Deepfake Kimliklerin Artan Tehditleri

Homeland Security

Deepfake Kimliklerin Artan Tehditleri

Özet

Deepfake'ler, daha geniş ve giderek yaygınlaşan **sentetik medya** başlığı altında yer alan yeni bir tehdit türüdür. Bu içerikler, **yapay zekâ/makine öğrenmesi (AI/ML)** kullanılarak gerçekte hiç yaşanmamış olaylara ait **inandırıcı ve gerçekçi video, görüntü, ses ve metinler** üretmektedir.

Sentetik medyanın birçok kullanım alanı masum eğlence amaçlı olsa da, bazı uygulamaları ciddi riskler barındırmaktadır.

Deepfake ve sentetik medyadan kaynaklanan tehdit, esas olarak bunları üretmek için kullanılan teknolojiden değil; **insanların gördüklerine inanma eğiliminden** kaynaklanmaktadır. Bu nedenle deepfake ve sentetik medya içeriklerinin etkili olabilmesi için son derece gelişmiş veya kusursuz olmaları gerekmemektedir. Yanlış ve yanıltıcı bilgilerin yayılmasında oldukça basit içerikler dahi etkili olabilmektedir.

Alanında uzman kişilerle yapılan çok sayıda görüşmeye dayanarak, sentetik medyadan kaynaklanan mevcut tehdidin **ciddiyetinin ve aciliyetinin**, soruyu kime sorduğunuza bağlı olarak değiştiği görülmektedir. Endişe yelpazesi, *"acil ve ciddi bir tehdit"* görüşünden *"panik yapmaya gerek yok, ancak hazırlıklı olunmalı"* yaklaşımına kadar uzanmaktadır.

Müşterilerin potansiyel bir tehdidin nasıl ortaya çıkabileceğini ve bu tehdidin neye benzeyebileceğini daha iyi anlayabilmeleri için; **ticaret, toplum ve ulusal güvenlik** alanlarına özgü çeşitli senaryolar ele alınmıştır.

Bu senaryolardan herhangi birinin gerçekleşme ve başarıya ulaşma olasılığı; **kullanılabilir deepfake'lerin üretilmesi için gereken maliyet ve diğer kaynaklar azaldıkça** artacaktır. Nitekim yapay zekâ/makine öğrenmesi içermeyen tekniklerin yaygınlaşmasıyla birlikte sentetik medya üretimi de zamanla daha kolay hâle gelmiştir.

Sorunun çok boyutlu doğası gereği, **tek ve evrensel bir çözüm bulunmamaktadır**. Ancak teknolojik yenilik, eğitim ve düzenleyici yaklaşımlar; tespit ve zarar azaltma önlemlerinin mutlaka bir parçası olmak zorundadır.

Başarıya ulaşabilmek için; özel ve kamu sektöründeki paydaşlar arasında güçlü bir iş birliği gerekecektir. "Kurumsal silo" (stovepiping) gibi mevcut engellerin aşılması ve **sivil özgürlükler korunurken** bu yeni tehditlere karşı toplumsal güvenliğin sağlanması ancak bu şekilde mümkün olacaktır.

EKİP

- **Tina Brooks**, Verizon
- **Princess G.**, Transportation Security Administration
- **Jesse Heatley**, JP Morgan Chase & Co.
- **Jeremy J.**, United States Secret Service
- **Scott Kim**, Experian
- **Samantha M.**, Federal Bureau of Investigation
- **Sara Parks**, National Cyber-Forensics & Training Alliance
- **Maureen Reardon**, Melian LLC
- **Harley Rohrbacher**, National Cyber-Forensics & Training Alliance
- **Burak Şahin**, Deloitte & Touche
- **Shani S.**, Federal Bureau of Investigation
- **James S.**, U.S. Department of Homeland Security
- **Oliver T.**, Federal Bureau of Investigation
- **Richard V.**, Federal Bureau of Investigation

TERİM KULLANIMINA İLİŞKİN AÇIKLAMALAR

“Kleenex”, “Xerox” ve “Photoshop” terimleri, başlangıçta tek bir üreticiye ait belirli ürünleri ifade ederken; günümüzde yaygın kullanım (ya da yanlış kullanım) sonucu, üreticisinden bağımsız olarak **bir ürün sınıfını temsil eden genel adlar** hâline gelmiştir. Geniş kitleler nezdinde “deepfake” teriminin de benzer biçimde, **her türlü sentetik medyayı kapsayan** bir anlama büründüğü görülmektedir.

Ekibimiz bu terimin bu şekilde yanlış veya genişletilmiş kullanımını **desteklememektedir**; ancak **pragmatik bir yaklaşım** benimsemektedir. Bu nedenle, bu çalışmada zaman zaman “deepfake” terimi, teknik olarak gerçek bir “deepfake” olup olmadığına bakılmaksızın, **her türlü sentetik medyayı ifade edecek şekilde** kullanılacaktır.

Benzer şekilde, bu çalışmada **Yapay Zekâ (Artificial Intelligence – AI)**, **Makine Öğrenmesi (Machine Learning – ML)** ve **Derin Öğrenme (Deep Learning – DL)** terimleri;

deepfake'ler ve diğer sentetik medya türleri bağlamında sıkça anılmaktadır. Başka bölümlerde de belirtildiği üzere, makine öğrenmesi ve derin öğrenme, **yapay zekâya imkân sağlayan alt kümeler** olarak değerlendirilebilir. Bu nedenle bazı durumlarda, bazı uzmanların "ML" veya "DL" terimlerini tercih edebileceği yerlerde, **"AI/ML" ifadesi** kullanılacaktır.

Giriş

2017 yılının sonlarında, Motherboard, internette ortaya çıkan bir videoyu haberleştirdi. Bu videoda, **Gal Gadot**'un yüzü mevcut bir pornografik videonun üzerine yerleştirilmiş ve oyuncunun videoda tasvir edilen eylemleri gerçekleştirdiği izlenimi yaratılmıştı.[1] Video sahte olmasına rağmen, görüntü kalitesi sıradan bir izleyicinin videoya inanmasına—ya da gerçeği umursamamasına—yetecek düzeydeydi.

Kendisini "deepfakes" olarak tanıtan anonim bir kullanıcı, sosyal medya platformu **Reddit** üzerinden bu videonun yaratıcısı olduğunu iddia etti.[2]

"Deepfake" terimi, bu tür manipüle edilmiş içeriklerin (ya da "sahte"lerin) üretilmesinde kullanılan teknolojinin **derin öğrenme (deep learning)** tekniklerine dayanmasından türemiştir. Derin öğrenme, **makine öğrenmesinin** bir alt kümesini oluşturur; makine öğrenmesi ise **yapay zekânın** bir alt alanıdır. Makine öğrenmesinde, bir model belirli bir görevi yerine getirebilmek için eğitim verilerini kullanarak geliştirilir. Eğitim verisi ne kadar kapsamlı ve güçlü olursa, model de o kadar iyi performans gösterir. Derin öğrenmede ise model, veriler içindeki özellik temsillerini otomatik olarak keşfedebilir; bu sayede verilerin sınıflandırılmasına veya ayrıştırılmasına olanak tanır. Bu anlamda modeller, "daha derin" bir düzeyde eğitilmiş olur.[3]

Derin öğrenme ile incelenebilen veriler yalnızca insanlara ait görüntü ve videolarla sınırlı değildir. Her türlü nesneye ait görüntü ve videoların yanı sıra **ses ve metin** de bu kapsama girer. 2020 yılında, **Dave Gershgorn**, OneZero için yaptığı bir haberde, **OpenAI** internet sitesinde ünlü sanatçılara ait "yeni" müziklerin yayımlandığını aktarmıştır.[4] Yaşayan ya da vefat etmiş tanınmış sanatçıların mevcut kayıtlarını kullanan programcılar; **Elvis**, **Frank Sinatra** ve **Jay-Z** gibi isimlere ait gerçekçi yeni şarkılar üretebilmiştir. Jay-Z'nin şirketi **Roc Nation LLC**, bu kayıtların kaldırılması için **YouTube**'a dava açmıştır.[5]

Yapay zekâ tarafından üretilen metinler, giderek büyüyen bir başka deepfake türü olarak karşımıza çıkmaktadır. Araştırmacılar, görüntü, video ve ses deepfake'lerinde tespit amacıyla kullanılabilecek çeşitli zayıflıklar belirlemiş olsa da, **deepfake metinlerin tespiti çok daha zordur**.[6] Kullanıcıların çoğu zaman gayriresmî olan mesajlaşma tarzlarının, deepfake teknolojisi kullanılarak taklit edilmesi de ihtimal dışı değildir.

Tüm bu deepfake medya türleri—**görüntü, video, ses ve metin**—belirli bir kişiyi veya o kişinin temsiliyetini simüle etmek ya da değiştirmek amacıyla kullanılabilir. Deepfake'lerin temel tehdidi de budur. Ancak bu tehdit yalnızca deepfake'lerle sınırlı

değildir; **dezenformasyon amacıyla kullanılan tüm "Sentetik Medya" alanını** kapsamaktadır.

Sadece "Deepfake"lerden İbaret Değil – "Sentetik Medya" ve Dezenformasyon

Deepfake'ler, aslında daha genel bir kavram olan **"sentetik medya"** veya **"sentetik içerik"** kategorisinin bir alt kümesini oluşturmaktadır. Konuya ilişkin birçok popüler makale, sentetik medyayı; özellikle otomatik biçimde olmak üzere, **yapay zekâ/makine öğrenmesi (AI/ML)** kullanılarak oluşturulan veya değiştirilen her türlü medya olarak tanımlamaktadır.

Ancak pratik açıdan bakıldığında; **kolluk kuvvetleri ve istihbarat toplulukları** içinde sentetik medya, genellikle daha geniş bir şekilde ele alınmaktadır. Bu kapsamda sentetik medya; ister **dijital ya da yapay yollarla üretilmiş** (örneğin bilgisayar tarafından oluşturulmuş kişiler), isterse **analog veya dijital teknolojiler kullanılarak değiştirilmiş ya da manipüle edilmiş** tüm medya türlerini içermektedir.

Örneğin, fiziksel ses bantları manuel olarak kesilip birleştirilerek kelimeler veya cümleler çıkarılabilir ve böylece bir kaydın genel anlamı değiştirilebilir. **"Cheapfake"** olarak adlandırılan içerikler de sentetik medyanın bir başka türüdür. Bu içeriklerde, gözlemcinin bir olaya ilişkin algısını değiştirmek amacıyla **basit dijital teknikler** kullanılır. Bu çalışmanın başka bölümlerinde yer alan cheapfake örneklerinde, konuşmanın yavaşlatıldığı ya da videonun hızlandırıldığı görülmektedir.

Bilim ve teknoloji sürekli olarak ilerlemektedir. Deepfake'ler ile otomatik içerik oluşturma ve değiştirme teknikleri, **görsel, işitsel ve metinsel içeriği değiştirmek veya üretmek için geliştirilen en güncel araçları** temsil etmektedir. Ancak bu teknolojilerin ortaya koyduğu temel fark, **ne kadar kolay üretilebildikleri ve bu üretimin ne kadar yüksek kalitede yapılabildiğidir**.

Geçmişte, sıradan izleyiciler (ya da dinleyiciler) sahte içerikleri kolaylıkla tespit edebiliyordu. Günümüzde ise bu her zaman mümkün olmayabilir. Bu durum, yanlış bilgi (misinformation) veya kasıtlı dezenformasyon yaymak isteyen her türlü aktörün; **öncekilere kıyasla çok daha gerçekçi görüntü, video, ses ve metin içeriklerini** kampanyalarında kullanabilmesine olanak tanımaktadır.

Deepfake'ler Nasıl Üretilir ve Nasıl Kullanılabilir?

2017 yılında ortaya çıkan ilk deepfake'ten bu yana, deepfake ve ilgili **sentetik medya teknolojilerinde** çok sayıda gelişme yaşanmıştır. Aşağıda yer alan zaman çizelgesi; en bilinen ve temsil gücü yüksek deepfake örneklerinin yanı sıra bazı **"cheapfake"** örneklerini ve ayrıca başlangıçta deepfake kullanıldığı düşünülen ancak sonradan bunun **kanıtlanamadığı** bir olaya ilişkin bir örneği de içermektedir.

Bu rapora ek olarak, söz konusu örneklerin özetlerini ve daha fazla bilgi için bağlantıları içeren bir **ek bölüm (addendum)** de sunulmaktadır.

Kaynak: Increasing Threats of Deepfake Identities, Homeland Security

https://media.licdn.com/dms/document/media/D4D1FAQFY4uuhlk5DdQ/feedshare-document-pdf-analyzed/0/1722636799794?e=1723680000&v=beta&t=ZnuTkIVVEsSjMVqwinmJCOLkCXG36pAPa72JCfypIG4

Derin Sahtecilik ve Sentetik Medya örnekleri internet bağlantıları

Obama Buzzfeed: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video debunk-buzzfeed

Jim Acosta Doctored Video: https://apnews.com/article/entertainment-north-america-donald trump-us-news-ap-top-news-c575bd1cc3b1456cb3057ef670c7fe2a

Jennifer Lawrence- Steve Buscemi: https://fortune.com/2019/01/31/what-is-deep-fake-video/

David Beckham Anti-Malaria PSA: https://www.campaignlive.com/article/deepfake-voice-tech-used good-david-beckham-malaria-campaign/1581378

World Leaders Sing “Imagine”: https://scifi.radio/2019/05/29/watch-world-leaders-sing-for-peace in-canny-ais-imagine-video/

Dali Museum: https://www.dezeen.com/2019/05/24/salvador-dali-deepfake-dali-musuem-florida/

Bill Hader Impressions: https://www.fastcompany.com/90353902/bill-haders-al-pacino impression-gets-even-more-real-and-creepy-with-the-help-of-deepfakes

Nancy Pelosi Doctored Video: https://www.usatoday.com/story/news/factcheck/2020/08/11/fact check-video-pelosi-altered-and-selectively-edited/3332920001/

Mark Zuckerberg: https://www.technologyreview.com/2019/06/12/134992/facebook-deepfake zuckerberg-instagram-social-media-election-video/

Joe Rogan: https://www.maxim.com/news/joe-rogan-audio-and-video-deepfake-2019-12 Nixon/Moon Landing: https://www.newsweek.com/richard-nixon-deepfake-apollo-disinformation mit-1475340

Queen's Christmas Speech: https://www.independent.co.uk/news/uk/home-news/queen deepfake-channel-4-christmas-message-b1778542.html

Tom Cruise TikToks: : https://www.theverge.com/22303756/tiktok-tom-cruise-impersonator deepfake

Pennsylvania Cheerleader Case: https://www.buckscountycouriertimes.com/story/news/2021/05/14/da-chalfont-woman-may-not have-used-deepfake-tech-harassment-vipers-cheerleading/4992798001/

Anthony Bourdain Documentary: https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of a-deepfake-anthony-bourdain-voice

Zaman çizelgesinde, deepfake üretiminde kullanılan en yaygın teknikler gösterilmektedir. Bu tekniklerin ilki, deepfake ve AI/ML teknolojilerinden önceye dayanan “yüz değiştirme (face swap)” yöntemidir.

1990’lı yıllarda, Adobe Photoshop gibi görüntü düzenleme yazılımlarının ticari olarak yaygınlaşmasıyla birlikte, bilgisayara sahip olan herkes için bir görüntüyü değiştirmek—örneğin bir kişinin yüzünü veya başını başka bir kişinin vücudu üzerine yerleştirmek—mümkün hâle gelmiştir.

