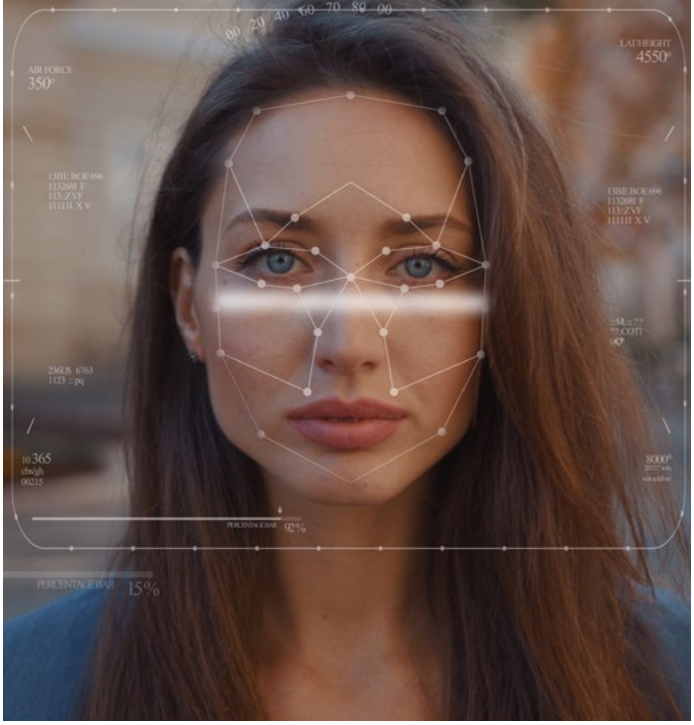
thinktech
STM Teknolojik Düşünce Merkezi



Yapay Zekâ ile Üretilen İçerikler Tespit Edilebilir mi?

Yapay zekâ günümüz teknolojilerinin en önemlilerinden biri hâline geldi. Akıllı telefonlardan akıllı ev sistemlerine ve elektrikli ev aletlerine kadar her yerde önem kazanan bu teknoloji dijital dünyada da içerik üretmek ve işlemleri hızlandırmak için kullanılıyor. Sosyal medyanın önemli bir güç hâline geldiği bu yüzyılda ise yapay zekâ ile desteklenen ve deepfake adı verilen sahte yüz veya içerik yaratma teknolojisi ise toplumun bazı kesimlerince endişeyle karşılanıyor.

Deepfake Nedir?

Deepfake, yapay zekâ desteği ve derin öğrenme teknolojisi yardımıyla fotoğraf ve diğer görsellerden faydalanılarak sahte olaylar veya görseller yaratma teknolojisidir. Uzun yıllardır özellikle kadınların hedef alındığı ve yetişkin film endüstrisinde kullanılan deepfake günümüzde herkesin kolaylıkla erişebileceği bir teknoloji hâline geldi. İlk versiyonlarında sadece fotoğraf veya basit videolara uygulanan deepfake ile yapılan içerikler artık ses klonlamalarıyla daha da gerçekçi hâle gelebiliyor[1].

Bir diğer tanımla aslında deepfake; makine öğrenmesi algoritmalarından yararlanan yapay zekâ ile gerçekçi medya içeriği yaratmak olarak değerlendiriliyor.

Deepfake teknolojisi temelde Çekişmeli Üretici Ağlar (Generative Adversarial Network -GAN) isimli makine öğrenmesi tekniğine dayanıyor. GAN, bir görseli tanımak üzere kendini eğitmek için bir dizi algoritma kullanır. Bu eğitim sahte görüntüler üretebilmek için gerçek özellikleri öğrenmesine yardımcı olur. İki farklı derin öğrenme algoritmasının karşılıklı çalıştığı bu yapıda biri en iyi sonucu bulmaya çalışırken diğeri bunu ayırt etmeye çalışır. Böylece son aşamada ayırt edilemez bir sonuç ortaya çıkmış olur.

Deepfake'in en bilinen uygulamaları, bir insan yüzünün başka bir insana video veya fotoğraf medyaları üzerinde uygulanmasıyla yapılan yüz değiştirme ve ses klonlaması olarak öne çıkıyor. Deepfake teknolojisi görüntü ve ses manipülasyonu dışında yazılı medyada da kullanılabiliyor. Özellikle sosyal medya fenomenleri veya ünlü kişilerin yazı yazma stillerinin kopyalanarak içerik üretilen bu yöntem reklam veya karalama kampanyalarında da yer edinebiliyor[2].

