tweetlerin %8,5'inin 23 milyon sosyal botun attığı tespit edilmiştir (Salge & Berente, 2017, ss. 29-31). Pew Research Center'ın (2008) 47 gün boyunca yapmış olduğu araştırmada 1,2 milyon tweet incelenmiş ve atılan tweetlerin %66'sının dezenformasyon amaçlı oluşturulduğuna dair bulgulara ulaşılmıştır (ss. 1673-1689) Bu bağlamda araştırmada, öncelikle yapay zekâ teknolojileri ile desteklenen dezenformasyon içerikleri incelenmiş, ne tür tehditlere yol açabileceği ve etkilerinin nasıl azaltılabileceği irdelenmiştir. Araştırma sonunda "deepfake" içeriklerini tespit etmek veya karşı koymak amacıyla kullanılan tekniklere odaklanılmış ve yapay zekâlı dezenformasyon teknolojileri hakkında yayınlanmış literatür taramasından elde edilen bilgilerle siyasal, etik ve toplumsal etkilere yönelik sonuç ve öneriler sunulmuştur.

Yapay Zekâ, İçerik Yönetimi ve Sosyal Botlar

İçerik yönetimi, derin öğrenmeden algı yönetimine, çevrim içi yayınlardan yapay zekâya ve toplum mühendisliğine kadar geniş bir alana yayılmıştır (Bukovská, 2020). Sosyal ağ mecralarının konuşma, yazışma ve davranış kurallarını düzenleyen hizmet şartlarını sağlaması için yapay zekâ sistemlerinden faydalanılır. Çocuğun cinsel istismarı, terör ve şiddet eylemleri içeren materyallerin otomatik olarak algılanması ve engellenmesi buna örnek gösterilebilir. Yapay zekâyla oluşturulan algoritmalar, çok sayıda yazılı, sözlü ve görüntülü veriyi eş zamanlı olarak düzenlemekte ve içeriğini denetlemektedir. Ayrıca bu sistemler, geniş bir algoritma ağı üzerine inşa edildiğinden dezenformasyonun yayılmasında kullanılmaktadır. Algoritmalara dayalı talimatlar otonom bir içeriğe sahip olduğundan hesap kurtarma, içeriği önceleme, tanıtım, içeriğin kısıtlanması ve iyileştirilmesinde oldukça etkilidir.

Yapay zekâya sahip sosyal botlar, dezenformasyonun yayılmasında önemli hâle gelmiştir. Bu botlar içerik oluşturmak için sosyal medyayı kullanan yeni bir bot türüdür (Boshmaf et al., 2011, ss. 93-102). Twitter ve Facebook gibi kullanımı kolay sosyal ağlar, botların yaygınlaşmasında oldukça etkilidir. Paylaşım içerikleri sosyal medyayı toplumsal algı yönetiminin önemli bir parçası hâline getirmektedir. Bu bağlamda sosyal botlar, toplumsal algı yönetiminin önemli bir bileşenidir (Russell & Norvig, 2016). Sosyal botların sosyal ağ mecralarında yaygınlaşmasının hem iyi hem de kötü etkileri olmuştur. Covid 19 döneminde sosyal medyada botlar aracılığıyla faydalı bilgilerin yayılması buna örnektir. Ancak kötü niyetli botlar, yalan ve yanlış haber kaynaklarıdır.

İçerik, dezenformasyon süreçlerinin yakıtıdır. Hedef kitle belirlendikten sonra operatörler, 24 saatlik bir haber döngüsü oluşturmak ve toplumun dikkatini çekmek için bilgi ortamını sosyal botlar aracılığıyla şekillendirerek içerik akışı oluşturmaya çalışırlar (Storozuk et al., 2020, ss. 472-481). İçerik geliştirme süreci interaktiftir, yoğun emek, kullanıcı etkileşimi ve kesintisiz bir içerik akışı gerektirir. “TSK kimyasal silah kullandı” dezenformasyonu sırasında grafik, blok, video üretimi ve sosyal medya uzmanları gibi etiketlere sahip çok sayıda kişinin varlığını Twitter paylaşımlarında görmek mümkündü. Bazı sözde siyasetçiler, doktorlar ve sanatçılar organik olarak viral olabilecek içerik oluşturma konusunda bu uzmanlara prim vermişlerdir. Gerçek kişilerin üretmiş olduğu ve kimyasal kullanımına atıfta bulunan binlerce gönderi paylaşılmıştır. Bu tür bir içerik yönetimi için hedef toplumun diline, ideolojisine ve kültürüne hitap edecek beceriye sahip özel bir ekip gerekir (Brief, 2021).

Operatörler, olumlu olumsuz, umutlu umutsuz, öfkeli hatta tarafsız içerikler yayınlarlar. Farklı türde içeriğin yayınlanması, toplumun dikkatini çekmek ve algısını yönetmek için gereklidir. Tarafsız ve konuyla ilgisiz içerik üretimi, troller ve provokatörlerin yönettiği hesaplara doğru kitlesel bir akışın başlamasını sağlarken, olumlu gönderiler kullanıcıların güvenlerini kazanma amacı güder. Ayrıca farklı siyasi düşün-

ceye ve dünya görüşüne sahip kullanıcıların hedef sayfa ya da profillere katılımlarını sağlayabilir. PKK terör örgütüne yakınlığıyla bilinen sosyal ağların paylaşımları, terör eylemleri öncesinde insaniyken, saldırılar sonrasında terör örgütünün amaçlarına hizmet etmesi için olumsuz içerikler de üretmiştir. Örneğin, Türkiye'nin dinamiklerini hedefleyen 28 Mayıs 2013 Gezi Parkı dezenformasyon çabaları, içerik üretimlerinde ağırlıklı olarak öfke gibi olumsuz ve duygusal içeriklere ağırlık vermiştir.

İçerik türleri kısa mesajlardan, orta uzunluktaki makalelere ve görsellere kadar değişir. Troller, kullanıcıların ekran görüntülerini kopyalayabilir, yapay zekâ teknikleri kullanarak değiştirebilir ve yeni içerikler üretebilirler. Ekran görüntüsü alma ya da kopyalama teknikleri provokatörlerin kimliklerini gizlemesine yardımcı olurken, içerik oluşturma çabalarını en aza indirir ve paylaşımın otantik özelliği korunur. Yapay zekâ teknikleri kullanılarak üretilen içerikler yaygın dil bilgisi hatalarından veya yanlış kullanılan deyimlerden arınıktır (Brief, 2021).