Günümüzde ise inandırıcı bir yüz değiştirme işlemi üretmek için kullanılan teknoloji büyük ölçüde yapay zekâya dayanmaktadır. Bu teknoloji, bir saldırganın bir kişinin yüzünü başka bir kişinin yüzü ve vücudu üzerine gerçekçi biçimde yerleştirmesine olanak tanır. Yüz değiştirme işlemi için Encoder (kodlayıcı) veya Derin Sinir Ağı (Deep Neural Network – DNN) teknolojileri kullanılabilmektedir.

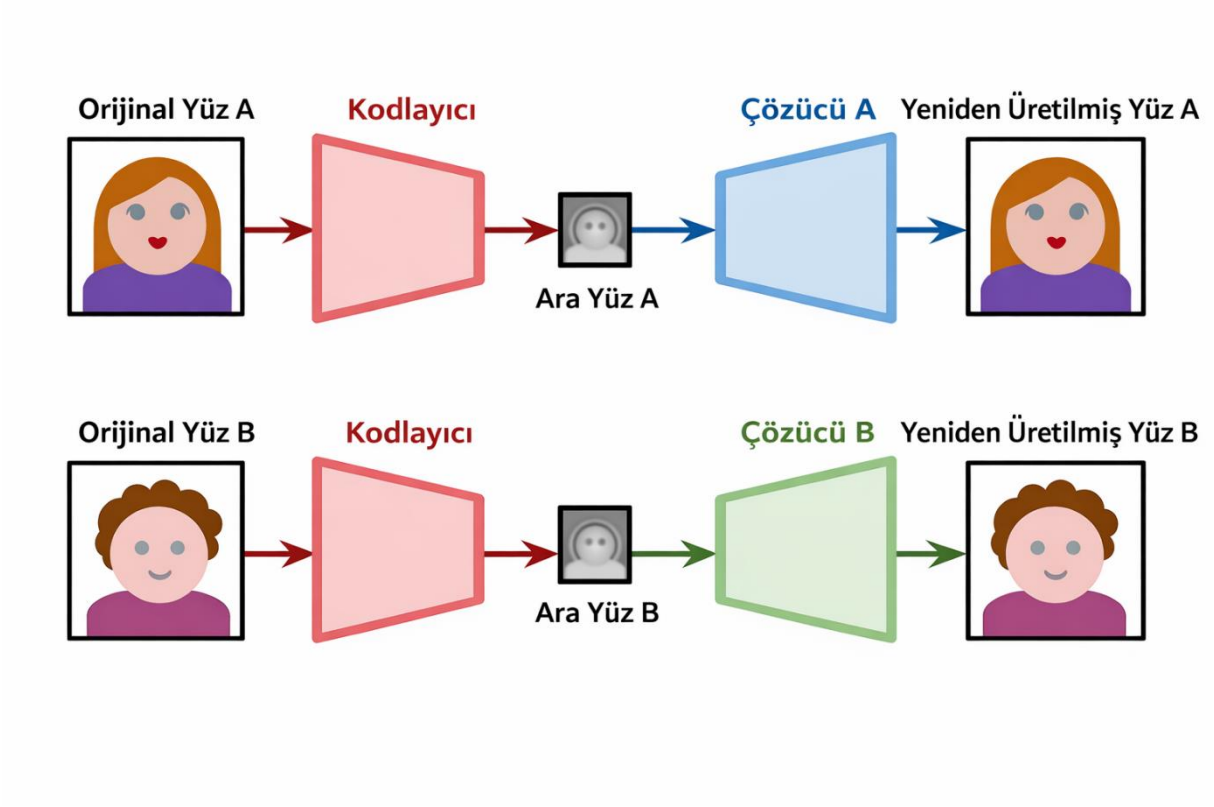
Bir autoencoder kullanılarak yüz değiştirme modelinin öğrenilmesi sürecinde, kişi A ve kişi B’ye ait ön işlenmiş örnekler, aynı kodlayıcı parametreleri kullanılarak ortak bir sıkıştırılmış gizli (latent) uzaya eşlenir.[9]

Üç ağ eğitildikten sonra, kişi B’nin yüzünü kişi A’nın üzerine yerleştirmek için; kişi A’ya ait hedef video (veya görüntü), kare kare ortak kodlayıcı ağa verilir ve ardından kişi B’ye ait çözücü (decoder) ağ tarafından yeniden üretilir.[10]

Yüz Değiştirme

Kullanıcıların yüz değiştirme işlemi yapmasına imkân tanıyan çok sayıda uygulama bulunmaktadır; ancak bu uygulamaların tümü aynı teknolojiyi kullanmamaktadır. Yüz değiştirme yapılmasına olanak sağlayan bazı uygulamalar şunlardır: FaceShifter, FaceSwap, DeepFace Lab, Reface ve TikTok.

Ayrıca Snapchat ve TikTok gibi uygulamalar, hesaplama gücü ve uzmanlık gereksinimini ciddi ölçüde düşürerek, kullanıcıların gerçek zamanlı çeşitli manipülasyonlar üretmesine imkân tanımaktadır.[1]



İnsanlar bir **deepfake yüz değiştirme (face swap)** uygulamasını düşündüklerinde, genellikle 2019 yılında **David Letterman'ın programında** Bill Hader'ın **Tom Cruise, Arnold Schwarzenegger, Al Pacino ve Seth Rogen**'a dönüştüğü videoyu hatırlar. Bu, **zarar vermeyen** bir deepfake kullanımına örnektir. Ancak gördüğümüz üzere, **Yapay Zekâ / Makine Öğrenmesi (AI/ML)** teknolojilerinin karanlık bir yönü de vardır ve bu durum teknolojinin kendisinden değil, **onu kullanan kişilerden** kaynaklanmaktadır.

Dünya, kötü niyetli aktörler tarafından kullanılan deepfake'lerin **doğasında bulunan risklerin** farkında olmak zorundadır.

Yüz değiştirme teknolojisinin en büyük zararlı kullanım alanlarından biri **deepfake pornografidir**. Yüz değiştirme teknolojisi, oyuncular **Kristin Bell** ve **Scarlett Johansson**'ın yüzlerinin birçok pornografik videoya yerleştirilmesi için kullanılmıştır. "Sızdırılmış görüntü" olarak etiketlenen sahte videolardan biri **1,5 milyondan fazla izlenme** elde etmiştir. Kadınların, kötü niyetli bir aktörün deepfake pornografisi üretmesini **önleyebilecek hiçbir etkili yolu yoktur**.

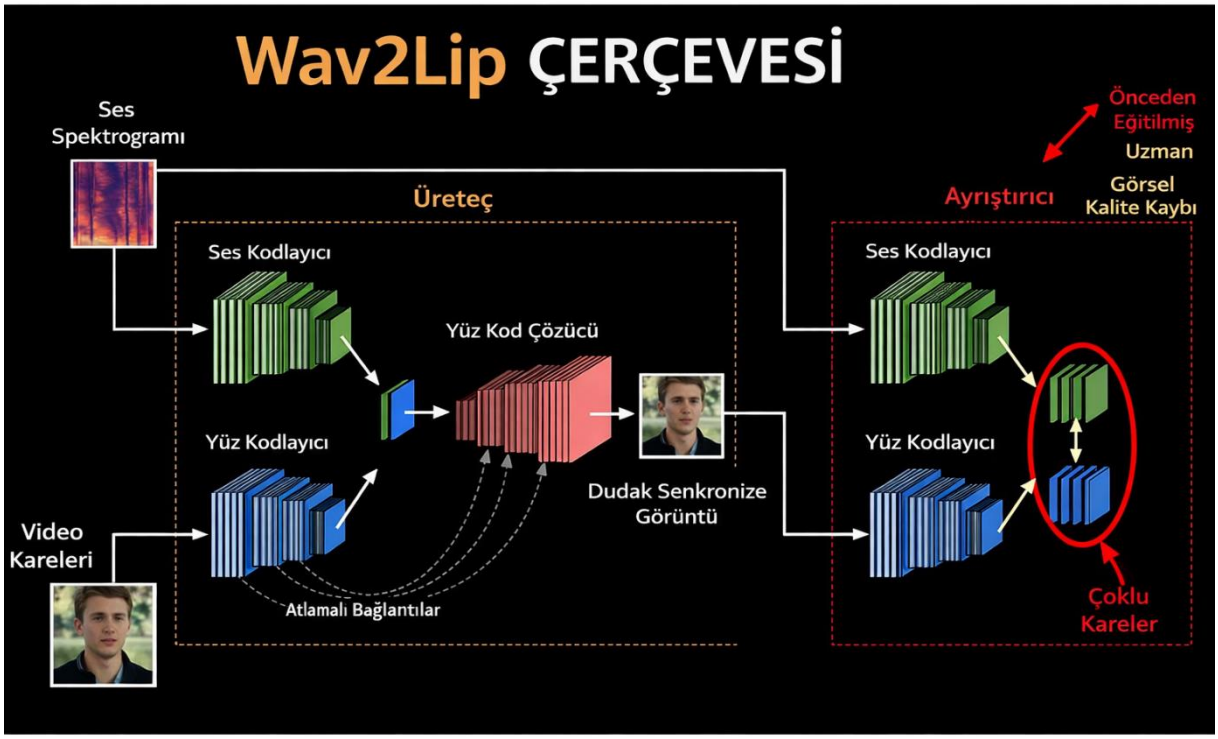
Kamuoyunda tanınmayan, kamu gücüne veya yanlış bilgileri çürütecek kaynaklara sahip olmayan **özel kişilerin taciz edilmesi veya zarar görmesi amacıyla** bu teknolojinin kullanılması son derece endişe vericidir. Deepfake pornografinin **sonuçları ve etkileri** henüz yeni yeni ortaya çıkmaya başlamıştır.

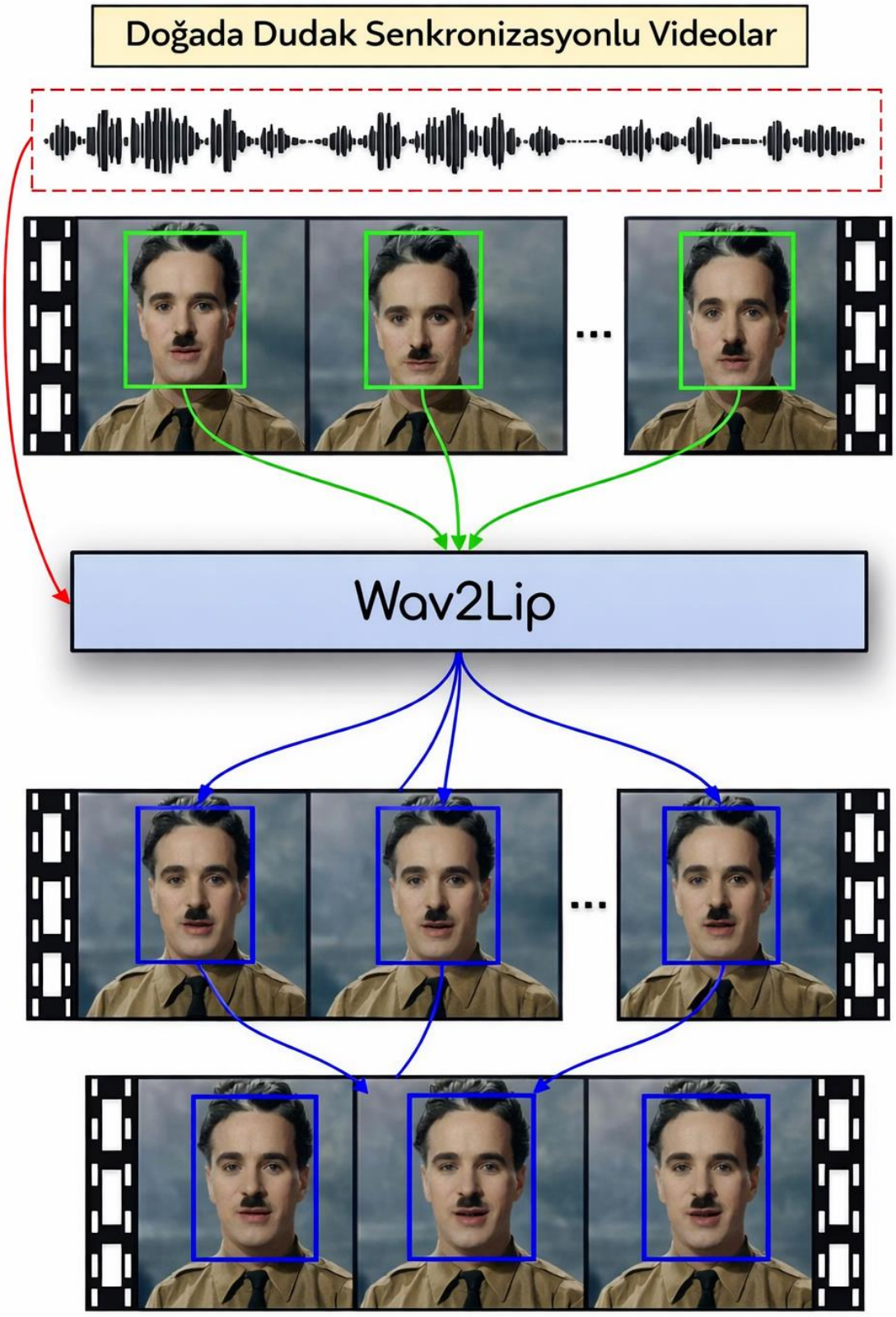
Bir diğer deepfake tekniği ise **“Dudak Senkronizasyonu (Lip Syncing)”**dur. Dudak senkronizasyonu, “bir ya da birden fazla bağlamdan alınan bir ses kaydının, başka bir bağlamdaki video kaydına eşleştirilerek videodaki kişinin gerçekten o sözleri söylüyormuş gibi görünmesini sağlama” işlemidir. Dudak senkronizasyonu teknolojisi, **tekrarlayan sinir ağları (RNN)** kullanılarak hedef kişinin **istenilen her şeyi söylemiş gibi gösterilmesine** imkân tanır.

Kasım 2017’de Stanford Üniversitesi araştırmacıları, **Face2Face** adlı, RNN tabanlı bir video üretim modeli üzerine bir makale ve model yayımlamıştır. Bu model, üçüncü kişilere **kamuoyunda tanınan kişilerin ağzından gerçek zamanlı olarak sözler söyletme** imkânı tanımaktadır.

O tarihten bu yana deepfake teknolojisi **hızla gelişmiş** ve **genel kullanıcılar için daha erişilebilir** hale gelmiştir. Bu teknikler yaygınlaştıkça, özellikle **politik, sosyal veya ekonomik olarak kırılgan** durumda olan özel kişilere yönelik zarar riski artacaktır.

Son olarak, **Wav2Lip** adlı yeni bir AI/ML teknolojisi, **dudak senkronizasyonu deepfake’lerinin** üretilmesini mümkün kılmıştır.