1 https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them
2 https://flashpoint.io/blog/what-is-deepfake-technology/

Addison Rae Easterling, Charlie D'Amelio ve Bella Parch gibi bilinen TikTok sosyal medya fenomenleri deepfake'in kötü amaçla kullanımlarından nasiplerini alan ünlülerden. Fotoğraflarına yapılan müdahaleler bu ünlüleri hayranlarına karşı zor durumda bırakmıştı. Benzer şekilde Julia Roberts, Emma Watson ve Scarlett Johansson gibi Hollywood ünlüleri de deepfake saldırılarına maruz kalmıştı[3].

Nisan 2018'de, Hindistan'da yaşayan bir gazeteci olan Rana Eyyub, iktidar partisi BJP hakkında eleştirel bir makale yazdıktan sonra, yüzünün farklı içerikte video üzerine eklenmesiyle bir deepfake saldırısıyla karşı karşıya kalmıştı. Video sonrasında suçlanması, yaşadığı taciz ve aşağılama Eyyub'un kalp rahatsızlığı ile hastaneye kaldırılmasına ve sosyal medyadan uzaklaşmasına neden olmuştu[4].

Nicolas Cage, Nancy Pelosi gibi ünlülerle Mark Zuckerberg gibi milyonerler de deepfake videoları ile son yıllarda gündeme gelen isimler[5].

2020'de hacker'ler, bir Hong Kong bankasından 35 milyon dolar çalmak için deepfake ses klonlama yöntemini kullandılar. Business Identity Compromise (BIC) adı verilen ve başkasının kimliğinin illegal olarak kullanıldığı bir yöntemle gerçekleştirilen soygun modeli dünya üzerinde en büyük maddi kayıpla sonuçlanan olaydır.

BIC, sentetik kurumsal kişilikler oluşturmak veya mevcut çalışanları taklit etmek için deepfake teknolojisini kullanıyor. Genellikle kuruluşta tanınmış, üst düzey bir yöneticinin bilgileri ve sesi klonlanarak gerçekleştiriliyor[6].

Deepfake'in olumlu amaçlarla kullanıldığı alanlar da bulunuyor. Eğitim, yakınlarını kaybedenlerin yaşadığı travmayı atlatmaları amacıyla yapılan terapiler, eğlence, sanat, endüstriyel amaçlar ve birçok sektörel uygulamalarda deepfake kullanılabiliyor. Hatta bazı teleskopik görüntülerin netleştirilerek iyileştirilmesinde deepfake teknolojisinin kullanıldığı biliniyor.

Deepfake teknolojisi çeşitli akıllı telefon uygulamaları aracılığıyla bir film, fotoğraf karesi veya müzik videosunda bulunan kişilerle kullanıcıların yüzlerinin değiştirilmesi gibi amaçlarla da yaygın bir şekilde kullanılıyor. Deepfake teknolojisinin kullanımı aslında 1990'lara dayanıyor. Ancak son yıllarda güçlenen yapay zekâ bu teknolojinin olumsuz amaçlarla kullanım tespitini oldukça zorlaştırıyor. Deepfake teknolojisinin olumsuz amaçlarla kullanımına karşı oluşturulan yaklaşım ve uygulamalara ise anti-deepfake adı veriliyor[7].

Anti-deepfake Nedir?

Deepfake ile oluşturulan her türlü medya aracının tespiti için kullanılan yöntemler anti-deepfake olarak tanımlanıyor. Deepfake teknolojisi her zaman kötücül amaçlarla kullanılmasa da tespit edilmesi önemli bir konudur. Günümüzde film endüstrisinde bir aktörün daha yaşlı veya genç görünmesi veya hayatını kaybetmiş bir aktörün devam filmlerinde yer edinmesine olanak sağlayan deepfake teknolojisi iyi amaçlar dışında kullanıldığında büyük riskler yaratabiliyor. Ukrayna savaşında Rus hacker'lar tarafından Ukrayna Devlet Başkanı için oluşturulan deepfake videosu ile Ukrayna askerlerine teslim olma çağrısı yapılması bunun örneklerinden sadece biri. Bu nedenle deepfake'e karşı anti-deepfake uygulamalarının çok iyi anlaşılması ve çalışılması gerekiyor. Anti-deepfake video, ses ve yazı içeriği açısından üç kademede uygulanabiliyor. Özellikle yüz manipülasyonlarında kullanılan tipik deepfake uygulamaları şu şekilde tanımlanabiliyor.