Dezenformasyon içeriği öncelikle belirsiz bir web sitesine veya video paylaşım platformuna yüklenir. Bu şekilde dezenformasyona konusu tekrar tekrar web siteleri ve video paylaşım platformları referans gösterilerek paylaşılır. Son yıllarda bazı düşünce kuruluşları, yalan haber ve araştırmalara referans olması bakımından devlet desteğiyle kurulmuştur (Center, 2020). Benzer yöntem Ermeni diasporası tarafından da kullanılır. Ermenistan tarafından desteklenen web siteleri, 1915 olayları hakkında dezenformasyon içerikleri üretmeye devam etmektedir. Oluşturulan içerikler, maksimum sayıda göz ve kulağın maruz kalması için sosyal medya üzerinden trend konular ve hashtagler altında toplanıp yayılmaları sağlanır.

Dezenformasyon içerikleri dinamik bir katılımla yayılır. Sosyal medya mecralarının ekosisteminde aldatıcı ve yalan haberler patlarken, gönderi maksimum insan sayısına ulaşır. Doğal tartışmalar dezenformasyonun bir dayanak kazanmasına ve gelişmesine yardımcı olur. Sosyal medya kullanıcıları, konuyu savunmak ya da eleştirmek için bilişsel ve duygusal enerjilerini harcamak zorunda kalırlar. Sosyal medya mecrasının uzun tartışma ortamlarına olanak sağlaması, dezenformasyon içeriğinin yayılmasına hatta yeni dezenformasyon içeriklerinin ortaya çıkmasına neden olur. Amaç dezenformasyon içeriğini sağlamlaştırmak ve doğal kullanıcılar arasında nefret tohumları ekmektir. Bu şekilde provokatörler ve troller doğal sosyal medya kullanıcılarını birbirleriyle ayrıntılı olarak tartıştırmaya ve dezenformasyon içerikleri üretmeye devam ederler.

Yapay Zekâ Sistemleri ve Dezenformasyon

Yapay zekâ teknikleri, dezenformasyonu çevrimiçi ortamlarda iki şekilde güçlendirir. Öncelikle yapay zekâ teknikleri kullanılarak manipüle edici metin, görüntü, ses

ve video içerikleri oluşturmak provokatörlere ve trollere kolaylık sağlar. Daha sonra doğal kullanıcıların üretilen içeriklerle etkileşimini artırmak ve dezenformasyon içeriklerinin çevrim içi ortamlara hızlı ve etkili bir şekilde yayılması amaçlanır. Bu nedenle yapay zekâ teknikleri, dezenformasyon sorununa neden olan ana aktör olarak kabul edilir. Bu durum derinlemesine incelenmesi gereken çok sayıda etik sonucun ortaya çıkmasına neden olur.

Dezenformasyon içeriklerinde yapay zekâ tekniklerinin kullanılması, deepfake ürün olarak tanımlanır. Sosyal medya ağlarının ortaya çıkmasıyla birlikte, nispeten kötü aktörlerin içerikler üzerinde sosyal, ekonomik, siyasi ve toplumsal aklı etkileyecek değişiklikler yapması, dezenformasyon sorununun büyümesine neden olmuştur. Yapay zekâ tekniklerindeki gelişmeler metinleri, görselleri ve sesleri insanları manipüle etmek için daha gerçekçi hâle getirmiştir (Akers et al., 2018, s. 4). Bu nedenle deepfake ürünlerin tam olarak ne olduğu açıklanmalıdır. Deepfake, derin ve sahte kelimelerinin bileşiminden oluşur. Deepfake ürünler, GAN (Generative Adversarial Network) adı verilen bir ağda çalışan iki yapay zekâ algoritmasının ürünüdür (Bontridder & Poullet, 2021, s. 32). GAN sayesinde mevcut içerik kümeleri, algoritmik olarak yeni veri türlerinin üretilmesini sağlar. Örneğin; GAN aracılığıyla Cumhurbaşkanı Recep Tayyip Erdoğan'a ait çok sayıda fotoğraf analiz edilir, ardından analiz edilen fotoğraflarla birbirinin tam kopyası olmayan yeni bir görüntü, ses ve videolar oluşturulabilir. Bu teknoloji, çeşitli içerik türlerine kolaylıkla uygulanabilir.

Deepfake ürünlerini oluşturmak kolaylaşmış ve sosyal medyada büyük bir hızla yayılmıştır. Mayıs 2019'da Moskova'daki Samsung Yapay Zekâ Laboratuvarındaki araştırmacılar, konuşabilen animasyonlu kafa videoları üretmişlerdir. Bu videolarda Albert Einstein ile Leonardo da Vinci'nin Mona Lisa'sı tek görüntü girdisi olarak kullanılmıştır (Alvares & Salzman, 2019, ss. 181-202). Bu deepfake ürünleri oluşturmak için önemli bir adımdır çünkü daha önce bu tür görüntüleri oluşturmak geniş bir veri ağı gerektiriyordu (Zakharov et al., 2019, ss. 9459-9468). Birinin kafasını, başka birinin vücuduna entegre etmek de yapay zekâ sayesinde kolaylaşmıştır. Yine de vücut hareketleri sergilemek veya taklit etmek için bir aktöre ihtiyaç vardır (Vincent, 2021, ss. 425-438).

Dezenformasyonda kullanılan çok sayıda yapay zekâ teknolojileri olmasına rağmen, deepfake görselleri, ses klonlama ve metin içerikleri en endişe verici olanlarıdır. Bu nedenle deepfake ürünleriyle dezenformasyonun normalleşmesi, genel bir güven erozyonuna yol açabilir. Deepfake videolarının sayısı katlanarak artmış ve son altı ayda iki katına çıkmıştır (Collins & Ebrahimi, 2021). Yalan haberlerin etkisi, gerçek bir olayın deepfake olarak reddedilmesine, devletlerin gelişen teknoloji merkezlerine, kurumlara ve demokratik süreçlere duyulan toplumsal güvenin azalmasına neden olabilir (Chesney & Citron, 2019). Toplumsal güvenin erozyona uğraması, gerçeği kurgudan ayırmadaki yetersizlik, şüphecilik döngüsünü daha da kötüleşebilir ve dezenformasyona karşı toplumların hassasiyeti artabilir.

Deepfake Manipülasyon Ürünleri

Görsel manipülasyon teknikleri yeni değildir. Fotoğraflar ve videolar üzerinde uzun yıllar restorasyon çalışmaları yapılmıştır. Yüz değiştirmenin en eski örneklerinden biri Amerika'nın 16. başkanı Abraham Lincoln ile Amerikalı siyasetçi John Calhoun'un kafalarının Amerikan iç savaşında hiciv amaçlı yeniden düzenlenmesidir. Ancak son yıllarda değişen şey, yapay zekâ algoritmaları başta olmak üzere gelişen yazılım formlarıyla dezenformasyon içeriklerinin üretilmesidir. Aynı zamanda bu teknolojik yazılımlar ticari bir metaya dönüşmüş, FakeApp ve Face Swap gibi uygulamalarla yaygınlaşmıştır. Görüntüleri veya videoları manipüle etmek için kullanılan temel teknik, iki fotoğraf arasındaki sınırların belirlenmesi ve görüntünün A noktasından B noktasına hareketle diğerinin dinamik görünümünü almasıdır (Ivakhiv, 2016, ss. 31-39).