Doğada Dudak Senkronizasyonlu Videolar
Wav2Lip

Uygulamalar

Seslendirme filmlerinde dudak senkronizasyon

Çeviri derslerde dudak senkronizasyonu

CGI Karakterlerde dudak senkor-nizasyonu

Basın toplantilarinda canli çeviri

Eksik görüntülü çağri segmentleri

Mizah içerikleri, sticker ve yüksek doğruluk

Wav2Lip'in Özellikleri

- Herhangi bir yüz için çalişir
- Herhangi bir ses için çalişir
- İlk dilde çalişir
- İlk dil video kareleriyle kusursuz şekilde harmanlama
- Doğadaki videolar için yüksek doğruluk
- CGI yüzler & sentetik sesler için çalişir

Wav2Lip, son derece gerçekçi ses-görüntü eşleşmesi üretir ve gerçek senkronize videolarla eşdeğer dudak senkronizasyon doğruluğuna sahip videolar üretebilen ilk konuşmacıdan bağımsız modeldir. Yapılan insan değerlendirmeleri, Wav2Lip tarafından üretilen videoların, mevcut yöntemlere ve senkronize edilmemiş sürümlere kıyasla vakaların %90'ından fazlasında tercih edildiğini göstermektedir.

Zaman çizelgesinde temsil edilen son bir deepfake tekniği ise "kukla (puppet)" tekniği olarak adlandırılmaktadır. Adından da anlaşılacağı üzere, kukla tekniği kullanıcının hedeflenen kişiyi gerçekte yapmadığı hareketleri yapıyormuş gibi göstermesine olanak tanır. Bu, yüz hareketlerini veya tüm vücut hareketlerini içerebilir. Kukla deepfake'leri, bilgisayar tabanlı grafiklerden oluşan Üretici Karşıt Ağ (Generative Adversarial Network – GAN) teknolojisini kullanır. GAN'ler hakkında daha fazla bilgi için aşağıdaki metin kutusuna bakınız.

Üretici Karşıt Ağlar (Generative Adversarial Networks – GAN'ler)

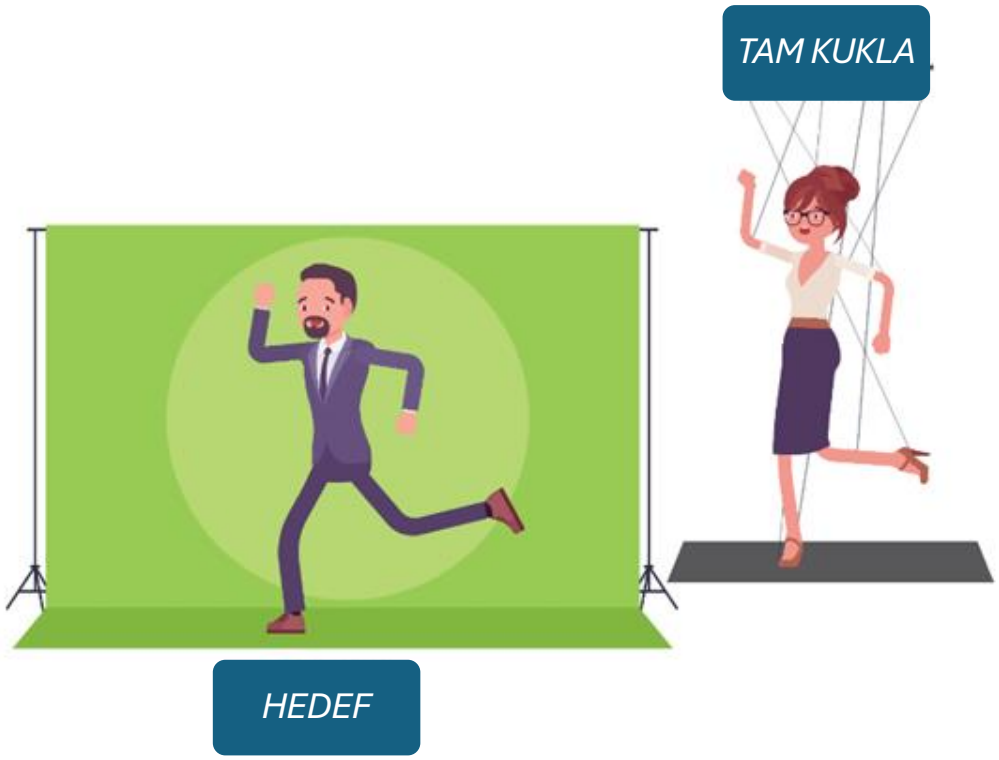
Deepfake ve diğer **sentetik medya** türlerinin üretilmesinde kullanılan temel teknolojilerden biri, **"Üretici Karşıt Ağ (Generative Adversarial Network – GAN)"** kavramıdır. Bir GAN yapısında, **karşıt (adversarial) bir süreç** yoluyla sentetik içerik üretmek üzere **iki ayrı makine öğrenmesi ağı** birlikte kullanılır.

Bu ağlardan ilki **"üretici (generator)"** olarak adlandırılır. Oluşturulmak istenen içerik türünü temsil eden veriler bu ilk ağa verilir; böylece ağ, bu veri türünün **özelliklerini öğrenir**. Üretici ağ daha sonra, orijinal verinin özelliklerini taşıyan **yeni veri örnekleri** oluşturmaya çalışır.

Oluşturulan bu örnekler, ikinci makine öğrenmesi ağına sunulur. Bu ikinci ağ da (biraz farklı bir yaklaşımla eğitilmiş olmakla birlikte) aynı veri türünün **ayırt edici özelliklerini tanımayı öğrenmiştir**. **"Karşıt (adversary)"** olarak adlandırılan bu ikinci ağ, sunulan örneklerdeki **kusurları tespit etmeye** çalışır ve orijinal veriye ait özellikleri yeterince taşımadığını düşündüğü örnekleri **reddederek "sahte" (fake)** olarak tanımlar.

Bu sahte olarak işaretlenen örnekler tekrar üretici ağa geri gönderilir; böylece üretici ağ, yeni veri üretme sürecini **geliştirmeyi öğrenir**. Bu karşılıklı süreç, üretici ağın oluşturduğu sahte içeriğin, karşıt ağ tarafından **gerçek olarak kabul edilmesine** kadar devam eder.

GAN'lerin ilk pratik uygulaması, **Ian Goodfellow** ve çalışma arkadaşları tarafından **2014 yılında** ortaya konulmuş; bu çalışmada **insan yüzlerine ait sentetik görüntüler** üretilebildiği gösterilmiştir. İnsan yüzleri GAN uygulamalarında popüler bir konu olsa da, bu teknoloji **her türlü içerik türüne** uygulanabilir. GAN'lerde ağların eğitilmesinde kullanılan içerik ne kadar **detaylı ve gerçekçi** olursa, elde edilen çıktı da o kadar **gerçekçi** olacaktır.



Öğrenme temelli yöntemlerin aksine, bazı yöntemler **deepfake** üretmek için daha **geleneksel bilgisayar grafikleri** yaklaşımlarına dayanır. Örneğin **Face2Face**, bir kişinin (**"efendi / master"**) yüz ifadeleri ve baş hareketlerinin, başka bir kişinin (**"kukla / puppet"**) yüzüne aktarılmasına imkân tanıyan ve **"kukla yöneticisi (puppet master) deepfake"** olarak adlandırılan içeriklerin oluşturulmasını sağlar.

Cheapfake'ler ile Deepfake'lerin İlişkisi ve

Dijital Manipülasyonda Güncel Durum

Toplum, **sentetik medyayı** yüzyıllardır manipüle edebilmektedir. **Görsel-işitsel cheapfake** örnekleri, dijital çağdan bile **önceye** uzanmaktadır.

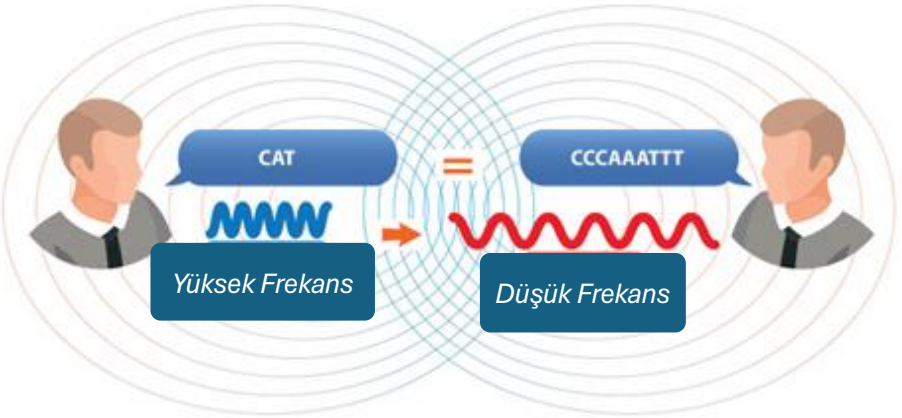
1983 Birleşik Krallık seçimleri öncesinde, İngiliz anarko-punk grubu **Crass** üyeleri; **Margaret Thatcher** ve **Ronald Reagan**'ın konuşmalarından alınan bölümleri birleştirerek, iki lider arasında geçtiği izlenimi veren **sahte bir telefon konuşması** oluşturmuştur. Bu kurguda her iki lider de **kışkırtıcı ve politik açıdan zarar verici** ifadeler kullanıyormuş gibi gösterilmiştir.

Yapay zekâ / makine öğrenmesi (AI/ML) ortaya çıkmadan önce de **Adobe© Photoshop™**, videoların **yavaşlatılması veya hızlandırılması** ve **benzer yüzlerin (look-alike)** kullanımı gibi yöntemler mevcuttu. Günümüzde bu teknikler **"cheapfake"** olarak

adlandırılmakta ve **“yüzeysel sahtekârlıklar (shallow fakes)”** olarak da bilinmektedir. Cheapfake’ler; deepfake’lere kıyasla **daha ucuz**, **daha erişilebilir yazılımlar** (hatta bazen hiçbir yazılım kullanmadan) üretilen **görsel-işitsel (AV) manipülasyonlardır**.

Bu teknikler **daha düşük maliyetlidir**, **daha az teknik uzmanlık gerektirir** ve **çok daha geniş bir kitle tarafından kullanılabilir** durumdadır. Yakın dönemin en bilinen cheapfake örneklerinden biri, **ABD Temsilciler Meclisi Başkanı Nancy Pelosi**’yi konu alan bir videodur. Bu videoda Pelosi’nin konuşması bilinçli olarak **yavaşlatılmış**, böylece **sarhoşmuş gibi** görünmesi sağlanmıştır.

Deepfake teknolojisi ilk geliştirildiğinde, bu tür içeriklerin üretimi **yüksek düzeyde yapay zekâ bilgisi**, eğitim, ileri teknoloji, gelişmiş ekipman, büyük miktarda **eğitim verisi** ve **zaman** gibi ciddi kaynaklar gerektiriyordu. Ancak **son yıllardaki gelişmeler**, deepfake teknolojisini **çok daha az kaynak gerektiren** bir hale getirmiş ve böylece **genel nüfus için çok daha erişilebilir** kılmıştır.

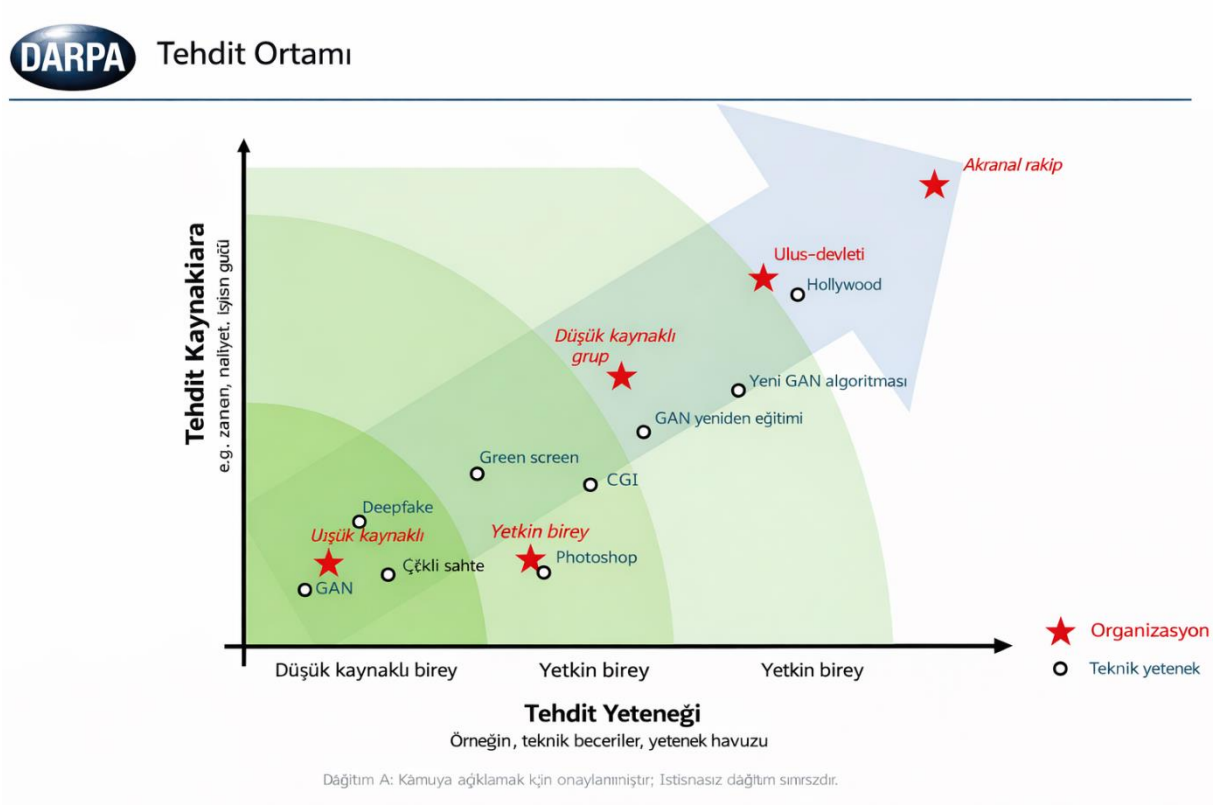


Ses / Video Hızı Düşürme

2015 yılında, **Savunma İleri Araştırma Projeleri Ajansı (Defense Advanced Research Projects Agency – DARPA)**, **Medya Adli İncelemesi** ya da **“MediFor”** olarak adlandırılan programı başlatmıştır. **MediFor** programının amacı; **görsel medyada (görüntü ve videolar)** yapılan manipülasyonları **otomatik olarak tespit etmek**, tespit edilen manipülasyonlar hakkında **ayrıntılı bilgi sunmak** ve şüpheli bir görüntü ya da videonun **özgünlüğünün doğrulanmasına yardımcı olmak** amacıyla medyanın **genel bütünlüğü** hakkında değerlendirme yapmaktır.

MediFor kapsamındaki gelişmeleri, **sentetik medya** ve **içerik manipülasyon teknikleri** bağlamında daha geniş bir çerçeveye oturtmak amacıyla, program; sentetik medya ve görsel manipülasyon teknolojilerinde **mevcut en ileri düzeyi (state of the art)** haritalandırmıştır. Aşağıdaki grafik, farklı teknolojileri; bu tekniklerin kullanılabilmesi için

gereken **beceri düzeyi** (**yatay eksen**) ile **kaynak düzeyi** (**dikey eksen** – yani zaman, maliyet ve hesaplama gücü) açısından konumlandırmak üzere tasarlanmıştır.



Yukarıdaki metin kutusunda da belirtildiği üzere, **Üretici Karşıt Ağlar (Generative Adversarial Networks – GAN'ler)**, **gerçekçi sentetik içerik üretiminde** önemli bir evrimi temsil etmektedir. Tek bir GAN türü bulunmadığından, yukarıdaki grafik birden fazla GAN örneğini ve **yenilerinin ortaya çıkma olasılığını** da göstermektedir. Grafik ayrıca, günümüzde **deepfake** üretiminin **düşük düzeyde beceri ve kaynak** gerektirdiğini ortaya koymaktadır.