3 https://www.marca.com/en/lifestyle/celebrities/2023/06/15/648afcc246163fac6d8b4599.html
4 https://towardsdatascience.com/deepfakes-harms-and-threat-modeling-c09cbe0b7883
5 https://www.garbo.io/blog/deepfakes
6 https://business.bofa.com/en-us/content/cyber-security-journal/deepfakes-business-risks.html
7 https://builtin.com/machine-learning/deepfake

Görünür geçişler: Bir kişinin yüzü başka bir kişinin yüzüyle değiştirildiğinde veya üst üste bindirildiğinde yüzün kenarlarında bazı noktalarda ten rengi ve dokusu geçişleriyle birlikte görünebiliyor ve bu durum bozulmalara neden oluyor. Bu sayede orijinal yüzün bazı bölümlerinde fark edilebilir deformasyonlar oluşuyor.

Keskin konturlarda bulanıklık: Deepfake yazılım algoritmalarının büyük bir kısmı daha yakından incelendiğinde bulanık görünen ve özellikle dişlerde ve gözlerde bulunan detaylarda fark edilebilir dokusal farklılıklar yaratan hatalar oluşturabiliyor.

Sınırlı yüz ifadeleri ve gölgeler: Deepfake için uygulanan modeldeki veri eksikliği, modelin yüz ifadelerini veya aydınlatma durumlarını doğru bir şekilde uygulayamayabiliyor. Özellikle yüzün profil görünümü yetersiz ise başın hızlı hareketleri bulanıklık veya görsel hatalarla sonuçlanabiliyor.

Ses ile ilgili deepfake uygulamalarında ise metalik ses, monoton konuşma şekli, dikte hataları veya doğal olmayan ses çıkışları kendini ele veriyor. Deepfake'in tespitinde en çok zorlanılan alan ise yazılı medya uygulamalarında ortaya çıkıyor. Bu durumda anti-deepfake'in kaynak kontrolü, mevcut trend ve yaklaşımların değerlendirilmesi veya doğrudan kaynağa erişimle onay alınması şeklinde olabiliyor[8].

Deepfake'in tespitinde geçmişte geleneksel insan kontrollü yöntemler kullanılıyordu. Günümüzde ise bu teknolojiye destek veren yapay zekâ aynı zamanda deepfake ile mücadelede kullanılabiliyor. Özellikle Facebook, Twitter ve Google gibi internet devleri ellerinde bulunan güçlü altyapılarla başarılı anti-deepfake çalışmaları yapmayı vadediyor. Anti-deepfake için kullanılacak yapay zekânın makine öğrenmesi algoritmaları çok yüklü miktarda veriye ihtiyaç duyuyor. Makine öğrenmesinin ihtiyaç duyduğu verileri en iyi şekilde sağlaması ise Facebook, Google, Amazon Web Hizmetleri ve Microsoft'un ortaklaşa oluşturduğu Deepfake Tespiti Meydan Okuması (Deepfake Detection Challenge -DFDC) gibi uygulamalarla mümkün kılınıyor[9].

2019 yılında başlatılan DFDC'nin tamamlandığı 2020 yılında, 2.000'den fazla takım, oluşturdukları veri setlerini kullanarak 1 milyon dolarlık toplam ödülün 500 bin dolarlık birincilik kısmı için yarıştı[10].

Yarışmada herkese açık veri setlerinde en iyi performans gösteren model yüzde 82,56 oranla deepfake medyalarını tespit etmeyi başardı. Ancak katılımcıların kara kutu veri kümesine göre değerlendirmelerinde ise en iyi performans gösteren modellerin sıralaması önemli ölçüde değişti. Bu kategoride en yüksek performans gösteren katılımcı Selim Seferbekov'un oluşturduğu model ancak yüzde 65,18'lik bir başarı elde edebildi[11].