Deepfake görselleri, istismar ve dezenformasyon için kullanılan güçlü araçlar oldukları gibi toplumsal düzeni ve güveni etkileyebilecek güce sahiptir. Temel yapay zekâ teknolojileri tespit edilmesi zor, düşük maliyetli deepfake görsellerinin geliştirilmesine olanak sağlamıştır. Temel deepfake teknikleriyle yüz değişikliği, yüz ifadesi değişikliği, yüz ve ses sentezi yapılabilir. Bu şekilde bir görsel, ses veya videodaki birinin aslında hiç söylemediği veya yapmadığı bir şeyi yapıyor ya da söylüyor gibi görünmesini sağlar. Deepfake tekniklerinin kendi alanlarında meşru uygulamaları olsa da genellikle eğlence, maddi kazanç ve dezenformasyon amaçlı kullanılır.

Yüz değiştirmenin üç aşaması vardır: kırpma, geliştirme ve yaratma (Jacobs et al., 2018, ss. 7-31). Kırpma aşamasında yüz modelini geliştirmek için yeterli görüntü toplanır. Hedef kişilerin videolarını oluşturmak için görüntüler kırpılır. Bu şekilde hedef kişilerin yüz yapıları ve kafa büyüklükleri birbirlerine benzer. Deepfake teknikleri bu aşamadan sonra kullanılmaya hazır hâle gelir. Geliştirme aşamasında toplanan görüntüler kullanılarak yüz değiştirme modelleri tasarlanır. Yüz değiştirme modeli, bir kodlayıcı ve bir kod çözücüden oluşur (Young et al., 2021, ss. 467-468). Kodlayıcı yüzün temsili bir görüntüsünü alır ve düşük boyutta sıkıştırır. Bu temsil üzerinden yüz orijinal formunda yapılandırılır. Yaratma aşamasında deepfake görüntüleri videoya ya da görsele eklenir. Videodaki her kare için sentezlenen yüz açısının hedef kişinin baş açısıyla eşleşmesi sağlanır.

Deepfake görsellerinde farklı yapay zekâ teknolojileri olan oto kodlayıcılar ve çekişmeli üretici ağlardan (GAN) faydalanılır. Oto kodlayıcılar, bir resmin temsili görüntüsünü yeniden yapılandırmak için geliştirilmiş yapay sinir ağlarıdır. Çekişmeli üretici ağlar, birbirlerine rakip iki sinir ağından oluşur, biri sahteyi üretmeye çalışırken, diğeri onu tespit etmeye çalışır. Bu döngünün devamında video ya da resimdeki yüzlerin daha makul görüntüleri ortaya çıkar. Çekişmeli üretici ağlar üst düzey sanal gerçeklik üretirler fakat kullanımları oldukça zordur.

İyi hazırlanmış bir deepfake ürünü, üst düzey bilgi işleme kaynakları, zaman, para ve beceri gerektirir (Helmus, 2022, s. 24). Her deneyim deepfake videolarını daha ger-

çekçi hâle getirirken, yapay bileşenlerin çıplak gözle tespit edilmesini de zorlaştırır. Çeşitli web siteleri deepfake eğitimlerini ücretsiz sunmaktadır. Popüler siteler arasında kullanıcıların mevcut videolar ve GIF'lerdeki yüzleri değiştirmesine izin veren Reface vardır. Kullanıcılar Reface sayesinde ölen akrabalarının fotoğraflarını canlandırırken, MyHeritage kullanıcıların çeşitli film karakterlerine kendi yüzlerini birebir entegre etmelerini sağlayan deepfake teknolojisine sahiptir.

Ses klonlama, deepfake üretiminin başka bir yoludur. Celebrity Voice Cloning ve Voicer Famous AI gibi çevrim içi uygulamalar, kullanıcılara ünlü kişilerin seslerini taklit etme kolaylığı sağlar. Bu tür hizmetlerin kötü niyetli kullanımına ilişkin örnekler mevcuttur. Örneğin; İngiltere merkezli bir enerji şirketinin CEO'su merkezden patronunun sesine benzeyen birinin aradığını ve ses klonlama yazılımının bir ürünü olduğu iddia edilen telefondaki sesin talimatıyla Macaristan'daki birine yaklaşık 220.000 avro para transferi gerçekleştirdiğini bildirmiştir. Başka bir örnekte ise ABD'den bir adam, ses klonlamanın kurbanı olmuş, telefonda oğlunun sesine benzeyen birinin hapishanede olduğunu ve kurtulmak için 9000 dolar avukat ücretine ihtiyacı olduğunu ifade etmiştir (Stupp, 2019).

Bir deepfake ürünü olan sahte fotoğraflar da manipülasyon ve endişe kaynağıdır. Deepfake görüntüleri genellikle gerçekçi görseller içerir. Kullanıcılar sahte, hızlı ve kolay deepfake görüntüleri oluşturabilecekleri web sitelerine kolayca erişebilirler. Kötü niyetli hazırlanmış deepfake görüntülerinin bazıları, uzmanların yürütülen casusluk operasyonlarının bir parçası olarak kullanıldığı düşüncesine neden olmuştur. Deepfake görüntülerin kullanıldığı sosyal medya profillerinin çoğu hatırı sayılır bir takipçi sayısına, stratejik ve uluslararası çalışmalar yapan sahte kullanıcı isimlerine sahiptir (Satter, 2019).

Deepfake görüntüler sahte sosyal medya hesaplarının bir parçasıdır. Facebook tarafından yapılan bir araştırma, sahte görselleri profil resmi olarak kullanan düzinelerce devlet destekli hesabın varlığını tespit etmiştir (Kopf et al., 2019, ss. 310-312). Provokatörler ve troller, dezenformasyon içerikleri üretiminde alternatif insan görüntülerini kullanırlar. Ancak kullanıcılar, çalıntı profil resimlerini tespit edebilecekleri bir araca sahiptir. Örneğin, Google'ın zıt görsel arama özelliğini kullanarak şüpheli bir fotoğraf için interneti taramak ve görsel kökenini tespit etmek mümkündür. Bu nedenle provokatörler, savunma önlemlerini aşmak için izi sürülemeyecek görseller üretirler.