Her türlü **sentetik medya manipülasyonu**, zarar vermek amacıyla **silah haline getirilebilir**. Ancak deepfake teknolojisinin **hızlı gelişimi**, bu tehdidi daha da artırmaktadır. Deepfake'ler, **cheapfake**'lere kıyasla **daha gerçekçi** olup **tespit edilmeleri daha zordur**.

Rochester Institute of Technology bünyesindeki **Global Cybersecurity Institute**'ın Araştırma Direktörü ve **Bilişim Güvenliği Profesörü Dr. Matthew Wright**, **AEP Ekibi**'ne yaptığı açıklamada şunları ifade etmiştir:

"Cheapfake'lerin daha kolay yapılması gerekir; çünkü onları çok daha fazla görüyoruz."

Bununla birlikte, **asıl yükselen ve daha ağır tehdit**, deepfake'lerdir. Bu durum, **bir hesap makinesi ile bir bilgisayar arasındaki farka** benzetilebilir: Bilgisayar **katlanarak daha güçlüdür**, çok daha fazla şeyi yapabilir; yapılabileceklerin kapsamı açısından daha etkilidir; **daha görsel olarak ikna edici**, **daha doğru** ve **daha yüksek çözünürlüklüdür**. Cheapfake'lerle yapılabilen her şey, **deepfake'lerle çok daha inandırıcı biçimde** gerçekleştirilebilir.

TEHDİT ORTAMI

Deepfake'ler ve **sentetik içeriğin kötüye kullanımı**, **ulusal güvenlik**, **kolluk kuvvetleri**, **finansal sistemler** ve **toplumsal alanlar** dâhil olmak üzere kamuoyuna yönelik **açık, mevcut ve giderek gelişen bir tehdit** oluşturmaktadır.

Deepfake'lerin yarattığı mevcut tehdit ortamını doğru biçimde tanımlayabilmek için; **çeşitli tehdit türleri**, **kötü niyetli aktörler**, **mağdurlar** ve **teknolojik bağlam** birlikte ele alınmalıdır. Sanayi, akademi ve kamu sektöründeki uzmanlar, sentetik medyanın oluşturduğu tehdit konusunda **farklı bakış açılarına** sahiptir. Bazı uzmanlar, bu tehdidin **abartıldığını**, **olduğundan fazla yansıtıldığını**, **ikna edici bir deepfake üretmenin teknik olarak zor** olduğunu ya da **şüpheci bir kamuoyu karşısında etkisiz kalacağını** savunmaktadır.

Ancak, **rızaya dayanmayan kötü niyetli sentetik pornografi saldırılarına** maruz kalmış bir kişiye sorulduğunda, tehdidin **son derece gerçek ve ciddi derecede zararlı** olduğu açıkça ifade edilecektir.

Gerçek şu ki, **deepfake saldırıları halihazırda mevcuttur** ve özellikle **rızaya dayanmayan pornografi** ile **sosyal medya platformlarındaki sınırlı etki/algı operasyonları** gibi alanlarda **yaygınlaşmakta** olduğu görülmektedir. **2019 yılından bu yana**, **Rusya ve Çin dâhil olmak üzere bazı ulus-devletlerle ilişkili kötü niyetli aktörler**, sosyal medya profillerinde **GAN ile üretilmiş görüntülerden yararlanarak etki operasyonları** yürütmüştür. Bu aktörler, **sentetik kimlikler (personalar)** kullanarak güvenilirlik ve inandırıcılık oluşturmuş; böylece **yerel veya bölgesel meseleleri** öne çıkarmayı amaçlamıştır.

Bu durum **tekil bir olay değildir** ve günümüzde **etki kampanyaları çağında yaygın bir teknik** haline gelmiş görünmektedir. **Facebook** gibi sosyal medya platformları ile **Graphika** gibi diğer **AI/ML araştırma şirketleri**, bu profilleri tespit edebilmiş ve kullanılan görüntülerin **yapay zekâ temelli ve sentetik olarak üretildiğini** değerlendirmiştir. Bu, sentetik kimliklerin tespit edilmesi açısından **önemli bir başarı** olmakla birlikte, tespitler **her zaman zamanında yapılamamıştır** ve **GAN ile üretilmiş görüntüler kullanan kaç sosyal medya profilinin hâlâ tespit edilmediğini söylemek neredeyse imkânsızdır**.

Aşağıda, **etki kampanyalarının bir parçası olarak kullanılan sentetik içeriklere** ilişkin bazı **spesifik örnekler** sunulmaktadır:

- 2020–2021 yılları arasında, GAN'ler ile üretilmiş görüntüler kullanan sosyal medya kimlikleri, Belçika'nın 5G kısıtlamalarına yönelik tutumunu eleştirmiştir. Bu faaliyetlerin, 5G altyapısı satmaya çalışan Çinli firmaları destekleme amacı taşıdığı değerlendirilmektedir.
- 2021 yılında, FireEye, siber aktörlerin sosyal medya platformlarında GAN ile üretilmiş görüntüler kullanarak Lübnan'daki siyasi partileri destekleyici içerikler yaydığını rapor etmiştir.
- Ulus-devletlerle bağlantılı siber aktörler tarafından yürütülen çok sayıda etki operasyonunda, yerel ve bölgesel meseleleri hedef almak üzere GAN ile üretilmiş görüntüler kullanılmıştır.

Rızaya dayanmayan pornografi, deepfake içeriklerin yaygınlaşmasında **katalizör** olarak ortaya çıkmış olup, günümüzde hâlâ **doğada (gerçek dünyada) karşılaşılan yapay zekâ destekli sentetik içeriklerin büyük çoğunluğunu** oluşturmaktadır.

- **Ekim 2020'de**, **Sensity AI**'ye (deepfake içerik ve tespiti konusunda uzman bir firma) göre araştırmacılar, **kadınlara ait 100.000'den fazla bilgisayar üretimi sahte çıplak görüntünün**, **ilgili kişilerin bilgisi ve rızası olmaksızın** oluşturulduğunu rapor etmiştir. Bu görüntülerden bazılarının **reşit olmayan bireyleri** tasvir ettiği de anlaşılmaktadır. İçerik üreticileri, deepfake içeriklerle ilişkili **paylaşım, takas ve satış hizmetlerini** kolaylaştırmak amacıyla **Telegram** mesajlaşma platformu üzerinde **botlardan oluşan bir ekosistem** kullanmıştır.
    - **Sensity AI**, **2018'den bu yana üretilen deepfake videoların yaklaşık %90–95'inin**, ağırlıklı olarak **rızaya dayanmayan pornografiye** dayandığını tespit etmiştir.
- **2021 yılında**, sentetik olarak üretilmiş metin kullanan **AI Dungeon** adlı platformun, **çocukların cinsel istismarını tasvir eden metinler** üretebildiği ortaya çıkmıştır. Bu durum, kullanıcı girdilerine ve çok büyük eğitim veri kümelerine dayanmakla birlikte, **yeterli kısıtlamalar olmadan geliştirilen AI/ML tabanlı içerik üretiminin istenmeyen sonuçlarını** açıkça göstermiştir. **AI Dungeon**, **OpenAI'nin GPT-3** adlı **otomatik artımlı (auto-regressive) dil modelini** kullanmaktaydı. Bu model; **doğal dil üretimi**, **metinden görüntü üretimi**, **çeviri** ve diğer **metin tabanlı uygulamalar** dâhil olmak üzere birçok farklı yetenek kapsamında uygulanabilmektedir.
- **Aralık 2019'da**, deepfake pornografi ve arkasındaki teknolojiyi araştıran bir **gazeteci**, bir **yüz değiştirme (face-swap) pazar yerine** katılmış ve sentetik içerik üreticisinden, **yüzünün pornografik bir videoya dijital olarak yerleştirilmesi** için ödeme yapmıştır. İçerik oluşturmak amacıyla çok sayıda fotoğraf göndermesi gerekmediğini; bunun yerine **450 ayrı kareden oluşan 15 saniyelik bir Instagram**

hikâyesinin, gerekli tüm kareleri sağladığını tespit etmiştir. Gazeteci, önden bakan bir kamerayla konuştuğu **13 saniyelik bir video** ile yüzünün yerleştirilmesini istediği **Pornhub videosunun bağlantısını** içerik üreticisine göndermiştir.

Tehdit Ortamını Etkileyen İlave Faktörler

Tehdit ortamını etkileyen çeşitli faktörler bulunmaktadır. Bunlar arasında; AI/ML ile üretilen sentetik içeriğin giderek artan yetenekleri, hukuki çerçeveler, ulus-devletler tarafından üzerinde uzlaşılan normlar ve eşikler, sentetik içeriğin dolandırıcılık faaliyetlerinde kullanılabilme imkânı ve kamuoyunun gördüğüne inanma eğilimi yer almaktadır. Kötü niyetli aktörlerin tamamı bu faktörlerin her birinden aynı ölçüde etkilenmez. Örneğin, finansal dolandırıcılık faaliyetlerine karışan siber suç aktörleri, ulus-devletlerin sentetik içeriğin kullanımına ilişkin belirlediği normları önemsemeyecektir.

Ulus-devletler, yüksek etkili bir deepfake saldırısını bir tırmanma (eskalasyon) olarak değerlendirebilir. Bu nedenle, en yetenekli aktörler arasında yer almalarına rağmen, deepfake saldırısı gerçekleştirme olasılıkları görece düşük olabilir. Buna karşılık, siber suçlular ve diğer bireysel aktörler, sentetik medya üretme konusunda daha az caydırılacaktır.

Bu bağlamda, “yüksek etkili–düşük olasılıklı” saldırıların, “düşük etkili–yüksek olasılıklı” saldırılardan sayıca fazla olması beklenmemektedir. Ancak bu durum, örneğin rızaya dayanmayan pornografinin mağduru olmanın, ilgili birey açısından düşük etkili olduğu anlamına kesinlikle gelmemektedir.

Genel kamuoyu, ilk bir deepfake saldırısını deneyimledikten sonra, sentetik içeriklere karşı çok daha dirençli hale gelebilir. Bu nedenle kötü niyetli aktörler, yüksek etkili bir deepfake saldırısını gerçekleştirmeden önce “büyük vurgun” için uygun zamanı beklemeyi tercih edebilir. Bu aktörler, kamuoyunun sentetik medyaya karşı zamanla daha dayanıklı hale geldikçe fırsat penceresinin daraldığını değerlendirebilir ve bu nedenle büyük etki yaratacak bir saldırıyı gecikmeden gerçekleştirmeye teşvik edilebilir.

DEEPFAKE SENARYO ÖRNEKLERİ

Deepfake’lerin oluşturabileceği potansiyel tehdit, **belirli senaryolar bağlamında** daha iyi anlaşılabilir. Bu bölüm, deepfake’lerin rol oynayabileceği senaryolara ilişkin **örnekler** sunmaktadır. Burada yer alan örnekler çok sayıda olsa da, **kapsamlı veya sınırlayıcı değildir**.

Bu Senaryolar Nasıl Seçildi?

Deepfake ve sentetik medyanın tehdidini değerlendirmek üzere **üç ana kategori** belirlenmiştir:

- **Ulusal güvenlik ve kolluk kuvvetleri**,
- **Ticaret**,
- **Toplum**.

Bu kategorilerin her biri için, **deepfake teknolojisinin gelecekteki durumu**, mevcut ve potansiyel tehdit dinamikleriyle birlikte ele alınarak senaryolar tasarlanmıştır. Aşağıdaki bölüm, her kategori için birkaç senaryoyu incelemekte ve analizi **dört temel soru** etrafında yapılandırmaktadır: **amaçlar**, **kullanım biçimi**, **beklenen kazanımlar** ve **yöntem veya adımlar**.

Bir Deepfake Saldırısı Nasıl Görünebilir?

Ulusal Güvenlik ve Kolluk Kuvvetleri – Senaryo 1: Şiddeti Kışkırtma

Deepfake bir videonun **toplumsal huzursuzluk ve şiddeti körükleyebileceği** bir senaryo, **Profesör Danielle Citron** tarafından **2019 yılında** ABD Temsilciler Meclisi **Daimî İstihbarat Seçim Komitesi**'ne verilen ifadede dile getirilmiştir. Bu senaryoda, **kötü niyetli bir aktör**, **Baltimore Polis Departmanı (BPD) Komiseri'nin**, **2015 yılında polis nezaretindeyken hayatını kaybeden Afro-Amerikalı Freddie Gray'e kötü muameleyi desteklediğini** gösteren bir deepfake video üretmektedir.

Profesör Citron tarafından tanımlanan senaryo şu şekilde gelişebilir: Kötü niyetli aktör, deepfake yayımlayarak **şiddeti kışkırtmaya karar verdikten sonra**, **Baltimore Polis Departmanı** hakkında araştırma yapar ve **Komiser'e ait fotoğraflar, videolar ve ses kayıtları** gibi **eğitim verilerini** toplar. Bu materyaller, geçmiş **basın toplantılarından** ve/veya **haber kayıtlarından** elde edilebilir.

Bir sonraki aşamada, kötü niyetli aktör topladığı bu bilgileri kullanarak, **Komiser'in yüzünü ve sesini taklit edebilen bir AI/ML modeli** eğitir. Model eğitildikten sonra, aktör **polis komiserinin Freddie Gray'e yönelik kötü muameleyi onayladığını gösteren bir deepfake video** oluşturur. Bu video, açıklamalara **inandırıcılık kazandırmak amacıyla özel bir konuşma gibi kurgulanır** ve **kışkırtıcı ifadeler** de içerebilir.

Son aşamada, kötü niyetli aktör videoyu **anonim olarak sosyal medya platformlarında paylaşır** ve **sahte sosyal medya hesapları** kullanarak videoya dikkat çekmeye çalışır.

Ulusal Güvenlik ve Kolluk Kuvvetleri – Senaryo 2:

İklim Değişikliği Hakkında Sahte Delil Üretilmesi

Bu senaryoda, **Çin'e ait uydular**, **Antarktika** ve çevresindeki **buz tabakasının** görüntülerini kaydeder. Uygulayıcılar, **AI/ML modelleri** kullanarak bu görüntülere, **buzul büyümesinin azalmış gibi değil; artmış ya da sabit kalmış gibi görünmesini sağlayan özellikler** ekler. Çin daha sonra bu **sentetik olarak üretilmiş uydu görüntülerini**, **Birleşmiş Milletler** ve diğer uluslararası aktörleri, **Çin'in ekonomik kalkınması**

açısından olumsuz sonuçlar doğurabilecek daha sıkı iklim anlaşmalarının uygulanmasını geciktirmeye veya iptal etmeye ikna etmek amacıyla kullanır.

Bu girişim başarısız olsa bile, Çin bu **sahte verileri**, söz konusu anlaşmaların gerekliliğini **inandırıcı biçimde tartışmaya açmak** ve **sera gazı emisyonları gibi çevresel kısıtlamaları** görmezden gelmek için kullanabilir. Her ne kadar **ABD'ye ait ve diğer sivil toplum kuruluşlarına (NGO) ait uydular**, Çin'in Birleşmiş Milletler'e sunduğu verileri **çürütebilecek başka verilere** sahip olsa da, bu durum **gecikmelere**, **kafa karışıklığına** yol açabilir ve **küresel mutabakatları zayıflatabilir**.