DFDC gibi yarışmalar günümüzde anti-deepfake modellemelerinin temelini oluşturuyor ve bu yarışmalarla oluşturulan veri setleri yeni nesil yapay zekâların deepfake ile mücadelelerine destek oluyor. Bu mücadelede bazı önemli noktaların bilinmesinde ise fayda var.

Deepfake'i Tespit Etme Yolları

Videolarda özellikle dikkat çeken anormal göz hareketleri, deepfake'in tespit edilmesine yardımcı oluyor. Gözlüklerde yaşanan aşırı veya orantısız parlama ile renk ve ışık uyumsuzlukları da deepfake'i ele veren diğer unsurları oluşturuyor.

Üst düzey Deepfake manipülasyonları neredeyse her zaman yüz dönüşümlerinden oluşuyor.
Deepfake ile bıyık, favori veya sakal eklenebiliyor veya kaldırılabiliyor. Ancak deepfake, sakal dönüşümlerini

8 https://golden.com/wiki/Anti-deepfake_technology-R9VDX65
9 https://www.vontobel.com/en-ch/impact/in-the-battle-against-deepfakes-ai-is-being-pitted-against-ai-18263/
10 https://www.kaggle.com/competitions/deepfake-detection-challenge/overview/prizes
11 https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

tamamen doğal hâle getirmede başarısız olabiliyor. Ayrıca deepfake uygulanan kişinin doğal sıklıkla göz kırpıp kırpmadığı belirgin bir şekilde görülebiliyor[12].

Kontrast ve ses kalitesinde yaşanan değişimler de deepfake'in tespit edilmesini sağlayan bazı önemli veriler olarak öne çıkıyor. Vücut orantısızlıkları, anormal hareketler, suni yüz mimikleri gibi başka konular da deepfake tespitinde önemli bir rol oynuyor[13].

Mevcut deepfake yazılımları ağız içini, dili veya dişleri yeterli detayda oluşturamıyor. Dolayısıyla bu alanlarda oluşan bozulmalar da deepfake'i ele verebiliyor[14].

Anti-deepfake için yapay zekâ tespit yazılımları kullanılabiliyor. Gelişmiş yapay zekâ uygulamaları ile donatılmış yazılımlar yine yapay zekâ ile oluşturulan deepfake medyalarında bırakılan izleri kolaylıkla tespit edebiliyor. Ayrıca oluşturulan orijinal içeriğe uygulanacak filigranlar da deepfake yöntemlerince engellenemediğinde kaynağa ulaşmada fayda sağlıyor. Son olarak medyanın kaynağının korunması ve kontrolü de anti-deepfake uygulamalarının önemli bir kısmını oluşturuyor. Ancak anti-deepfake alanında özellikle uluslararası regülasyon ve standartların olmamasından kaynaklı teknik sorunlar yaşanabiliyor. Son yıllarda yaygınlaşan açık kaynak yapay zekâ yazılımlarında ise filigran kullanılmadığından bir güvenlik sorunu ortaya çıkıyor[15].

Yapay zekâ tarafından oluşturulan deepfake görüntülerini tespit etmenin önemli yollarından biri de kaynağın doğrulanmasıdır. Resmin sosyal medya dışında herhangi bir yerde öne çıkarılıp çıkarılmadığını görmek için Google veya Bing gibi arama motorlarında görsel araması yapılabilir. Bilinen ve güvenilir bir medya kuruluşunun ortaya çıkan olay hakkında haber yapıp yapmadığı kontrol edilebilir[16].

ChatGPT gibi sohbet robotları, resmi adı ile Büyük Dil Modelleri (Large Language Models -LLM) tarafından oluşturulan metinler de iş dünyasında soru işaretleri yaratıyor. Çalışanların işlerini doğru yapmaması veya içeriği kontrol etmemesiyle sonuçlanabilecek veya işyerini zarara uğratabilecek içeriklerin sisteme dahil olmasına sebep olabilecek yapay zekâ destekli chatbot'lar ayrı bir riski ortaya çıkarıyor. Neyse ki chat robotlarının oluşturduğu içerikleri tespit edebilen yazılımlar mevcut. ChatGPT'yi geliştiren OpenAI firması aynı zamanda yapay zekâ sınıflandırıcı bir yazılım ile insan ve yapay zekâ kullanılarak yazılan metinlerin farklarını ayırt edebiliyor. Bu yazılım yapay zekâ uygulamasını yüzde 99 oranında tespit ediyor[17].