Yapay zekâ, doğal dil modelleri kullanarak gerçeğe yakın metinler üretebilir. The Guardian gazetesi, 8 Eylül 2020 tarihinde "Bunu Bir Robot Yazdı" başlıklı bir haber makalesi yayınlamıştır. Haber metninde GPT-3 (Generative Pre-Trained Transformer-3) ve OpenAI tarafından geliştirilmiş doğal bir dil modeli kullanılmıştır. GPT-3, CommonCrawl, WebText ve Wikipedia'dan alınan verilerden bir haber makalesi oluşturmuştur (Tom B. Brown et al., 2020). The Guardian Gazetesinin haber editörleri, GPT-3'e dili basit,

öz ve 500 kelime civarında kısa bir köşe yazısı yazması talimatını verdiğinde genel olarak paragraf düzeyinde gerçekçi ve anlamlı bir haber makalesi ortaya çıkmıştır.

Yapay zekâ algoritmaları aracılığıyla üretilen metinler, korkutucu olabilir. Sosyal medya bot ağları güçlendirilerek troll ya da provokatörlerin metin içerikleri hazırlamalarına gerek kalmayabilir. Örneğin, FireEye araştırmacıları, Rus trollerin 2016 ABD seçimlerine müdahale etmek için sosyal medya paylaşımlarına GPT-3'ün öncüsü GPT-2'yi kullanarak karşı koymuşlardır (Simonite, 2019). Manipülasyon içeren metinlerle mücadele, elektronik harp tekniklerindeki düşman radar sistemlerinin körleştirilmesine benzer. Manipülatif metinler gerçek haberleri bastırmak ve sahte içerikleri yaymak için sosyal medyada dolaşıma sokulurlar. Gerçek sosyal medya kullanıcılarının yapmış oldukları dil bilgisel hatalar, deepfake teknikleri kullanılarak oluşturulan metinlerde bulunmaz. Bu durum yazılı propagandanın daha inandırıcı ve zor tespit edilmesine neden olur.

İçerik üretim teknikleri geliştirildikçe dezenformasyon yaymayı amaçlayan bazı kötü aktörler, hem içerik üretimi öncesinde hem de sonrasındaki denetim sistemlerini atlatmanın yollarını bulmuşlardır (Pierson et al., 2023, ss. 125-157). Örneğin; bazıları denetim mekanizmalarından kaçmak için "COVID-19" yazmak yerine "COv1D" yazarak içeriği gizlemeye çalışmışlardır. Provokasyon amaçlı içeriklere karşı birçok ülke ve şirket, yapay zekâ temelli sistemlerini değiştirmiş, dezenformasyon içeriklerini tespit etmek için eğitim faaliyetlerine önemli kaynaklar ayırmıştır. Dezenformasyona karşı geliştirilen çok sayıda algoritmanın varlığına rağmen bazı içerikler sistem açıklarından geçebilir. Bu nedenle yapay zekâ ve makine öğrenimi tabanlı içerik denetleme araçlarının handikapları göz önüne alındığında, en iyi çevrim içi içerik denetiminin, yapay zekâ ve insan incelemesinin bir bileşimi olduğu anlaşılmalıdır. İçerik denetimlerinden insan gözetiminin kaldırılması, otonom sistemlerden kaynaklanan nefret, aşırıcılık, önyargı, ayrımcılık ve provokasyonun güçlenme riskini artırabilir.

Dezenformasyon İçeriklerinin Grup ve Duygusal Dinamikleri

İnsanlar sosyal varlıklar olduğundan, bir gruba ait olma hissi motive edicidir. Sosyal medya gruplarındaki aitlik duygusu, kullanıcıların dezenformasyon içeriklerine verdikleri duygusal tepkiler ve tartışmalar esnasındaki yorumlarına yansır. Her iki tarafta da haklı olduğunu düşünen insanlar vardır. Öte yandan ekonomik düzenlemelere, aşırıcılığa, popülizme ve özgürlükçülüğe yönelik söylemlerden faydalanan gruplar daha yaygındır. Kutuplaştırıcı tartışmalar, sosyal medyadaki farklı gruplar arasında büyürken, siyasi kutuplaşmalar dezenformasyon salgınının yayılmasını körükler (Osmundsen et al., 2021, ss. 999-1015). Böylece toplumsal kutuplaşma sosyal medyada daha güçlü ortaya çıkar ve devam eden tartışmalar nefreti ve taraflar arasındaki yanlış anlaşılmaları artırır. Sosyal medya mecralarının farklı görüşlerin örtüşmesini sağlayacak şekilde yeniden tasarlanması, dezenformasyonun azaltılmasına yardımcı olabilir (Metzler et al., 2022).

Birçok ülkede, karmaşık siyasi ve ekonomik konular hakkında kamuoyunun görüşlerini etkilemek için şüpheli haberler yayınlanır. Bazı sosyal medya kullanıcıları, gruplarda paylaşılan manipülatif duygusal bilgileri eleştirmeden inanmazlar (Roy et al., 2020, ss. 56-79). Haberi kimin paylaştığı, kaynağı ve güvenilirlik, kullanıcıların bilgiye yaklaşımını etkiler. Haberin yanlış olduğu fark edildiğinde, öfke ortaya çıkabilir (Hameleers et al., 2023, ss. 1699-1715). Dolayısıyla güçlü duygulara sahip insanlar, otomatik olarak bir gönderiye inanmaz ve onu paylaşmaya devam etmezler. Kaynağa duyulan güven ve dünya görüşü paylaşılacak bilgilerin değerlendirilmesini sağlar. Dezenformasyon içeriğini araştırmak ve kanıta dayalı çözümlere başvurmak kutuplaşmayı ve dezenformasyonun duygusal etkisini azaltabilir.

Bilişsel psikolojideki yanlış bilgiye olan inanca yönelik baskın fikir, bilginin zihindeki işleyişinden kaynaklanmaktadır (Ecker et al., 2022, ss. 13-29). Bu fikre göre insanlar, farklı derecelerde bilgiye yönelik doğru yargılarda bulunma yeteneğine sahiptir. Karşılaşılan bilginin tutarlılığını değerlendirmek için ön bilgiler kullanılır. Fakat katılım ve bilişsel kapasite genellikle bireysel farklılıklardan dolayı sınırlıdır (Lang, 2000). Sonuç olarak analitik düşünme becerisi azaldıkça bilgi duygusal işlenme eğilimine girer.