Ulusal Güvenlik ve Kolluk Kuvvetleri Senaryosu 3: Deepfake Kaçırma

Bu senaryoda, Meksika'daki turistik bir bölgede faaliyet gösteren bir suç çetesi, sentetik görüntü ve videolar kullanarak bir kişinin esaret altında olduğu izlenimini yaratmak suretiyle hedefli ve fırsatçı dolandırıcılık faaliyetleri yürütmektedir. Kötü niyetli failler gerçekte kimseyi kaçırmaz; bunun yerine, internette buldukları ya da çalınmış bir cihazdan elde ettikleri görüntü ve bilgileri kullanarak bu dolandırıcılık planını uygularlar.

Failler daha sonra hedef kişinin ailesiyle iletişime geçerek fidye talep eder. Mağdurun bir otel odasında, muhtemelen bağlı ve gözleri kapalı şekilde tutulduğunu gösteren "yaşam kanıtı" niteliğinde görüntüler sunarlar. Ayrıca, mağdurun yaralandığına dair izler taşıyan ek görüntüler göndererek aile üzerinde daha fazla baskı kurabilirler.

Bu süreçte mağdurun kendisi gerçekte herhangi bir tehlike altında değildir ve yaşananlardan tamamen habersiz olabilir.

Ulusal Güvenlik ve Kolluk Kuvvetleri Senaryosu 4: Bir Ceza Davasında Sahte Delil Üretilmesi

Bu senaryoda, kendisine ait bir binada işlenen bir cinayetle suçlanan varlıklı bir sanık; gizli (latent) parmak izleri, saçtan elde edilen DNA ve saik (motiv) gibi çeşitli delillere dayanılarak suçlanmaktadır. Sanık, binanın lobisinde bulunan kameralar tarafından kaydedilen video görüntülerindeki yüzün düşük çözünürlüklü ve net olmaması gerekçesiyle bu görüntülere dayalı kimlik doğrulamasının kullanılmasına itiraz etmiştir.

Sanık, savunma (alibi) olarak, suçun işlendiği anda binanın başka bir bölümünde bulunduğunu tartışmasız biçimde gösterdiğini iddia ettiği video görüntülerini mahkemeye sunmaktadır.

Bu vakadaki amaç, biyometrik delilleri zayıflatmak ve sanığın sağlam bir mazereti (alibisi) olduğunu ileri süren çelişkili bir kanıt ortaya koymaktır. Buradaki deepfake içerik, bizzat mahkemeye sunulan video görüntüsünün kendisidir. Alibi oluşturmak için yalnızca zaman damgası (timestamp) değiştirilemez; aynı zamanda videonun gerçekmiş gibi görünmesini sağlayacak özgün koşullar da kurgulanabilir.

Bu deepfake video, dolaylı olarak, bina ve olay yerinin kendine özgü koşulları nedeniyle, normalde güçlü sayılabilecek delillerin (örneğin biyometrik kanıtların) dahi durumsal (dolaylı) delil olarak değerlendirilmesine yol açabilir.

Ticaret Senaryosu 1: Kurumsal Sabotaj

Bu senaryoda, bir şirketin ürünü, pazardaki konumu, yöneticileri, genel marka algısı ve benzeri unsurlar hakkında yanlış bilgi yaymak amacıyla deepfake teknolojisinin kullanımı ele alınmaktadır. Bu yaklaşım; şirketin pazardaki konumunu olumsuz etkilemek, piyasayı manipüle etmek, rekabeti haksız şekilde zayıflatmak, rakip bir şirketin hisse senedi değerini olumsuz yönde etkilemek veya bir şirketin planlanan birleşme ve satın alma (M&A) süreçlerini hedef almak amacıyla tasarlanmıştır.

Ticaret Senaryosu 2: Kurumsal Düzeyde Geliştirilmiş Sosyal Mühendislik Saldırıları

Bu senaryoda, sosyal mühendislik saldırılarının daha inandırıcı şekilde gerçekleştirilmesi amacıyla deepfake teknolojisinin kullanımı ele alınmaktadır.

Öncelikle kötü niyetli bir aktör, şirketin faaliyet alanı, yöneticileri ve çalışanları hakkında araştırma yapar. Şirketin İcra Kurulu Başkanı'nı (CEO) ve Finans Direktörü'nü tespit eder. Ardından şirketin kısa süre önce duyurduğu yeni bir ortak girişim (joint venture) hakkında bilgi toplar. Aktör, CEO'ya ait TED konuşmaları ve çevrim içi videoları kullanarak, CEO'nun sesini taklit eden bir deepfake ses modeli oluşturur.

Kötü niyetli aktör ayrıca Finans Direktörü'nün sosyal medya profillerini inceler. Finans Direktörü'nün, yeni doğan bebeğiyle ilgili bir fotoğraf paylaştığını ve işe geri dönmenin zor olduğuna dair bir mesaj yayımladığını görür. Daha sonra bu kişi, hileli şekilde para elde etmek amacıyla Finans Direktörü'nü telefonla arar. Görüşme sırasında, işe dönüş sürecinin nasıl gittiğini ve bebekle ilgili durumu sorar.

Finans Direktörü telefonu açar ve arayan kişinin patronunun sesini tanır. Kötü niyetli aktör, ortak girişim kapsamında kullanılmak üzere 250.000 ABD doları tutarında bir meblağın bir hesaba havale edilmesini ister. Para transferi gerçekleştirilir ve ardından aktör bu fonları birden fazla farklı hesaba aktarır.

GERÇEK ZAMANLI SES DEĞİŞTİRME
(TELEFON ÜZERİNDEN DOLANDIRICILIK İÇİN)

Ticaret Senaryosu 3: Finansal Kurumlara Yönelik Sosyal Mühendislik Saldırısı

Bu senaryoda, kötü niyetli bir aktör maddi kazanç elde etmek amacıyla bir finansal kuruma saldırmak için deepfake ses teknolojisini kullanmaya karar verir. Öncelikle karanlık ağda (dark web) araştırma yapar ve birden fazla kişiye ait isim, adres, sosyal güvenlik numarası ve banka hesap bilgilerini ele geçirir.

Kötü niyetli aktör, bu kişilerin TikTok ve Instagram gibi sosyal medya profillerini tespit eder. Sosyal medya platformlarında paylaşılan videoları kullanarak modeli eğitir ve hedef kişilere ait deepfake ses kayıtları oluşturur. Ardından finansal kurumun doğrulama politikalarını araştırır ve sesle kimlik doğrulama (voice authentication) sistemi kullandığını belirler.

Daha sonra finansal kurumu arar ve ses doğrulamasını başarıyla geçer. Bir müşteri temsilcisine yönlendirilir ve dark web üzerinden elde ettiği müşteri gizli (özel) bilgilerini kullanır. Kötü niyetli aktör, temsilciye çevrim içi hesabına erişemediğini ve şifresini sıfırlaması gerektiğini söyler. Bunun üzerine kendisine geçici bir şifre verilir ve çevrim içi hesaba erişim sağlanır.

Sonuç olarak kötü niyetli aktör, hedef kişinin finansal hesaplarına erişim elde eder ve bu hesaplardan yurt dışındaki hesaplara para transferleri gerçekleştirir.

Ticaret Senaryosu 4: Kurumsal Sorumluluk (Liability) Endişeleri

Deepfake'lerin yeterince inandırıcı ve yaygın hale gelmesi durumunda, şirketler; bu saldırılardan etkilenen tüketicilerin, ortaya çıkan ihlaller, kimlik hırsızlığı ve benzeri nedenlerle uğradıkları mali zararlar için tazminat ve telafi talep etmeleri sonucunda artan hukuki risklerle karşı karşıya kalabilir. Ayrıca tüketiciler veya müşteriler, bir şirketi dolandırmak amacıyla sahte bir olay (örneğin bir markette kayıp düşme vakası) kurgulayabilirler.

İlk senaryo, doğrudan bir saldırıdan ziyade, önceki deepfake saldırılarının tetiklediği bir sonuç olarak değerlendirilmektedir.

İkinci senaryoda ise, bir şirketi; arızalanan ve yaralanmaya neden olan bir üründen dolayı sorumluymuş gibi göstermek amacıyla tasarlanmış bir deepfake video ele alınmaktadır. Kötü niyetli aktör, araştırma yaparak geçmişte fiziksel yaralanmalara yol açtığı bilinen bir ürünü tespit eder. Önceki olaylardan elde edilen özgül ayrıntılar, deepfake videoya dahil edilmek üzere seçilir. Ardından, hızlı bir uzlaşma olasılığını değerlendirmek amacıyla eyalet düzeyindeki haksız fiil (tort) hukuk düzenlemeleri incelenir.

Kötü niyetli aktör, YouTube ve diğer sosyal medya platformlarında aynı ürünün arızalanarak insanları yaraladığına dair videolar bulur. Yüz değiştirme (face swap) modeli eğitilir ve deepfake video oluşturulur. Aktör videoyu sosyal medyada paylaşır ve yoğun bir destek dalgası ile karşılaşır. Şirketin sosyal medya ekibi bu paylaşımı fark eder ve kötü niyetli aktörle iletişime geçer. Sosyal medyadan gelen baskı, şirketin sorumluluk kabul etmesi yönünde giderek artar.

Son aşamada, yaralanmalar nedeniyle şirkete tazminat talebi iletilir. Şirket, ödeme yapmadan önce videonun bir deepfake olup olmadığını tespit edebilecek midir? Videoyu araştırmak için yeterli zamana sahip midir? Yoksa şirketin marka itibarına verilecek zararı azaltmak adına kötü niyetli aktöre ödeme yapmak daha mı mantıklıdır?

Ticaret Senaryosu 5: Hisse Senedi Manipülasyonu

Bu senaryoda, hisse senedi piyasasını manipüle etmek ve kötü niyetli aktörün yasa dışı kazanç elde etmesini sağlamak amacıyla üretilen bir deepfake ele alınmaktadır.

Kötü niyetli aktör, hisse senedi manipülasyonu yoluyla hızlı bir kâr elde etmek ister. Aktör, ilgili hisse senedini ayrıntılı biçimde araştırır ve düşük fiyattan satın alır. Ardından Reddit ve Stockaholics gibi borsa forumlarında, şirket çalışanı izlenimi veren birden fazla sahte (deepfake) profil oluşturur. Bu profiller, kullanıcıların şirketin çalışanı olduğunu göstermektedir.

Kendini bu çalışanlar gibi tanıtan aktör, yakında yapılacak "büyük" bir duyuruya ilişkin paylaşımlar yapar. Şirketin CEO'sunu tespit eden aktör, çeşitli televizyon ve radyo programlarında yayımlanan röportajlardan yararlanarak CEO'nun konuşma tarzına ilişkin bir model eğitir. Daha sonra, CEO'nun söz konusu "büyük" duyuruyu konuştuğu

izlenimini veren bir ses deepfake'i üretir ve bu kaydı sosyal medyada paylaşır; ayrıca borsa forumlarında ses kaydına yönlendiren bağlantılar yayımlar.

Kötü niyetli aktör, forumları takip ederek yoğun bir hareketlilik artışı olduğunu ve deepfake ses kaydının amacına ulaştığını görür. Hissenin değeri %1000 oranında yükselir ve aktör, hisse senedi düşüşe geçmeden önce kârını realize eder. Bu durum, diğer yatırımcıların maddi kayıplar yaşamasına ve şirketin itibarının zarar görmesine yol açabilir.

Şirket, CEO'ya ait ses kaydının sahte olduğunu belirten bir açıklama yapmak zorunda kalabilir. Yatırımcılar ise uğradıkları zararların telafisi için şirketten sorumluluk üstlenmesini talep edebilir.

Toplum Senaryosu 1: Siber Zorbalık

Bu senaryoda, bir kişiyi itibarını zedeleyecek ya da belirli gruplara, hizmetlere veya haklara erişimini olumsuz etkileyecek bir durumda gösteren; örneğin kişiyi suç teşkil eden bir davranış içinde tasvir eden bir deepfake içeriğin üretilmesi ele alınmaktadır.

Saldırganın amacı, hedef kişinin itibarını sarsmaktır; bu durum aynı zamanda saldırganın tercih ettiği başka bir kişinin statüsünü güçlendiren ikincil bir etki de yaratabilir. Pensilvanya'da yakın zamanda geniş yankı uyandıran bir olayda, bir kadının, sınırlı kontenjanı bulunan bir amigo takımına seçilmek için rekabet eden kızının arkadaşlarının itibarını zedelemeye çalıştığı görülmüştür.

Bu senaryoda, hedef kişiyi suç teşkil eden bir davranışta bulunuyormuş gibi gösteren bir deepfake video üretilir ve hedefin faaliyetleri üzerinde yetki sahibi olan kişilere gönderilir. Video temel alınarak bu yetkililer, hedef kişinin belirli faaliyetlere katılımını kısıtlar veya tamamen sonlandırır.

Siber Zorbalık Senaryosu – Deepfake'lerin Rolü

Siber zorbalık, özellikle sosyal medyanın yoğun kullanımı nedeniyle genç nesiller arasında yaygın bir sorundur. Söylentiler sosyal medya ve çevrim içi platformlar üzerinden kolayca yayılabilmekte; bu söylentiler, doğruymuş izlenimi vermek amacıyla sahte görüntü veya videolarla desteklendiğinde çok daha inandırıcı hâle gelmektedir. Bu durum, itibarların zedelenmesine ve mağdurların kendilerine zarar vermelerine kadar varabilen ciddi psikolojik etkilere yol açabilmektedir.

Mart 2021'de, iddia edilen deepfake'lerin siber zorbalık aracı olarak kullanıldığı bir olay, haklarında dava açılmasının ardından uluslararası basında gündeme gelmiştir. Pensilvanya'da bir annenin, kızının amigo takımındaki takım arkadaşlarına ait görüntü ve videoları manipüle ettiği ileri sürülmüştür. Söz konusu iddia edilen deepfake'ler, amigo takımındaki bazı üyeleri alkol alırken, elektronik sigara (vape) kullanırken ve çıplak poz verirken göstermekteydi; bu tür davranışlar, onların amigo takımından çıkarılmasına neden olabilecek nitelikteydi.

Mağdurlardan birkaçı siber zorbalığa maruz kaldıklarını açıklamış; bir mağdur ise annenin, iddia edilen deepfake'lerin ötesine geçerek intihara teşvik ettiğini ve tacizi bu şekilde daha da ileri taşıdığını belirtmiştir.

Ancak Mayıs 2021 itibarıyla, video kanıtlarının sahte olduğunun ispatlanamaması nedeniyle deepfake suçlamalarından vazgeçilmiştir. Sentetik medya araştırmacıları, videolarda yüz çevresinde manipülasyona işaret eden tipik bozulma (artefakt) izlerinin bulunmadığını; buna karşın bir kişinin dışarı verdiği buhar bulutu gibi taklit edilmesi zor gerçekçi ayrıntılar içerdiğini belirtmiştir. Birden fazla dijital adli bilişim uzmanı, videoların özgün göründüğünü ve büyük olasılıkla gerçek deepfake olmadığını ifade etmiştir.

Bir diğer önemli sorun ise deepfake üretimine yönelik araç ve kaynakların kamuya son derece kolay erişilebilir hâle gelmiş olmasıdır. Ücretsiz ya da düşük maliyetli olarak indirilebilen mobil ve web tabanlı uygulamalar sayesinde deepfake üretmek mümkündür ve bu içerikler siber zorbalık gibi kötü niyetli senaryolarda kullanılabilmektedir. Deepfake'lerin siber zorbalık vakalarında kullanımının zamanla artması ve özellikle teknoloji ile sosyal medyayı yoğun kullanan genç nesiller için daha büyük bir tehdit hâline gelmesi muhtemeldir.