Kullanıcı: Bana bir metin yaz
ChatGPT: metin yazar
Kullanıcı : Bu metni sanki sen değil insan eliyle yazılmış gibi yeniden yaz
ChatGPT: Metin yazar
Kullanıcı: Bunu sen mi yazdın?
ChatGPT: Evet

Anti-deepfake için kullanılan yöntem ve anahtar konular yapay zekâ uygulamalarıyla birleştirildiğinde, günümüz medya içeriklerinde deepfake ile mücadelede bir adım daha önde olunabiliyor.

12 https://www.media.mit.edu/projects/detect-fakes/overview/
13 https://q5id.com/blog/how-to-spot-and-identify-a-deepfake-7-proven-techniques
14 https://www.telefonica.com/en/communication-room/blog/what-is-a-deepfake-and-how-to-detect-it/
15 https://www.analyticsvidhya.com/blog/2023/05/how-to-detect-and-handle-deepfakes-in-the-age-of-ai/
16 https://bdtechtalks.com/2023/05/12/detect-deepfakes-ai-generated-media/
17 https://www.businessinsider.com/how-to-detect-ai-generated-content-text-chatgpt-deepfake-videos-2023-3

Anti-deepfake Uygulama Örnekleri

Deepfake gibi kötücül içerikleri tespit etmek için Google'ın uyguladığı "Tecrübe, Uzmanlık, Yetki ve Güvenilirlik (Experience, Expertise, Authoritativeness, and Trustworthiness -E-E-A-T) politikası gibi yaklaşımlar özellikle internet ortamında bulunan medyaların kalitesinin kontrol altında tutulmasını sağlayabiliyor. Bu politikaya uymak isteyen içerik üreticilerinin ise ciddi tecrübe ve teknik bilgiye ihtiyacı bulunuyor. Ancak bu yolla yapay zekâ müdahaleleri kolaylıkla tespit edilebiliyor[18].

2022'nin Kasım ayında önemli teknoloji firmalarından ve ileri gelen çip üreticilerinden biri olan Intel'in açıkladığı Gerçek Zamanlı Deepfake Algılayıcısı (Real-Time Deepfake Detector) yapay zekâ kullanarak çoklu katmanlı medya içeriklerinden değişikliğe uğrayanları tespit edebiliyor. Intel'in bir işlemcisi bu konuda aynı anda 72 adet videoyu analiz edebiliyor. Fake Catcher adı verilen yazılım fotopletismografi (photoplethysmography -PPG) işlemiyle yüzdeki renk değişimlerini algılayabiliyor. Şu ana kadar yüzde 91 başarı elde eden yöntem gelecek için umut vadediyor[19].

Sensity gibi çevrimiçi platformlar anti-deepfake konusunda genel bir fayda sağlıyor. Platform çoklu formatta medya yüklemeye izin vererek ortalama bir saniye gibi kısa bir sürede yüzde 98,1 başarı gösterebiliyor. Sensity özellikle arkadaşlık siteleri gibi ortamlarda sahte kimliklerin tespitinde önemli bir rol oynuyor.

Deepware Scanner ise açık kaynak kodlu adli bir yazılım olarak biliniyor. Deepware Scanner 2018 yılından beri topladığı veri setleriyle canlı video yayınlarında dahi deepfake tespiti için kullanılabiliyor.

Microsoft'un Demokrasiyi Savunma Programı (Microsoft Defending Democracy Program) çerçevesinde oluşturduğu Microsoft Video Doğrulayıcı statik resim ve videoları gerçek zamanlı olarak inceleyebiliyor. Bu yazılımın DFDC veri setleri Face Forensics++ adlı bir platform ile geliştiriliyor. Microsoft bu tarz yazılımların tek başına yeterli başarı elde edemeyeceğini de düşünerek NewsGuard gibi girişimlerle işbirliği içinde çalışıyor[20].

Deepfake-o-Meter ve Reality Defender gibi diğer yazılımlar da anti-deepfake konusunda önemli katkılar sunuyor[21].