Analitik düşüncenin zayıf olması, bireyi dezenformasyona hem bilişsel hem de duygusal olarak savunmasız hâle getirir. Bu durum, bilgiye yönelik yargıda bulunurken onu duygularına güvenmeye teşvik eder (Holland et al., 2012). Duygular uyarılma konusunda daha hassas olduğundan (Berger & Milkman, 2013, s. 18) sık paylaşılırlar. Ayrıca olumsuz duygular uyandıran bilgiler, insanları partizanlığa daha duyarlı bir hâle getirebilir (Weeks & Garrett, 2019, ss. 236-250) ve içerik bireyin duygusal durumuyla etkileşime girebilir. Bu nedenle bireyin duygusal durumu, dezenformasyonu daha çekici hâle getirir (DeSteno et al., 2004, s. 43). Gollwitzer (2020) yapmış olduğu bir araştırmada, duygusal olarak hassas olan kişilerin, dezenformasyona maruz kalma ve yalan haberlere inanma olasılığını artırdığına dair bulgulara ulaşmıştır (ss. 1186-1197).

Duygunun dezenformasyon bağlamında uyarlanabilir olması, dezenformasyonun anlaşılmasında temel bir öneme sahiptir. Dezenformasyon içeriği, bireyin inancı ve dünya görüşlerinin yanı sıra yeni bilginin hangi duyguyu ortaya çıkaracağını belirler (Mercier, 2020, ss. 56-79). Bireyin sosyal kimliğiyle ilişkili dünya görüşünü tehdit eden bilgiyle karşılaşması, yoğun olumsuz duyguları tetikleyebilir ve bilginin reddedilmesine neden olabilir (Robertson et al., 2022). Bireyin bilgiyi reddetmesi ya da tutarlı bilgilere inanması, kişisel kimliği üzerinde onarıcı ve koruyucu bir etkiye yol açar (Wischnewski et al., ss. 1-11). Dahası birey, dezenformasyona karşı güdülere dayalı bir yetenek geliştirebilir. Güdüsel yetenek, bireyin tartışmalı konular hakkında dik bir duruş sergilemesine hizmet eder (Hareli & Hess, 2012, ss. 385-389). Sadece bilginin doğruluğuna odaklanmak, yanlış bilgiye duyulan inancın ve sosyal kimliğe dayanan sosyo-duygusal dinamiklerin yönetilmesine yardımcı olur (Rathje et al., 2022).

Dezenformasyonla Mücadele Bağlamında Yapay Zekâ Teknikleri

Çevrim içi dezenformasyon endişe verici boyutlara ulaştığından, sosyal medya platformları ve arama motorlarından özellikle infodemik bağlamda çözüm önerileri talep edilmektedir (Bontridder & Poullet, 2021). Bu nedenle çevrim içi dezenformasyonla mücadele için çeşitli teknolojik yöntemler geliştirilmiştir. Geliştirilen teknolojik yöntemler hem yanlış ve yanıltıcı içerikleri tespit etmek hem de bu içerikleri yapay zekâ kullanarak düzenlemeyi amaçlamaktadır. Dikkat edilmesi gereken nokta, şu ana kadar geliştirilen yapay zekâ tekniklerinin gerçek bilgiyi yalan bilgiden ayırt etme konusundaki yetersizliği ve uygunsuzluğudur. Bu teknikler özellikle bilgi edinme ve ifade özgürlüğü açısından sorunludur (Bontridder & Poullet, 2021).

Potansiyel dezenformasyon içeriklerini tespit etmek amacıyla oluşturulan metin analiz programları, içerikteki karmaşık ve yalan bilgileri ayrıştırmanın zorluğu nedeniyle negatif ya da pozitif eğilim gösterir. Ayrıca 2018 itibarıyla Facebook'un yapay zekâ sistemlerinin, İngilizce ve Portekizce dışındaki dillerde etkili olmadığı belirlenmiştir (Marsden et al, 2020, s. 17). Bu nedenle hatalı ve yalan haberlerin tespit edilmesi için hâlâ insan müdahalesi gerekse de teknoloji şirketleri yapay zekâ temelli akıl yürüten, öğrenebilen, bilgi toplayan, iletişim kuran ve algılayan yazılımlar geliştirmeye odaklanmıştır (Feldstein, 2019). Özellikle de deepfakeleri tespit etmek amacıyla sahte içerikten gerçek içeriği ayırt edebilen teknolojiler geliştirilmiştir. Bu amaçla, deepfake ürünler üreten algoritmaların benzerleri kullanılmıştır (Walorska, 2020, ss. 149-160).

2018 yılında yapılan araştırmalarda, yapay zekâ algoritmaları kullanılarak oluşturulan videolardaki aktörlerin, göz kırpmadıkları gözlemlenmiştir. Bu nedenle anormal göz kırpma davranışı arayan teknolojiler geliştirilmiştir. Fakat bu yöntem ortaya çıktığında, deepfake videolarına göz kırpabilen aktörlerin yerleştirildiği görülmüştür (Kertysova, 2018, s. 71). Deepfake ürünleri tespit etmek için kullanılan bir diğer yöntem, içeriğin yayılmadan önce doğrulanmasıdır. Görüntülerin, seslerin ve videoların oluşturuldukları anda tespit edilmesi, içerik yayılmadan önce doğruluğu teyit etmek üzere bir referans olarak kullanılabilir (Kertysova, 2018). Ancak bu yöntem farklı bir soruna neden olabilir. Bireyin her söylediğini, her yaptığını, nerede olduğunu doğrulama çabası, mahremiyet dâhil olmak üzere çok sayıda insan hakları ihlaline yol açar.

Dezenformasyonun teknoloji desteğiyle ortadan kaldırılması amacıyla oluşturulan raporlar, deepfake ürünleri saptamak amacıyla insan ve bot hesaplar arasındaki farkları bulmak için makine öğreniminden yararlanılmasını vurgulamaktadır. Deepfake ürünlerini tespit etmek için sosyal ağ grafik yapılarından sosyal medya hesap verilerine ve gönderi parametrelerine kadar metin içeriklerini analiz etmek amacıyla doğal dil işleme teknikleri kullanılmaktadır. Bu amaçla kitle kaynaklı teknikler de denenmektedir. Bu teknik sosyal ağ sağlayıcıları için de umut verici bir yaklaşım olarak benimsenmiştir. Sosyal ağ sunucuları, gönderileri görüntülemek için harcanan za-

man ve takipçi sayıları gibi hesap hareketlerini incelemeye almıştır (Akers et al, 2018, s. 5). Yapay zekâ teknikleri sayesinde, 2020 yılında Facebook'ta kullanılan sahte hesapların %99,6'sı tespit edilmiştir (Bontridder & Poullet, 2021, s. 32).

Filtreleme, içeriğin yüklenmesini ve yayınlanmasını engellemek için sunucular tarafından alınan bir önlemdir. Sunucular içeriklerin kaldırılmasını genellikle kullanıcı talepleri veya şikâyetleri doğrultusunda yaparken, içerik yayılmadan engellenir. Ancak bu önlem, sanal özel ağ adı verilen VPN (Virtual Private Network) aracılığıyla aşılabilir. Yine de kullanıcının görmek istediği içeriklere öncelik vermesi, içeriğin algoritmalar ve ağ tabanı aracılığıyla otomatik olarak sıralanmasına, belirli hesaplar tarafından paylaşılan dezenformasyona dayalı içeriklerle karşılaşmamasına ya da daha az karşılaşmasına olanak sağlar.