Toplum Senaryosu 2: Deepfake Pornografi

Bu senaryoda, rıza dışı deepfake pornografi ele alınmaktadır. MIT Technology Review Kıdemli Editörü Karen Hao, "en büyük tehdidin kadınlar ve savunmasız gruplar için olduğunu" belirtmiştir. Deepfake'lerin açık ara %95'i, kadınlara yönelik rıza dışı pornografik içeriklerden oluşmaktadır. Bireysel düzey, en yüksek tehdit seviyesini temsil etmektedir. Bu oran, "görüntüsü dijital olarak yakalanmış ve internete yüklenmiş olan herkesi" kapsamaktadır. Dolayısıyla bu durum, ülkedeki—hatta dünyadaki—neredeyse her kadını kapsamakta ve bu nedenle katlanarak büyüyen bir risk oluşturmaktadır.

Günümüzde, sosyal medya profili bulunan herkes sahte içerik üretimi için hedef hâline gelebilmektedir. Örneğin, ayrılmak istediğini ve başka biriyle görüşmek istediğini dile getiren bir kız arkadaşını şantajla kontrol altında tutmak isteyen öfkeli bir erkek arkadaşı düşünelim. Erkek arkadaş (saldırgan), ayrılık gerçekleşirse mağdurun çıplak fotoğraflarını yayımlamakla tehdit ederek kız arkadaşını (mağduru) korkutmak ve ilişkide kalmaya zorlamak ister.

Mağdur ilişkiyi sürdürmeyi reddettiğinde, saldırgan; birlikte oldukları dönemde çektiği fotoğraflardan ve mağdurun sosyal medya hesaplarından elde ettiği yüz fotoğraflarını toplar. Ardından internetten çıplak kadınlara ait çeşitli görüntüler temin eder.

Saldırgan daha sonra "Reflect" adlı ücretsiz bir uygulamayı kullanarak, mağdurun yüzünü internette bulduğu çıplak kadınlardan birinin vücuduna monte eder. Saldırgan acemi olsa bile, bu görüntüyü oluşturması beş dakika kadar kısa bir sürede mümkün olabilir.

Görüntüyü oluşturduktan ve sahte olduğuna işaret edebilecek bağlamsal unsurları (örneğin uygulamaya ait bir filigran ya da mağdurun hiç bulunmadığı bir konumu gösteren bilgiler) kırptıktan sonra, saldırgan fotoğrafı mağdurun ailesine ve arkadaşlarına gönderir. Ayrıca mağduru daha da küçük düşürmek amacıyla fotoğrafı sosyal medyada paylaşır. Mağdur fotoğrafın sahte olduğunu bilse de, bazı kişilerin görüntünün gerçek olduğuna inanması nedeniyle ağır bir aşağılanma ve utanç yaşar.

VAKA ÇALIŞMASI – NOELLE MARTIN

Ekibimiz, rızası dışında deepfake teknolojisinden ciddi şekilde etkilenen Avustralyalı aktivist ve avukat Noelle Martin ile röportaj yaptı.

2016 yılında, 17 yaşındayken, Martin, yüzünün bir selfie'sinin pornografik bir görüntüye yerleştirildiğini ve birkaç porno sitesinde dağıtıldığını keşfetti. Martin, sesini yükselttiği için ikinci kez hedef alındığını ve rızası dışında deepfake pornografik bir video çekildiğini belirtti. Martin, videonun onu susturmak için bir silah olarak kullanıldığına inandığını söyledi.

Martin ayrıca, rızası dışında çekilen deepfake pornografik videonun kendisine e-posta ile gönderildiğini ve çeşitli sitelerde yayınlandığını belirtti. "... Bir kadının bir erkeğin üstünde seks yaptığı bir videoydu, kadının çıplak vücudu tamamen görünüyordu, vücudu hareket ederken gözleri doğrudan kameraya bakıyordu ve yüzü bu eyleme tepki veriyordu. Ancak bu yüz bir yabancının yüzü değildi... benim yüzümdü." Martin, deepfake pornografinin bir kişinin itibarı, haysiyeti, istihdam edilebilirliği ve kişilerarası ilişkileri üzerinde ömür boyu süren etkileri olduğunu belirtiyor. Röportaj sırasında , kendi çalışma alanında iş bulamadığını belirtti. Bayan Martin, iş aramada karşılaştığı zorluğun, kendisiyle ilgili deepfake pornografik görüntü ve videolardan kaynaklandığına inanıyor.

Bayan Martin, kendisine yönelik deepfake saldırılarından sorumlu olan kişileri tespit etmeye çalışırken karşılaştığı bitmek bilmeyen mücadelenin üzerine düşündü. Beş yıl sonra Bayan Martin , kendisine saldıran kötü niyetli kişiyi veya kişileri hala tanımadığını belirtti. Videolar hala sitelerde ortaya çıkıyor ve videoları kalıcı olarak kaldırmak çok zor. Martin, deepfake pornografiyle ilgili hiçbir yasa olmadığını belirtti. Mevcut birçok yasal talep, çok özel durumlarda yeterli telafi sağlayabilir, ancak hiçbiri deepfake pornografiyi genel olarak ele almaya yetmiyor ve yeni çözümlerin gerekliliğini ortaya koyuyor. Martin, tüm paydaşların bu tehdidi ele almak için bir araya gelmesi gerektiğini, kolluk kuvvetleri arasında küresel bir işbirliği olması gerektiğini ve mağdurların daha fazla ruh sağlığı desteğine ihtiyacı olduğunu belirtti.

Toplum Senaryosu 3: Seçimlere Etki

Bu senaryoda, bir seçim döneminde yanlış bilgi yaymak amacıyla kullanılan bir deepfake ele alınmaktadır. Seçim sürecine girilirken, A Adayı'nı destekleyen teknoloji konusunda yetkin bir grup, B Adayı'na karşı bir dezenformasyon kampanyası başlatabilir. Bu

senaryoda kötü niyetli aktörler, hedeflerine ulaşmak için ses, video ve metin tabanlı deepfake'lerden yararlanabilir.

Tek tek ele alındığında, ses ve video deepfake'leri genellikle daha sansasyonel başlıklar üretir ve insanların dikkatini daha kolay çeker. Buna karşılık metin tabanlı deepfake'lerin oluşturduğu tehdit, çoğu zaman alarm vermeden bilgi ortamına sızabilme ve yayılabilme kabiliyetlerinden kaynaklanmaktadır.

Metin deepfake'lerinin bir diğer önemli kullanım alanı ise sosyal medya platformlarında anlatının (narrative) kontrol altına alınmasıdır. Bu yaklaşım; toplumsal gerilimleri artırabilir, rakip bir adayın itibarına zarar verebilir, belirli bir siyasi tabanı kışkırtabilir ya da seçim sürecine olan güveni zayıflatabilir.

Toplum Senaryosu 4: Çocuk İstismarcısı Tehdidi Senaryosu

Deepfake'ler ve sentetik medya tarafından mümkün kılınan ek bir tehdit senaryosu, teknolojiyi kullanarak çevrim içi konuşmalarda çocukları hedef almak amacıyla gerçekte olduğundan çok daha genç görünen bir avatar oluşturan bir çocuk istismarcısını içermektedir. Bu tür bir saldırganın, belirli ve gerçek bir kişiyi birebir taklit etmesi zorunlu değildir; gerçekçi görünen bir çocuk yüzü ve sesi oluşturması yeterlidir.

Bir sohbet odasında ya da başka bir sosyal medya ortamında karşılaşıldığında, potansiyel mağdur olan çocuk, karşısındaki kişinin bir deepfake olduğunu fark edemeyebilir.

GERÇEK ZAMANLI SES VE VİDEO DEĞİŞTİRME
(ÇEVRİM İÇİ SOHBETLER İÇİN)

Senaryolar – Özet

Görsel-işitsel manipülasyon araçlarının sınırsız kullanımının, kadınlar, beyaz olmayan topluluklar ve güçlü sistemleri sorgulayan kişiler üzerinde sıklıkla olumsuz etkiler yarattığını şimdiden görmekteyiz. Deepfake teknolojisinin çok sayıda kişi tarafından geniş ölçekte erişilebilir hâle gelmesi önemli bir zorluk oluşturmaktadır.

Daha önce yalnızca uzmanların erişebildiği her bir teknik yaklaşım, teknolojik gelişmeler ve sosyal medyanın yaygın kullanımı sayesinde artık amatörler için de ulaşılabilir hâle gelmiş; üretilen içerikler çok daha geniş kitlelere, çok daha hızlı biçimde ulaşabilmektedir. Videoları üreten kişilerin niyetinden bağımsız olarak, teknolojinin ortaya çıkardığı etkiler belirleyici olmaktadır.

Yapay zekâ tarafından üretilen pornografik videolar, bir teknolojinin "demokratikleşmesi" ile bağlantılı, birbiriyle ilişkili sorunlar bütününü açıkça ortaya koymaktadır.

Zarar Azaltmaya İlişkin Değerlendirmeler
Senaryolar Arasındaki Ortak Noktalar

Deepfake kullanılarak gerçekleştirilebilecek saldırıların çok çeşitli olmasına rağmen, bu çalışmada sunulan senaryolar; ilgili paydaşlarla etkileşime geçmek ve zarar azaltma önlemlerini şekillendirmeye yardımcı olmak için kullanılabilecek, genellenebilir bir ilerleme düzenini takip etmektedir. Bu adımlar, Jon Bateman tarafından kaleme alınan bir çalışmada belirtilenlerle de benzerlik göstermekte olup şunları içermektedir:

1. **Niyet (Intent):** Deepfake kullanan her saldırı, kötü niyetli bir aktörün belirli bir hedefe karşı saldırı gerçekleştirmeye karar vermesiyle başlar.

2. **Hedefin Araştırılması:** Bu aşamada kötü niyetli aktör, hedefe ait durağan görüntüler, video ve/veya ses kayıtlarını toplamak amacıyla araştırma yapar. Bu içerikler; arama motorları, sosyal medya platformları, video paylaşım siteleri, podcast'ler, haber siteleri gibi çok çeşitli kaynaklardan elde edilebilir.
3. **Deepfake Oluşturma, Bölüm 1 – Modelin Eğitilmesi:** Kötü niyetli aktör, topladığı bilgileri kullanarak hedefin görünümünü ve/veya sesini taklit eden bir yapay zekâ / makine öğrenmesi (AI/ML) modeli eğitir. Aktörün sahip olduğu kaynaklara ve teknik yeterliliğe bağlı olarak bu işlem, özel olarak geliştirilen modellerle ya da ticari olarak sunulan uygulamalarla gerçekleştirilebilir.
4. **Deepfake Oluşturma, Bölüm 2 – Medyanın Üretilmesi:** Kötü niyetli aktör, hedefin gerçekte yapmadığı ya da söylemediği bir şeyi yapıyormuş veya söylüyormuş gibi gösteren bir deepfake içerik üretir. Bu işlem, aktörün kendi donanımıyla, ticari bulut altyapılarıyla ya da üçüncü taraf uygulamalar aracılığıyla yapılabilir.
5. **Deepfake'in Yayılması:** Kötü niyetli aktör deepfake içeriği yayımlar. Bu yayılım, belirli bir kişiyi hedef alacak şekilde ya da e-posta gönderimi veya sosyal medya paylaşımı gibi yöntemlerle geniş bir kitleye yönelik olarak gerçekleştirilebilir.
6. **İzleyicilerin Tepkisi:** İzleyiciler deepfake içeriğini görür ve buna tepki verir.
7. **Mağdurun Tepkisi:** Deepfake'in mağduru, çoğu zaman "zarar kontrolü" şeklinde bir karşılık verir. Mağdur aynı zamanda bir "izleyici" olsa da, saldırıdaki özel konumu nedeniyle verdiği tepki diğer izleyicilerden oldukça farklı olabilir.

Bu adımların her birinde zarar azaltmaya yönelik fırsatlar bulunmaktadır.

Günümüzde ve deepfake tehdidinin erken aşamaları olarak nitelendirilebilecek mevcut durumda, zarar azaltma stratejileri ağırlıklı olarak teknolojik çözümlerin geliştirilmesine, özellikle de otomatik deepfake tespitine odaklanmaktadır. Ancak deepfake'ler geliştikçe, daha yaygın ve her yerde bulunur hâle geldikçe, bu tek yönlü yaklaşım artık yeterli olmayacak; bireyleri ve kurumları sürekli savunma pozisyonunda, en son tehdide yetişmeye çalışan bir mücadele içinde bırakacaktır. Bu tür tepkisel bir yaklaşım hem verimsiz hem de gereksiz yere risklidir.

Proaktif bir yaklaşım benimsenmesi hâlinde, kapsamlı bir zarar azaltma önlemleri bütünü; deepfake'in ilerleme sürecindeki farklı adımları ele almalı ve ideal olarak aşağıdaki unsurları içermelidir:

Niyet Aşamasına Yönelik Zarar Azaltma Önlemleri:

Kötü niyetli içeriğin yayılmasını suç sayan yasa, politika ve düzenlemeler;

Kötü niyetli içerik paylaşıldığında uygulanabilecek cezai ve hukuki yaptırımların kamuoyuna duyurulması, kötüye kullanım ihtimalini belirli ölçüde azaltabilir;

Kötü niyetli içerik ürettiği tespit edilen bireyler üzerinde oluşacak toplumsal baskı, bazı kişileri bu tür içerikler üretmekten caydırabilir (topluluk temelli öz-denetim?).

Politika ve hukuk:

Deepfake'lerin ortaya çıkardığı yeni tehdidi ele almak üzere uyarlanmış uygun politika ve yasaların araştırılması, geliştirilmesi ve uygulanmasının gerekli olduğu açıktır. Daha fazla bilgi için "Medeni Yargılama (Civil Litigation)" başlıklı metin kutusuna bakınız.

Medeni Yargılama (Civil Litigation)

Son dönemde, Amerika Birleşik Devletleri Kongresi'nde ve eyalet meclislerinde deepfake'leri ele alan çeşitli yasa tasarıları gündeme gelmiştir. Virginia eyaleti, intikam pornosu yasasını genişleterek rıza dışı deepfake pornografiyi de suç kapsamına almıştır; Texas, seçimlere müdahale eden deepfake'leri suç sayan yasalar çıkarmıştır; California ise rıza dışı deepfake pornografinin mağdurlarına tazminat davası açma hakkı tanıyan ve kamu görevi adaylarına, seçimden önceki 60 gün içinde seçimle ilgili deepfake üreten veya paylaşan kişi ya da kuruluşlara karşı dava açma imkânı veren iki ayrı yasa kabul etmiştir. Maryland, New York ve Massachusetts eyaletleri de deepfake'lere ilişkin kendi özgün yasal düzenlemelerini değerlendirmektedir.

Bununla birlikte, deepfake sorununu çözmek için eyalet bazlı yasaların en uygun yol olmadığı ileri sürülmektedir; zira her bir yasal düzenleme deepfake'lerin farklı yönlerini hedef alacak ve yalnızca ilgili eyalet sınırları içinde geçerli olacaktır. Ayrıca, deepfake'lere özgü yasaların aceleyle çıkarılmasının mutlaka gerekli olmayabileceği de savunulmaktadır.