ID R&D tarafından geliştirilen IDLive Face Plus ise yüz değiştirme gibi yüksek risk yaratan deepfake medyalarını gerçek zamanlı olarak tespit edebiliyor. Yüz tanıma sistemleri, biyometrik bir selfie'nin canlı olduğundan emin olmak için biometrik sistemleri aldatmayı amaçlayan Presentation Attack Detection (PAD) yöntemini uyguluyor. Enjeksiyon saldırıları (injection attacks) uygun bir görüntü yakalama sürecini atlamak için donanım ve yazılım saldırılarının kullanıldığı farklı bir güvenlik açığı oluşturuyor. Güvenlik önlemlerine rağmen dolandırıcılar, belirli canlılık algılama önlemlerini alt edebilecek şekilde, canlı olmayan dijital yüz görüntüleriyle kamera görüntü kayıtlarını taklit edebiliyor. Deepfake, dolandırıcıların sahte hesaplar açmak veya kurbanlarının hesaplarına yetkisiz erişim elde etmek için kullandıkları sentetik kimlikler oluşturmak için kullanılabiliyor. IDLive Face Plus bu gibi deepfake risklerini alt etmek için kullanılıyor[22].

18 https://contenthacker.com/can-google-detect-ai-content/
19 https://spectrum.ieee.org/deepfake
20 https://antispoofing.org/deepfake-detection-software-types-and-practical-application/
21 https://www.emergingtechbrew.com/stories/2021/09/15/antideepfake-company-truepic-authenticates-images-rather-detecting-fakes
22 https://www.idrnd.ai/idlive-face-plus-injection-attack-detection-deepfake-protection/?gad=1&gclid=CjwKCAjwqZSlBhBwEiwAfoZUIEoFP9KbxujIlcacDXgSXZeF39kjAUQsg75yDcXD3pS-xMT3XbU6BRoCr9EQAvD_BwE&gclsrc=aw.ds

Deepfake Gelecekte Ne Gibi Tehditler Oluşturuyor?

Teknolojinin gelişim hızına bakıldığında, kötücül deepfake protokolleri yakın gelecekte her şirketin güvenlik stratejisinin bir parçası hâline gelecek gibi görünüyor. Şirketlerin deepfake risklerine proaktif bir şekilde tepki verebilmesini sağlamak için bir adım önde olması ve şimdiden hazırlanmaya başlaması gerekiyor. Deepfake konularının güvenlik bilinci eğitimlerine dahil edilmesi, bir yanıt stratejisi taslağının oluşturulması ve bir tespit modelinin uygulanması bu konuda atılacak önemli adımları oluşturuyor[23].

Deepfake teknolojisi günümüzde anti-deepfake yöntemlerinden daha hızlı gelişiyor ve bu durum hem toplum hem de şirketler için bir tehdit oluşturma potansiyeli gösteriyor[24].

Önemli anket kuruluşlarından biri olan Europol, deepfake'in hangi bilgi kaynaklarının güvenilir olduğu konusunda toplumsal bir kafa karışıklığı yaratabileceğinden endişe ediyor.

Şirketler görsel/işitsel bir medyayı kabul etme veya reddetme noktasında sistemsel önlemler alabildikleri için anti-deepfake konusunda sosyal toplum açısından biraz daha güçlü bir konumda bulunuyor. Önemli olan konunun anti-deepfake uygulamalarının yaygınlaştırılarak toplumun her kesiminin kullanımına sunulması ve daha güvenli bir teknoloji geleceğinin oluşturulması olarak düşünülüyor[25].

Deepfake teknolojisi geliştikçe büyük teknoloji şirketlerinin de desteğiyle anti-deepfake yöntemlerinin güçlenmesi ve bu teknolojinin kötücül amaçlarla kullanımının önüne geçilmesi için daha geçerli uluslararası regülasyon ve standartlar getirilmesi gerekiyor. Bu konuda her ülkeden farklı görüşlerin ve bilim insanlarının katılımıyla oluşturulacak bir ortak platform hızla güçlenen deepfake teknolojisiyle mücadelede kritik bir rol oynayabilir. Geleceğin güvenli teknolojilerinin oluşumunda teknolojiyi geliştiren sorumluların alacağı ve uygulayacağı önlemler deepfake ile mücadelede önem kazanacak gibi görünüyor.

23 https://www.accenture.com/nl-en/blogs/insights/deepfakes-how-prepare-your-organization
24 https://services.global.ntt/en-us/insights/blog/the-real-danger-of-deepfakes
25 https://www.securityweek.com/deepfakes-are-growing-threat-cybersecurity-and-society-europol/