Hesapların devre dışı bırakılması veya askıya alınması, dezenformasyonla mücadelede sosyal ağ sunucularının kullandığı başka bir tekniktir. Teknoloji sağlayıcıları, kullanıcılar hizmet şartlarına uymadıklarında veya kötüye kullandıklarında önlem alırlar. Örneğin; ABD eski başkanı Donald Trump'ın Twitter ve Facebook hesapları, sosyal ağ platformlarının hüküm ve koşullarını karşılamadığı gerekçesiyle askıya alınmıştır (Jang et al., 2021, ss. 25-31). Ayrıca yenilikçi teknolojiler de kullanıcıların veri tabanlarını korumak için geliştirilmektedir. Prensipte bu teknolojiler, kullanıcı verilerini özel şirketler ve devletlerden daha iyi korumayı amaçlamaktadır. Çünkü yapay zekâ tabanlı sorunu çözmenin en etkili yollarından biri dezenformasyonda kullanılan teknolojinin kendisidir. Dezenformasyon içeriğini tanımlamak, aynı yapay zekânın ana mekanizması kullanılmaktadır. Sinir ağları yapay metin ürettiğinden, metnin benzerliklerini ve özelliklerini bilirler, bu da onları bu ağlardan çıkan içeriğin tespit edilmesinde etkili hâle getirir (Kreps et al., 2020, ss. 104-117). Zellers (2019) sinir ağlarının sahte haberi algılama araştırmalarında, insan ve makine tarafından yazılmış metinleri tespit etme konusunda %92 doğruluğa ulaştığını ifade etmiştir.

Teknoloji tabanlı çözümler, sahte metinleri tanımlamak ve üst verilerin analizini içerir. Web sitesinin kendisi, algoritmalar aracılığıyla üretilen metnin IP adresinin dezenformasyon amaçlı olup olmadığını tanımlamak için kullanır. 2016 ABD seçimlerine müdahalede kullanıldığı iddia edilen Facebook, bu iddialara yanıt vermek için dijital tıp ve istihbarat uzmanlarından faydalanmıştır. Elde edilen veriler, gönderilerin tümünün büyük ölçüde Kuzey Afrika, Latin Amerika ve ABD'deki hedef kullanıcıları kutuplaştırmayı amaçladığını göstermiştir (Isaak & Hanna, 2018, ss. 56-59). Yanlış bilgi yaymak isteyenlerle, buna karşı koymak isteyenlerin mücadelesi bir kedi-fare oyununa dönmüştür. Karşı önlemler için orijinal fikirlerin kullanılması, dezenformasyon kaynaklarının tespitinde çeşitli zorlukların ortaya çıkmasına neden olmuştur. Asıl zorluk, yapay zekâ tarafından üretilen sahte içeriklerin karmaşıklığıyla ilişkilidir.

Dezenformasyonun Neden Olduğu Etik Kaygılar

Dezenformasyon içerikleri birden fazla hedefe odaklanır. Yanlış, önyargılı ve yanıltıcı bilgiler bireyi, kurumu ve toplumu hedef alırken, tık tuzağı kullanarak sitenin tıklanma oranını artırmak ve ekonomik çıkar elde etmek de amaçlanabilir (Vizoso e al., 2017, ss. 291-300). Örnekler, dezenformasyonun farklı kullanım amaçları olduğunu göstermektedir. İnternet özellikle de sosyal medya, kullanıcıların her türlü bilgiyi kolayca ve hızlı bir şekilde yaymasını sağlar. Bu nedenle siyasetçiler, gazeteciler, siyasi parti başkanları, milletvekilleri ve şirketler dezenformasyonla karşı karşıya kalırlar. Literatürden elde edilen bilgiler, sosyal medyanın yalan haber içeriklerini yaymak için etkili bir şekilde kullanıldığını göstermektedir. ABD'de yapılan bir araştırma, nüfusun %55'inin sahte sosyal medya hesaplarının yalan haberlerin yayılmasından sorumlu olduğuna inandığını ortaya koymuştur (Richter, 2019).

Sahte haberler bilginin gerçekliğini değiştirme eğilimindedir ve insanlarda bir doğruluk hissi uyandırabilir (Jackson, 2019). Doğruluk duygusunun insanlarda yalan haberleri özümsemeye neden olması, etik dışı bir davranış olarak görülür. Etik olmayan bilgi davranışı sadece dezenformasyonda değil, aynı zamanda siyaset, bilim, ekonomi ve finans endüstrisinde de mevcuttur. Etik olmayan bilgi, politik ekonomi söz konusu olduğunda, halkın yanıltıcı bilgilere dayanarak ekonomik seçimler yapmasına ve refahlarını kaybetmelerine neden olabilir. Halktan bilgi saklamak da ekonomik refah gelişimini etkileyebilir. Bu nedenle siyasi dezenformasyon içeriklerini engellemek, ekonomik refahın azalma riskini de engeller.

Sahte haberler ekonomi, sağlık, politika, istihdam ve ticaret gibi alanlarda ülkeleri derinden etkiler. Hindistan hükûmeti dezenformasyonla mücadele için yanlış bilgiler yayan web sitelerini kapatmakla kalmamış, aynı zamanda halka internete erişim engeli de getirmiştir. Bu şekilde ekonomi üzerinde olumsuz etkileri olan etik dışı dezenformasyon içerikleri ekonomik rekabete yol açan maliyetli sorunun çözümüne katkı sağlamıştır (Dhir et al., 2018, ss. 141-152).

Etik açıdan en önemli endişelerden biri gizlilik konusudur. Yapay zekâ ile üretilmiş dezenformasyona karşı kullanılan tekniklerin geliştirilmesi için büyük ölçekli verilerin toplanması gerekir. Toplanan veriler, konum, tarama geçmişi ve satın alma gibi hassas kişisel bilgileri içerir. Bu bilgiler başta reklamcılık olmak üzere birçok alanda kullanılabilir. Büyük ölçekli veri toplayan şirketlerin, kuruluşların ve sosyal medya ağlarının bireylerin gizliliğini koruduklarından ve onların rızası olmadan bu bilgileri paylaşmayacaklarından veya satmayacaklarından emin olunmalıdır.