Deepfake'lere özgü federal yasalar kabul edilene kadar, kolluk kuvvetleri mevcut ceza yasalarına dayanabilir. Bazı durumlarda mağdurlar; şantaj, taciz, telif hakkı ihlali (örneğin telif koruması altındaki görüntülerin izinsiz kullanılması), kasten duygusal zarar verme, iftira, yanıltıcı reklam ve "false light" (mahremiyetin ihlali türlerinden biri) gibi iddialarla medeni dava açabilirler.

Örneğin bir deepfake şantaj amacıyla kullanılmışsa şantaj suçuna ilişkin yasalar uygulanabilir; deepfake'lerin bir kişiyi taciz etmek için kullanılması hâlinde ise taciz yasaları devreye girebilir. Ayrıca kötü niyetli bir aktör, bir kişi hakkında rahatsız edici ve gerçeğe aykırı bilgileri doğruymuş gibi yayımlarsa (örneğin fotoğraf manipülasyonu ve abartma yoluyla), mağdur "false light" kapsamında mahremiyetin ihlali gerekçesiyle medeni dava açabilir.

Araştırma (ve Yayım Öncesi Diğer) Aşama Zarar Azaltma Önlemleri

Kurumsal planlama:
Deepfake olayları kaçınılmaz olarak ortaya çıktığında bunların etkisini azaltabilmek için etkili iletişim yapılarının ve kanallarının kurulması ve sürdürülmesi kritik öneme sahiptir. Nasıl ki çoğu kurum, olası bir itibar krizinden çok önce böyle bir duruma karşı halkla

ilişkiler (PR) planları hazırlıyorsa, aynı şekilde kendi anlatısını (narrative) izlemek ve kontrol etmek, yanlış bilginin farklı biçimlerinden kaynaklanabilecek potansiyel felaketleri önlemek için de planlara sahip olmalıdır.

Bu tür bir plan kapsamındaki somut adımlar; genel bir bilgi güvenliği politikasının parçası olarak bir dezenformasyonla mücadele politikası geliştirilmesini ve sosyal medya ile diğer mecralarda hem bilgi hem de dezenformasyonun özel olarak izlenmesi ve raporlanmasını içerebilir.

- **Hedef olabilecek kuruluşlar ve bireyler** – siyasi ya da ticari – özellikle yayımlanan ya da herkese açık çoklu ortam (multimedya) içeriklerini izleme ve düzenleme konusunda proaktif davranabilirler. Bu sıkı içerik denetiminin iki temel gerekçesi vardır:

    1. Mevcut bir içeriğin bir deepfake içinde yeniden kullanılması hâlinde, özgün kaynak içeriğin hızlıca tespit edilmesi ve bunun "gerçek" (otantik) medya olarak sunulabilmesi;

    2. Günümüzdeki en iyi deepfake tespit modellerinden bazılarının, bir videodaki konuşmacının yüz hareketlerini ölçerek bunların önceki gerçek konuşma örnekleriyle tutarlı olup olmadığını belirleyebilmesidir. Ağız değiştirme veya "kukla" (puppet) teknikleri kullanılarak bir deepfake oluşturulduğunda, değiştirilmiş videodaki genel yüz hareketleri gerçek yüzden yeterince farklı olacağından, içeriğin deepfake olduğu tespit edilebilmektedir.

Eğitim ve farkındalık:
Kimlik avı (phishing) saldırılarının önlenmesi ve etkilerinin azaltılması örneğinde olduğu gibi, işverenler; çalışanları iş ortamında dezenformasyon ve deepfake dâhil olmak üzere ilgili tehditleri engelleyebilecek ve/veya raporlayabilecek sahadaki "ilk müdahale ekipleri" hâline getirecek bilgi ve becerilerle donatmak için kaynak ayırmayı değerlendirmelidir.

Kolluk kuvvetleri ve diğer yetkililere yönelik özel eğitimler ise, mağdurların deepfake saldırılarının itibarları, sağlıkları ve refahları üzerindeki etkilerini azaltmalarına yardımcı olmaya odaklanabilir. Daha fazla bilgi için aşağıya bakınız.

Oluşturma Aşamasına Yönelik Zarar Azaltma Önlemleri

Deepfake üretiminde kullanılan modelleri geliştiren kurum ve bireyler, zarar azaltma konusundaki sorumluluklarını da dikkate almalıdır. Geliştirdikleri teknolojilerin sorumsuz veya suç teşkil eden amaçlarla kullanılmasından endişe duyan kişi ve kuruluşlar, kendi modellerinin kullanıldığını tespit etmeyi kolaylaştıracak adımlar atabilirler.

Örneğin geliştiriciler, modellerinin kolayca tespit edilmesini sağlayan bir zayıflık veya ayırt edici bir imza (signature) özelliğinin farkında olabilirler. Bu bilgiyi gizlemek yerine, teknolojik olarak gelişmiş bir toplumun sorumlu bir üyesi olduklarını göstermek amacıyla, söz konusu imzayı kodlarıyla birlikte kamuoyuyla paylaşabilirler.

Ayrıca deepfake içerik üreten kişiler de, oluşturdukları içeriği açıkça "deepfake" olarak işaretleyebilirler.

Yayım (Dağıtım) Aşamasına Yönelik Zarar Azaltma Önlemleri

Ortaklıkların geliştirilmesi: Sanayi, akademi, kolluk kuvvetleri ve diğer ilgili kurumlar arasındaki ortaklıklar; rıza dışı görüntüler ve diğer iftira niteliğindeki sentetik medya içeriklerinin ortaya çıktığında daha hızlı tespit edilmesi, etiketlenmesi ve kaldırılması sürecini hızlandırabilir.

Tespit ve diğer teknolojik yenilikler: Teknolojinin eğlence, eğitim ve ifade özgürlüğü kapsamındaki amaçlarla da kullanılabilmesi nedeniyle, yalnızca deepfake tespitine dayalı bir yaklaşım başlı başına yeterli bir zarar azaltma protokolü oluşturamaz. Bununla birlikte, bu tür teknolojik araçların genel önemi ve etkisi göz ardı edilemez. Başarılı bir tespit, erken müdahaleye ve etkilerin azaltılmasına imkân tanır.

Sosyal medya platformları, internet servis sağlayıcıları ve diğer iletişim sistemleri sağlayıcıları — yani multimedya içerik akışını mümkün kılan altyapıyı sunanlar — ilettikleri içeriğin niteliğini belirleme konusunda en avantajlı konumdadır. Microsoft'un PhotoDNA teknolojisi gibi, daha önce tanımlanmış dijital görüntülerin kopyalarını tespit eden araçlar, bazı sağlayıcılar tarafından çocuk cinsel istismarı materyallerinin (CSAM) yayılımını engellemek amacıyla kullanılmaktadır; dolayısıyla bu konuda bir emsal mevcuttur.

Tespit araçları geliştirildikçe, bunlar sosyal medya şirketleri, internet servis sağlayıcıları ve iletişim sistemi sağlayıcılarıyla paylaşılabilir ve açık kaynak araçlar olarak kullanıma sunulabilir.

Ayrıca, insanları ve modelleri tespit konusunda eğitmek amacıyla kullanılabilecek deepfake içeriklerin üretilmesi ve kontrollü biçimde dağıtılması da bu sürece katkı sağlayabilir.

Sentetik içeriğin tespit edilmesinin karşı tarafında ise, doğrulama (authentication) önlemlerinin teşvik edilmesi fırsatı bulunmaktadır. Daha açık bir ifadeyle, bireyler ve kurumlar; ürettikleri ve tükettikleri medyanın özgünlüğünü göstermek ve doğrulamak için adımlar atabilirler.

Örneğin 2019 yılında Content Authenticity Initiative (CAI) kurulmuştur. CAI kendisini; "içerik özgünlüğü ve kaynağı (provenance) için açık bir endüstri standardının benimsenmesini teşvik etmek üzere çalışan medya ve teknoloji şirketleri, sivil toplum kuruluşları, akademisyenler ve diğer paydaşlardan oluşan bir topluluk" olarak

tanımlamaktadır. Bu standartlar, kullanıcıların medya içeriklerinin kaynağını ve atfını, standartları kullanan herkes için erişilebilir bir biçimde göstermesine imkân tanır. Bu sayede tüketiciler, içerikleri bir "özgünlük mührü" üzerinden kontrol edebilir ve içeriğe duyulan güven artar.

CAI, herkes tarafından benimsenebilecek açık standartlar sunmayı hedeflerken, ticari hizmet sağlayıcılar da benzer doğrulama imkânları sunabilir. Truepic bu alanda faaliyet gösteren şirketlerden biridir.

Son olarak, kullanıcıların tüm içeriği kontrol ettiği güvenli iletişim kanalları (örneğin kapalı ağlar) üzerinden de bu alanda başarı sağlanabilir.

Gerçek Zamanlı Etkileşimlere ve Medyaya Olan Güveni Nasıl Artırabiliriz?

Kamuoyunun gerçek zamanlı etkileşimlere ve medyaya duyduğu güvenin artırılması uzun vadeli bir hedef olmakla birlikte, toplumu ve kurumları dezenformasyondan korumak açısından son derece kritik bir adımdır. İki faktörlü kimlik doğrulama (2FA) ve cihaza dayalı kimlik doğrulama gibi güvenlik protokollerine yeniden sıkı biçimde uyulması, bu sürecin temel ve ilk adımını oluşturmaktadır.

Buna ek olarak, demokratik kurumlarımızın, medya kurumlarının ve yeni ortaya çıkan teknolojilerin güçlendirilmesine yönelik yatırımların artırılmasını hedefleyen bir stratejinin izlenmesi faydalı olacaktır. Blok zinciri (blockchain) tabanlı kimlik doğrulama bu bağlamda öne çıkan bir olasılıktır; zira doğrulama ve özgünlüğün standartlaştırılmasını ve teşvik edilmesini sağlayarak güvenilir bir alan oluşturma potansiyeline sahiptir. Bu sayede, görülen ve duyulan içeriklere yönelik tüketici güveninin artırılması mümkün olabilir.

İzleyici Aşamasına Yönelik Zarar Azaltma Önlemleri

Toplumsal eğitim:
Yanlış bilgiye karşı direnci ele alan ve kamuoyunun gerçeği kurmacadan ayırt etme yeteneğini güçlendiren, daha kapsamlı eğitim faaliyetleri sunmak amacıyla politika ve programların hayata geçirilmesi gerekmektedir. Buna ek olarak, medya okuryazarlığı ve eleştirel düşünme becerilerine yönelik erken yaşta eğitim programlarının geliştirilmesi ve önceliklendirilmesi önemli faydalar sağlayacaktır.

Gerçek Olan ile Manipüle Edilmiş Olanı Nasıl Ayırt Edebiliriz?
Tespiti Nasıl Geliştirebiliriz?
Kamuoyunu Deepfake'leri Tespit Etme Konusunda Eğitebilir miyiz? Eğitmeli miyiz?

Teknoloji ilerledikçe, manipüle edilmiş medyayı tespit etmek giderek daha zor hâle gelecektir. Sahte medyanın tespitine yardımcı olabilecek ticari araçlar mevcuttur; ancak bu araçların, farklı değişkenleri ve çeşitli manipülasyon yöntemlerini yakalayabilmek için sürekli olarak yeniden eğitilmesi ve güncellenmesi gerekecektir. Ayrıca her aracın, neyi

“deepfake” olarak tanımladığı farklılık gösterebilir; bu durum da hangi tür medya manipülasyonlarının işaretleneceğini etkileyecektir.

Deepfake’in ne kadar gelişmiş olduğuna bağlı olarak, bazen kamuoyu kendi gözleriyle bunu fark edebilirken, bazı durumlarda adli bilişim uzmanlarının içeriği çok daha ayrıntılı biçimde analiz etmesi gerekebilir. Bu çalışmanın baştan itibaren yapay zekâ / makine öğrenmesi (AI/ML) araçları tarafından gerçekleştirilmesi, insanların bunu manuel olarak yapmasına kıyasla daha verimli olacaktır; ancak deepfake tespitinin geliştirilmesi, herkesin katkı sunması gereken kolektif bir çaba olacaktır.

Her AI/ML modelinde olduğu gibi, bu süreç deneme–yanılma yoluyla ilerleyecektir. Aracın neyi tanıyabildiğini görmek için sürekli olarak sahte medya içerikleri sistemden geçirilir, kaçırdığı noktalar tespit edilir ve ardından model, bir şeylerin gözden kaçma ihtimalini azaltacak şekilde uyarlanır. Toplumsal açıdan bakıldığında ise, bireylerin bir içeriği paylaşmadan önce onun güvenilirliğini kontrol etmeleri gerekir; böylece yanlış bilginin yayılması engellenebilir.

AI/ML alanındaki uzmanlar — Partnership on AI (PAI) dâhil — tespiti geliştirmek için, “neyin sahte olduğunu bulmaya odaklanmak” yerine, medyada “neyin gerçek olduğunu güçlendirmeye” ve içeriğe bağlam eklemeye yönelik bir paradigma değişiminin gerekli olduğunu önermektedir. Bu yaklaşım, bireylerin içeriğin kaynağına ilişkin bağlamsal ipuçlarını veya meta verileri kullanarak medyanın özgünlüğünü kendi başlarına incelemelerine imkân tanır.

PAI tarafından yapılan görüşmeler, bireylerin neye inanacaklarının kendilerine söylenmesini ya da gerçek–sahte ayrımının üstten bir dille öğretilmesini istemediklerini; bunun yerine, gerçeği kendilerinin keşfetmek istediklerini göstermektedir. Bu durum, paylaşılan medyanın meşruiyetini bireylerin kendilerinin doğrulayabilmesi hâlinde, kurumlara duyulan güveni de artırabilir.

Bireyleri özgün medyaya erişme ve onu tanıyabilme konusunda güçlendirmek için, toplumun deepfake’ler hakkında eğitilmesi gerekmektedir. İnsanlar deepfake’lerin gerçek anlamını ya da ne kadar ciddi zararlar doğurabileceğini bilmiyor olabilir; ancak nelere dikkat etmeleri gerektiği öğretilirse, bu içerikleri kendi başlarına tespit edebilir hâle gelebilirler. Deepfake’lerin ne kadar kolay üretilebildiğinin topluma anlatılması önemlidir; çünkü sosyal medyanın yoğun kullanımı nedeniyle sahte medya, gerçekliği sorgulanmadan çok hızlı biçimde paylaşılabilmektedir.

Eğitimle ilgili temel sorun ise, toplumun deepfake’ler hakkında öğrenmeye istekli olması gerekliliğidir. Bazı bireyler bu konuyla ilgilenmeyebilir ve yalnızca inanmak istedikleri anlatıya inanmayı tercih edebilirler. Ancak toplum, deepfake’leri analiz etmeyi ve anlamayı öğrenemese bile, en azından gördüğü her şeyi paylaşmamayı öğrenebilir.

Bir görüntü veya videonun sahte olup olmadığını belirlemeye çalışırken bireyler aşağıdaki işaretlere dikkat etmelidir:

- Yüz bölgesinde bulanıklık varken görüntü veya videonun diğer kısımlarında olmaması (ya da tam tersi)
- Yüz kenarlarına yakın bölgelerde cilt tonunda ani değişiklikler
- Çift çene, çift kaş veya yüz hatlarında çift kenar görünümü
- Yüzün, bir el ya da başka bir nesne tarafından kısmen kapatıldığında bulanıklaşıp bulanıklaşmadığı
- Aynı video içinde farklı kalite seviyelerine sahip bölümler bulunması
- Ağız, gözler ve boyun çevresinde kutu benzeri şekiller veya kesilip yapıştırılmış (crop) etkileri
- Göz kırpma (ya da göz kırpmama), doğal olmayan hareketler
- Arka plan ve/veya ışıklandırmada ani değişiklikler
- **Bağlamsal ipuçları:** Arka plan sahnesi, ön plandaki kişi ve konu ile tutarlı mı?

Bir ses kaydının sahte olup olmadığını belirlemeye çalışırken bireyler aşağıdaki işaretlere dikkat etmelidir:

- Kesik kesik veya kopuk cümleler
- Konuşma sırasında tonlama ve vurgu değişiklikleri
- **İfade biçimi:** Konuşmacı bunu gerçekten bu şekilde söyler miydi?
- **Bağlamsal ipuçları:**
    - **Mesajın bağlamı:** Mesaj, yakın zamanda yapılan bir konuşmayla tutarlı mı?
    - İlgili veya devam niteliğindeki sorulara mantıklı ve tutarlı yanıtlar verebiliyor mu?

Bir metnin sahte olup olmadığını belirlemeye çalışırken bireyler aşağıdaki işaretlere dikkat etmelidir:

- Yazım hataları
- Cümlelerde akışın olmaması veya kopukluklar
- Gönderenin bilinen bir telefon numarası veya e-posta adresi olup olmadığı
- **İfade biçimi:** Meşru gönderici bunu gerçekten bu şekilde ifade eder miydi?
- **Mesajın bağlamı:** Mesaj, yakın zamanda yapılan bir görüşme veya yazışma ile tutarlı mı?

Bir içeriğin "sahte" ya da manipüle edilmiş olduğu tespit edilirse ne yapılabilir?
İçeriğin paylaşılması veya iletilmesi engellenmeli mi?
İçerik kaldırılmalı mı?
İçerik etiketlenmeli mi?

DeepTrust Alliance'ın kurucusu ve CEO'su Kathryn Harrison'a göre, bir deepfake tespit edildikten sonra şu anda yapılabilecekler oldukça sınırlıdır. Seçeneklerden biri, medyanın ne ölçüde manipüle edildiğinin belirlenmesidir; ancak bu süreç zaman alıcı ve maliyetli olabilir. Harrison, deepfake'in paylaşıldığı platformlardaki Hizmet Şartlarının (Terms of Service) incelenmesine yardımcı olması için bir avukatla çalışılmasını önermektedir. Ancak bu da zaman alıcıdır; çünkü her platformun farklı koşulları vardır ve içeriğin siteden kaldırılabilmesi için fikri mülkiyet (IP) ve telif hakları konularında hukuki müzakereler yapılması gerekebilir.

Bazı durumlarda deepfake'ler, raporlanmak üzere kolluk kuvvetlerinin yetki alanına taşınabilir; ancak tehdidin çok yeni olması nedeniyle henüz yeterli düzenleme bulunmamaktadır. Çoğu kolluk birimi, deepfake vakalarını ele alacak protokollere veya araçlara sahip değildir ve kimi zaman bu tür vakalar yetki alanlarının dışında kalabilmektedir.

Önde gelen AI/ML şirketleri ve sosyal medya platformları, deepfake zararlarının azaltılması açısından önemli odak noktaları olabilir. AI/ML araçları, gelişen teknolojilere uyum sağlamak ve deepfake'leri tespit edebilmek için modellerini zaman içinde güncelleyebilir. Ancak toplumun tamamını, medyanın özgünlüğünü belirlemek amacıyla bu araçları kullanmaya zaman ayırmaya ikna etmek zor olacaktır. Sosyal medya platformları bu araçları sistemlerine entegre ederek içerikleri işaretleyebilir ve deepfake içeriklerin yayılmasını sınırlayabilir. Büyük platformlar AI/ML şirketleriyle birlikte çalışırsa, deepfake içerikler en baştan tespit edilebilir ve hiç yayımlanmayabilir ya da en azından kamuoyunda viral hâle gelmeden önce çok daha hızlı bir şekilde kaldırılabilir.

Deepfake'lerin oluşturduğu tehdidi azaltmak amacıyla tespit ve diğer proaktif önleme çalışmalarına yönelik devam eden araştırmalar ve destekler bulunmaktadır. Ancak deepfake saldırılarından etkilenen mağdurları (bireyler veya işletmeler) desteklemeye daha fazla önem verilmesi gerekmektedir. Avustralya'da hukuk mezunu ve aktivist olan Noelle Martin'e göre, 17 yaşındayken kendisine ait bir fotoğrafın pornografik görüntülere monte edilerek çeşitli pornografi sitelerinde yayıldığını fark etmiştir. Daha sonra bu fotoğraflar deepfake videolara dönüşmüş; bu videolar kendisine e-posta yoluyla gönderilmiş ve çeşitli internet sitelerinde paylaşılmıştır.

Noelle Martin, deepfake mağduru olmanın bir bireyin itibarı, onuru, istihdam edilebilirliği ve kişilerarası ilişkileri üzerinde ömür boyu sürebilecek etkileri olabileceğini belirtmiştir. Bir dönem, kendisi hakkında üretilen deepfake videoların niteliği nedeniyle hukuk alanında iş bulmakta ciddi zorluklar yaşamıştır. Noelle saldırılara karşı sesini

yükseltmeyi tercih etmiş; ancak konuşması, daha fazla çevrim içi saldırıya maruz kalmasına yol açmıştır.

Bugün dahi Noelle, kendisini kimin ya da neden hedef aldığını bilmemektedir ve saldırının gerçekleştiği dönemde bu durumu ele alacak herhangi bir yasal düzenleme bulunmamaktaydı. Peki, deepfake saldırılarından etkilenen mağdurların iyileşmesine yardımcı olmak için ne yapılabilir?

Mağdur Aşamasına Yönelik Zarar Azaltma Önlemleri

Sentetik medya saldırılarının mağdurları, özellikle rıza dışı cinsel içerikli medya saldırılarına maruz kalanlar, içeriğin tüm olası kaynaklardan kaldırılmasının ne kadar zor olduğunu sıklıkla dile getirmektedir. Mağdurlar, aynı içeriğin farklı sitelerde tekrar tekrar ortaya çıkmasının yarattığı kabusu ve içeriğin kaldırılabilmesi için resmî devlet müdahalesini talep etmek zorunda kalmanın yarattığı hayal kırıklığını tanımlamaktadır.

Noelle Martin, sorumlu devlet kurumlarının; resmî olarak kötü niyetli olduğu tespit edilen ve kaldırılması gerektiğine karar verilen içerikleri aktif biçimde arayıp tespit etmede proaktif bir rol üstlendiği bir senaryo önermiştir. Devletler, internet servis sağlayıcıları ve sosyal medya şirketleriyle iş birliği yaparak çocuk cinsel istismarı materyallerine (CSAM) ait **hash** kümelerini belirleme konusunda daha önce çalışmalar yürütmüştür. Aynı sürecin, kötü niyetli deepfake içerikler için de teknolojik olarak tekrarlanması mümkündür.

Deepfake'lerin Bildirilmesi

Deepfake saldırılarına maruz kalan mağdurların bu olayları bildirebilecekleri çeşitli yollar bulunmaktadır. Deepfake mağdurları şu adımları atabilir:

- **Kolluk kuvvetleriyle iletişime geçmek:** Kolluk birimleri, mağdurlardan toplanan deliller ve polis raporları üzerinden adli incelemeler yürüterek yardımcı olabilir.
- **ABD Federal Soruşturma Bürosu (FBI) ile iletişime geçmek:** Olaylar, yerel FBI ofislerine ya da FBI'ın 7/24 hizmet veren Siber İzleme Merkezi'ne (Cyber Watch) **CyWatch@fbi.gov** adresi üzerinden bildirilebilir.
- **Menkul Kıymetler ve Borsa Komisyonu (Securities and Exchange Commission – SEC)** hizmetlerinden yararlanmak: Finansal suçların soruşturulması amacıyla başvuru yapılabilir.
- **Sosyal medya platformlarında uygunsuz içerik ve suistimali bildirmek:** Facebook, Twitter (X), Instagram gibi platformların kendi bildirim prosedürleri kullanılarak içerik raporlanabilir.
- **Mağdur 18 yaşından küçükse:** Olaylar, **National Center for Missing and Exploited Children (NCMEC)** tarafından işletilen siber ihbar hattı üzerinden **https://report.cybertip.org** adresine bildirilebilir.

Mağdurlar İçin Mevcut Kaynaklar

Bobby Chesney ve Danielle Citron'a göre, ikna edici delillerin bulunmaması hâlinde mağdurların medeni sorumluluk (hukuki tazminat) yoluna başvurması zor olabilir; ayrıca kötü niyetli aktör tespit edilse bile, eğer bu kişi Amerika Birleşik Devletleri dışında bulunuyorsa ya da yerel hukuki yolların etkisiz olduğu bir yargı alanındaysa, medeni hukuk yollarının kullanılması imkânsız hâle gelebilir. Bununla birlikte, çevrim içi istismar mağdurlarını farklı şekillerde destekleyebilecek çeşitli kaynaklar mevcuttur. Mağdurlara yardımcı olmayı amaçlayan kuruluşlar arasında şunlar yer almaktadır (bunlarla sınırlı olmamak üzere):

- **Cyber Civil Rights Initiative:** 7/24 hizmet veren bir kriz yardım hattı sunan, avukat yönlendirmeleri yapan ve sosyal medya platformları ile diğer web sitelerinden görüntülerin kaldırılmasına yönelik rehberler sağlayan bir kuruluştur.
- **EndTab:** Üniversiteler, kolluk kuvvetleri, sivil toplum kuruluşları, yargı sistemleri ve sağlık ağları dâhil olmak üzere mağdurlara eğitim ve istismar bildirimi konusunda kaynaklar sunan bir kuruluştur.
- **National Suicide Prevention Lifeline:** Zor durumda olan bireylere ücretsiz ve gizli duygusal destek sağlayan, yerel kriz merkezlerinden oluşan ulusal bir ağdır.
- **Cybersmile:** Siber zorbalık ve çevrim içi nefret kampanyalarının mağdurlarına uzman desteği sağlayan, kâr amacı gütmeyen bir kuruluştur.
- **Identitytheft.gov:** Kimlik hırsızlığı mağdurları için federal hükümet tarafından sunulan tek duraklı bir kaynaktır.
- **Withoutmyconsent.org:** Medeni davalarda kullanılabilecek delillerin korunmasına yönelik rehberler sunan, kâr amacı gütmeyen bir kuruluştur.
- **Google Help Center:** Google tarafından sunulan ve mağdurların Google arama sonuçlarından sahte pornografik içerikleri kaldırmasına imkân tanıyan bir kaynaktır.
- **Imatag:** Görüntü ve video izleme ile takip hizmetleri sunan bir şirkettir.

SON DEĞERLENDİRMELER

"Yalancının Temettüsü" ve Silahlandırılmış Güvensizlik Tehlikesi

Bu çalışma, deepfake'lerin kötü niyetli kullanımından doğabilecek olası sonuçları ele almıştır; ancak birçok okuyucu için sezgisel olmayabilecek başka bir tehdit daha bulunmaktadır. Bu sonuç bölümünde, deepfake'in fiilen kullanılmasından değil, kullanılabilme ihtimalinin bizzat kendisinden kaynaklanan olası sonuçlar değerlendirilmektedir. Deepfake'lerin varlığı dahi; basın, devlet ve akademi gibi

geleneksel toplumsal kurumların güvenilirliğini ve otoritesini zayıflatma potansiyeline sahiptir.

Akademisyenler Danielle K. Citron ve Robert Chesney tarafından Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security adlı çalışmada ilk kez ortaya konan "Yalancının Temettüsü (The Liar's Dividend)" kavramı çerçevesinde, deepfake tehditleriyle mücadeleye yönelik iyi niyetli çabaların nasıl ters etki yaratabileceği incelenmektedir. Kamuoyu, kötü niyetli aktörlerin geniş çaplı panik ya da zarar yaratmayı amaçlayan ayırt edilemez deepfake'lerine karşı zamanla dayanıklılık geliştirebilir. Deepfake olgusuna ilişkin medya haberleri bu farkındalığı artırmakta ve toplumsal direnci güçlendirmektedir.

Ancak sarkaç bu kez ters yöne aşırı savrulursa — yani toplum yalnızca eleştirel bir bakış geliştirmekle kalmayıp varsayılan olarak şüphe ve inkâr refleksiyle hareket etmeye başlarsa — bu durum da kötü niyetli aktörler tarafından gerçekliği bulandırmak amacıyla istismar edilebilir. Bu kalıcı tehdit; güvenilir kurumları zayıflatan, gerçek içerik ve medyanın meşruiyetini ve özgünlüğünü sorgulayan yaygın bir kuşkuculuk iklimi yaratabilir. Kötü niyetli aktörler, bilinçli biçimde güven erozyonu yaratarak, gerçek ve meşru içeriklerin aslında karmaşık bir deepfake olduğu iddiasını ortaya atabilir.

Siyasal yelpazenin kutuplaştığı, çatışmacı bir atmosferin ve 7/24 medya döngülerinin hâkim olduğu bir ortamda; örneğin ciddi bir skandalla karşı karşıya kalan bir siyasetçi, kolaylıkla "Bu olay hiç yaşanmadı. Bu, siyasi düşmanlarım tarafından üretilmiş açık bir deepfake'tir" diyerek hesap vermekten kaçınabilir ve itibarını zedelenmeden koruyabilir.

Daha sofistike aktörler ise, daha önce kaydedilmiş gerçek bir olayı yapay zekâ / makine öğrenmesi (AI/ML) teknolojileriyle sentetik olarak yeniden üretebilir ve kasıtlı olarak tespit edilebilir bir imza ekleyerek doğrulama ve tespit sistemlerini tetikleyebilir. Bu durum, gerçek içeriğin meşruiyetinin baştan sorgulanmasına yol açabilir. Böylece kötü niyetli aktörler, yeniden ürettikleri içeriği işaret ederek olayın aslında hiç yaşanmadığını iddia edebilir.

Tarihin hem iyi hem de kötü yönlerinin kayda geçirilmesi, toplumun ahlaki gelişimi açısından hayati bir araçtır. Günümüzde dahi Holokost'un hiç yaşanmadığını iddia edenler bulunmaktadır. Deepfake'ler, tarihin güvenilirliğini ve kolektif hafızayı baltalamak için kullanılabilecek son derece tehlikeli bir araç hâline gelebilir.

Bundan Sonra Ne Yapılmalı?

Deepfake'ler, sentetik medya ve genel olarak dezenformasyon, toplumumuz için ciddi zorluklar oluşturmaktadır. Bu tehditler; küçük işletmelerden ulus-devletlere kadar bireyleri ve kurumları etkileyebilmektedir. Herkes bu etkilerden payını alabilir.

Yukarıda da tartışıldığı üzere, bu zorlukların etkisini azaltmaya yardımcı olabilecek bazı yaklaşımlar bulunmaktadır ve henüz tespit edemediğimiz başka yaklaşımların da varlığı muhtemeldir. Ancak hangi yaklaşım benimsenirse benimsensin, herhangi birinin başarılı olabilmesi; etkilenen tüm taraflar arasında **iş birliği** gerektirecektir.

Artık **koordine ve iş birliğine dayalı bir yaklaşımın** zamanı gelmiştir. Ekibimiz, önümüzdeki yıllarda bu iş birliğinin bir parçası olmayı umut etmektedir.