Şeffaflık, etik bir konudur. Veri toplayan şirketler ve kuruluşlar, verileri nasıl kullanacakları konusunda kullanıcılarına şeffaf olmalıdır. Kullanıcı verilerin ne için toplandığını, nerelerde kullanılacağını ve hangi kurum ya da kuruluşlarla paylaşılacağını bilmelidir (Malik & Dhiman, 2022, ss. 1-4). Toplanan veriler, şirketlerin ya da kurumların veri

tabanlarında depolanırken, veri tabanlarının siber saldırılara karşı savunmasız kalması, verilerin çalınmasına ya da kaybolmasına neden olabilir. Bu durum kullanıcılar açısından ciddi sonuçlar doğurur. Bu açıdan veri toplayan kurum ve kuruluşlar, topladıkları verileri korumak için yeterli güvenlik önlemleri almalıdırlar. Etik ve sorumlu davranışlar, bireylerin mahremiyetinin korunması ve güvenliklerinin sağlanması açısından özellikle önemlidir.

SONUÇ VE ÖNERİLER

Yapay zekâ ile oluşturulmuş dezenformasyon içerikleri, toplumun temel yönelimlerini etkileyerek siyasi, etik ve toplumsal sorunlara neden olmuştur. Yapay zekâ sayesinde otonom olarak oluşturulan dezenformasyon içerikleri, insanın problem çözme yeteneğinin ötesine geçmiştir. Yapay zekâ dezenformasyona karşı çeşitli avantajlar sunmasına rağmen, bir kontrol mekanizmasının bulunmaması özellikle siyasi, etik ve toplumsal alanda ciddi riskler oluşturmuştur. Bu risklerin başında tamamen otonom sistemlerden oluşan ve insan müdahalesini devre dışı bırakan yapay zekâlı dezenformasyon içerikleri gelmektedir. Yapay zekâya sahip dezenformasyon içeriklerinin siyasi, etik ve toplumsal sonuçlarını ve bu sonuçların etkilerini azaltmak amacıyla tespit edilen çözüm önerilerini şu şekilde sıralamak mümkündür:

- Yapay zekâlı dezenformasyon içerikleri provokatörlerin ve trollerin etkileşimiyle ortaya çıkarak bireyi ve toplumu hedef almıştır. Bu nedenle devlet ve özel sektör paydaşları dezenformasyonu bir silah hâline getirenlerin kabiliyetlerini azaltacak politikalar benimsemelidir.
- Dezenformasyon günümüzde gelişmiş teknolojilerin bir sonucudur. Bu nedenle deepfake teknolojilerine yatırım yapanlar mercek altına alınmalı ve dezenformasyona karşı istihbarat stratejileri geliştirmelidir.
- Yapay zekâya sahip dezenformasyon içerikleri hızlı bir gelişim içindedir. Bu nedenle siyaset kurumunun korunması, etik ve toplumsal kaygıların azaltılması için araştırmacıların dezenformasyon içeriklerini tespit etme çalışmaları devam etmelidir.
- Siyasi, etik ve toplumsal etkileri olan yapay zekâlı dezenformasyon içeriklerini üretmek günümüzde oldukça kolaylaşmıştır. Bazı çevrim içi paydaşların sunduğu deepfake içerik üretimine izin veren yapay zekâ algoritmalarının halkın erişimine kapatılması, dezenformasyonla mücadele seçeneklerinin arasında olmalıdır.
- Yapay zekâlı dezenformasyon içerikleri üst düzey sanal gerçeklik sağladığından, toplumun bilgi alma ve bilgiyi değerlendirme kapasitesini artırmak amacıyla okullarda medya okuryazarlığı dersleri teşvik edilmelidir.
- Devlet ve özel sektörün iş birliği, dezenformasyonun toplumsal etkilerine karşı yararlı sonuçlar alınmasını sağlamıştır. Paydaşların iş birlikli çabalarını koordine et-

mek zor olsa da iş birliği içinde çalışılarak dezenformasyonun toplumsal güvenliğin zayıflamasına neden olan etkileri azaltılabilir.

- Dezenformasyonla mücadele tek taraflı değildir. Bu nedenle yapay zekâya sahip dezenformasyon içeriklerinden korunmak için gazetecilerin, medya kuruluşlarının, sivil aktörlerin ve diğer paydaşların iş birliği içinde çalışması gerekmektedir.
- Dezenformasyon içerikleri siyasi, etik ve toplumsal olarak birden fazla amaçla üretilir. Hedef çeşitliliğine karşı mücadele etmenin en etkili yollarından biri OSINT (Open Source Intelligence) adı verilen açık kaynak istihbaratıdır. Açık kaynak istihbaratı, belirli periyodlarla akışa sokulan dezenformasyon akışını engelleyebilir.
- Yapay zekâlı dezenformasyon içerikleri tarafsızlığı, kamu değerleri ve demokratik ilkeler hakkındaki endişeleri artırmıştır. Bu nedenle teknoloji endüstrisindeki önemli oyuncular, özellikle internet güvenliği açısından sosyal medya platformlarına anti-dezenformasyon araçları geliştirmeyi finanse etmelidir.
- Ticari amaçlarla kullanılan yapay zekâya sahip dezenformasyon içerikleri, yeni sorunların ortaya çıkmasına neden olmuştur. Bu nedenle ticari amaçlı dezenformasyon içeriklerine karşı hükûmetlerin mevcut yasaları destekleyici yasalar çıkarması gerekmektedir.
- Yapay zekâya sahip dezenformasyon içerikleri güç kontrolünün kaymasına neden olmuş, dolayısıyla demokratik girdilerin meşrutiyetini sorgulatmıştır. Bu nedenle farklı siyasi ve yasal sonuçlar dikkate alınmalıdır.
- Bazı sosyal medya ağları, yapay zekâ algoritmaları kullanarak büyük çaplı veriler toplamaktadır. Bu verilerin çalınma ve satılma riski etik sorunların ortaya çıkmasına neden olmuştur. Bu nedenle sosyal medya sunucuları, toplanılan verilerin ne amaçla kullanıldığı ve nasıl korunduğu hakkında kamuoyuna karşı hesap verebileceği politikalar benimsemelidir.

KAYNAKÇA

Akers, L., & Gordon, J. S. (2018). Using Facebook for large-scale online randomized clinical trial recruitment: effective advertising strategies. *Journal of Medical Internet Research, 20(11)*, e290.

Alvares, J., & Salzman-Mitchell, P. (2019). The succession myth and the rebellious AI creation: Classical narratives in the 2015 film Ex Machina. *Arethusa, 52(2)*, 181-202.

Berger, J., & Milkman, K. L. (2013). Emotion and virality: what makes online content go viral? *NIM Marketing Intelligence Review, 5(1)*, 18.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy, 3*, e32.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The socialbot network: when bots socialize for fame and money. *In Proceedings of the 27th annual computer security applications conference*, (s. 93-102).

Brief, C. P. (2021). *AI and the Future of Disinformation Campings*.

Bukovská, B. (2020). The European Commission's Code of conduct for countering illegal hate speech online. *Algorithms*.

Candi, M. R. (2018). Social strategy to gain knowledge for innovation. *British Journal of Management, 29(4)*, 731-749.

Center, G. E. (2020). Pillars of Russia's disinformation and propaganda ecosystem. *US Department of State*.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Clif, L. Rev.* , 1753.

Collins, A., & Ebrahimi, T. (2021). Risk governance and the rise of deepfakes.

de Lima Salge, C. A., & Berente, N. (2018). Is that social bot behaving unethically? *Communications of the ACM, 60(9)*, 29-31.

DeSteno, D., Pretty, R. E., Rucker, D. D., Wegener, D. T., & Braverman, J. (2004). Discrete emotions and persuasion: the role of emotion-induced expectancies. *Journal of Personality and Social Psychology, 86(1)*, 43.

Dhir, A., Yossatorn, Y., Kaur, P., & Chen, S. (2018). Online social media fatigue and psychological well-being—A study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International Journal of Information Management, 40*, 141-152.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1(1)*, 13-29.

Feldstein, S. (2019). *The global expansion of AI surveillance.* Washington: Carnegie Endowment for International Peace.

Gollwitzer, A., Martel, C., Brady, W. J., Parnamets, P., Freedman, I. G., Knowles, E. D., & Van Bavel, J. J. (2020). Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour, 4(11)*, 1186-1197.

Hameleers, M., Humprecht, E., Möller, J., & Lühring, J. (2023). Degrees of deception: The effects of different types of COVID-19 misinformation and the effectiveness of corrective information in crisis times. *Information, Communication & Society, 26(9)*, 1699-1715.

Hareli, S., & Hess, U. (2012). The social signal value of emotions. *Cognition & Emotion, 26(3)*, 385-389.

Howard, P. N., & Kollanyi, B. (2016). *Bots, Strongerin and Brexit: Computational Propaganda During the UK-EU Rerefendum*.

Ireton, C., & Posetti, J. (2018). *Journalism, fake news & disinformation: handbook for journalism education and training.* Paris : Unesco Publishing .

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer, 51(8)*, 56-59.

Ivakhiv, O. (2016). Information state of system estimation. *International Journal of Computing, 15(1)*, 31-39.

Jackson, P. C. (2019). *Introduction to artificial intelligence. Courier Dover Publications*.

Jacobs, G., Caraça, J., Fiorini, R., Hoedl, E., Nagan, W. P., Reuter, T., & Zucconi, A. (2018). The future of democracy: Challenges & prospects. *Cadmus, 3(4)*, 7-31.

Jang, H., Rempel, E., Roth, D., Carenini, G., & Janjua, N. Z. (2021). Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research, 23(2)*, 25-31.

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights, 29(1-4)*, 55-81.

Kopf, R. K., Nimmo, D. G., Ritchie, E. G., & Martin, J. K. (2019). Science communication in a post-truth world. *Frontiers in Ecology and the Environment, 17(6)*, 310-312.

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science, 9(1)*, 104-117.

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences, 467*, 312-322.

Malik, D. P., & Dhiman, D. B. (2022). Science Communication in India: Current Trends and Future Vision. *Journal of Media & Management, 4(5)*, 1-4.

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review, 34(4)*, 754-772.

Marsden, C., Meyer, T., & Brown, I. (2020). Platform values and democratic elections: How can the law regulate digital disinformation? *Computer Law & Security Review, 36*, 105373.

Metzler, H., Pellert, M., & Garcia, D. (2022). Using social media data to capture emotions before and during COVID-19.

Miranda, S. M., & Yetgin, E. (2016). Are social media emancipatory or hegemonic? Societal effects of mass media digitization in the case of the SOPA discourse. *MIS quartrely, 40(2)*, 303-330.

Mork, A., Hale, J. A., & T., R. (2020). *Fake for real: a history of forgery and falsification*.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Scinece Review, 115(3)*, 999-1015.

Pierson, A. E., Brady, C. E., Clark, D. B., & Sengupta, P. (2023). Students' epistemic commitments in a heterogeneity-seeking modeling curriculum. *Cognition and Instruction, 41(2)*, 125-157.

Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (2022). *People think that social media platforms do (but should not) amplify divisive content*.

Richter, A. (2019). Accountability and media literacy mechanisms as a counteraction to disinformation in Europe. *Journal of Digital Media & Policy, 10(3)*, 311-327.

Roy, M., Moreau, N., Rousseau, C., Mercier, A., Wilson, A., & Atlani-Duault, L. (2020). Ebola and localized blame on social media: Analysis of Twitter and Facebook conversations during the 2014–2015 Ebola epidemic. *Culture, Medicine, and Psychiatry, 44*, 56-79.

Russell, S. J., & Norving, P. (2016). *Artificial Intelligence: A Modern Approach*. London: Pearson Education Limited.

Salzman, J., & Ruhl, J. B. (2019). Environmental Law. *Currencies and the commodification of environmental law* (s. 3-90). içinde

Satter, R. (2019). *Social media timeout as French election reaches final state*.

Simonite, T. (2019). Are You For Real? *Wired, 27(7)*, 24-25.

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology, 16(5)*, 472-481.

Stupp, C. (2019). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *The Wall Street Journal, 30(8)*.

Vincent, V. U. (2021). Integrating intuition and artificial intelligence in organizational decision-making. *Business Horizons, 64(4)*, 425-438.

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against Hi-Tech misinformation. *Media and Communication, 9(1)*, 291-300.

Walorska, A. M. (2020). Redesigning Organizations: Concepts for the Connected Society. *The Algoritmic Society* (s. 149-160). içinde

Weeks, B. E., & Garrett, R. K. (2019). Emotional characteristics of social media and political misperceptions. *Journalism and truth in an age of social media*, 236-250.

Wischnewski, M., Bernemann, R., Ngo, T., & Kramer, N. (2021). Disagree? You must be hot! How beliefs shape twitter profile perceptions. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (s. 1-11).

Young, S. D., Crowley, J. S., & Vermund, S. H. (2021). Artificial intelligence and sexual health in the USA. *The Lancet Digital Health, 3(8)*, 467-468.

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. *In Proceedings of the IEEE/CVF international conference on computer vision*, (s. 9459-9468).

Zellers, R., Holtzman, A., Rashking, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advnces in Neural Information Processing Systems, 32*.

Hakem değerlendirmesi/Peer review:
Dış bağımsız/Externally peer reviewed

Çıkar çatışması/Conflict of interest:
Yazar çıkar çatışması bildirmemiştir/The author have no conflict of interest to declare

Finansal destek/Grant support:
Yazar bu makalede finansal destek almadığını beyan etmiştir/The author declared that this article has received no financial support.