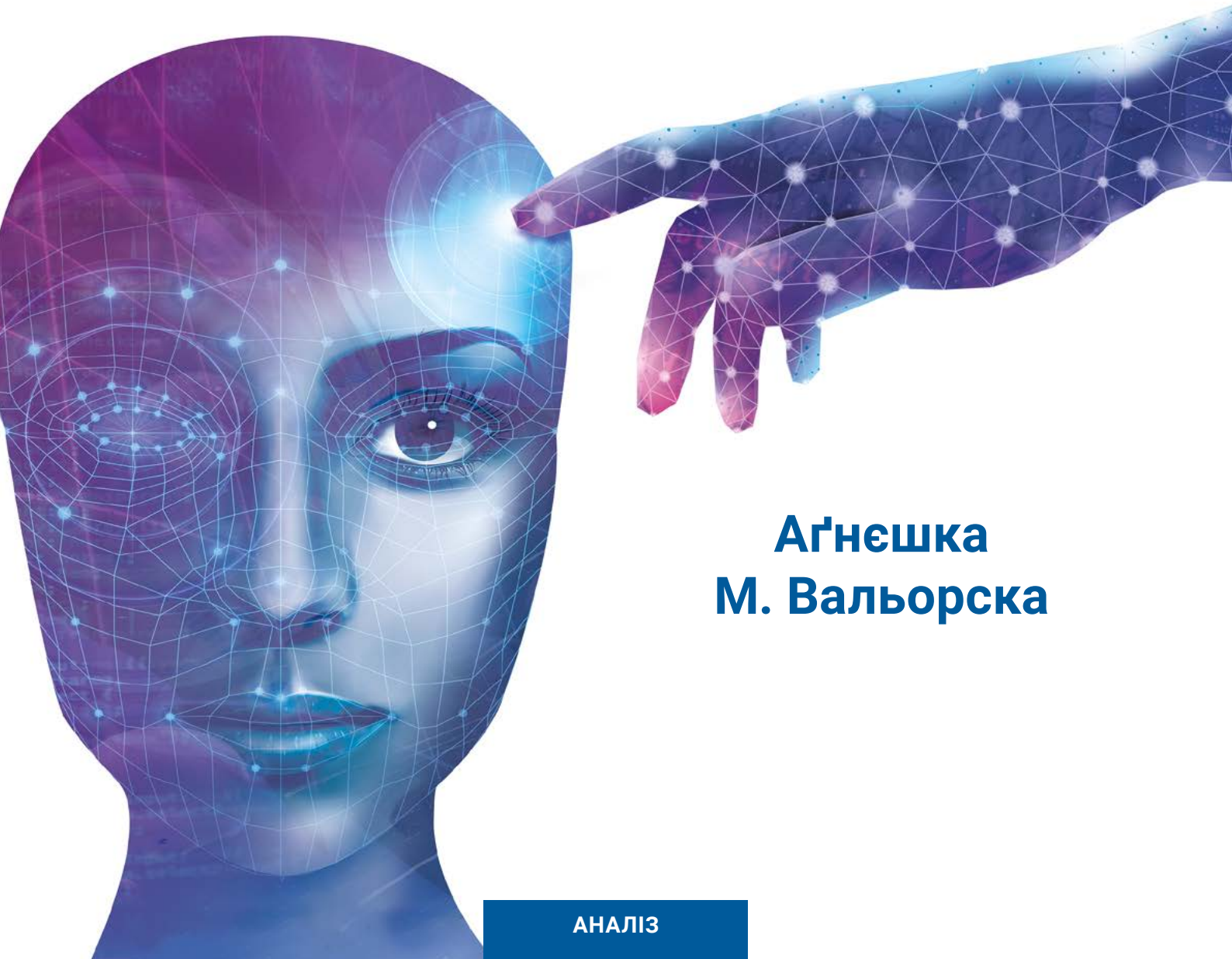
FRIEDRICH NAUMANN FOUNDATION For Freedom.
Ukraine and Belarus

Академія Української Преси

ДІПФЕЙК
ТА ДЕЗІНФОРМАЦІЯ

Бібліотека масової комунікації та медіаграмотності Академії Української Преси



Аґнєшка М. Вальорска

АНАЛІЗ

УДК 316.772.5-043.98:07]+070.16](07)
В16

Текст друкується за виданням:
Agnieszka M. Walorska «DEEPFAKES & DESINFORMATION», Friedrich-Naumann-Stiftung für die Freiheit, S. 31

Переклад з німецької мови *Володимира Олійника*
Коректура *Тетяна Заіченко*

Вальорска М. Аґнєшка
В16 Діпфейк та дезінформація : практ. посіб. / Аґнєшка М. Вальорска ; пер. з нім. В. Олійника – К. : Академія української преси ; Центр Вільної Преси, 2020. – 36 с.

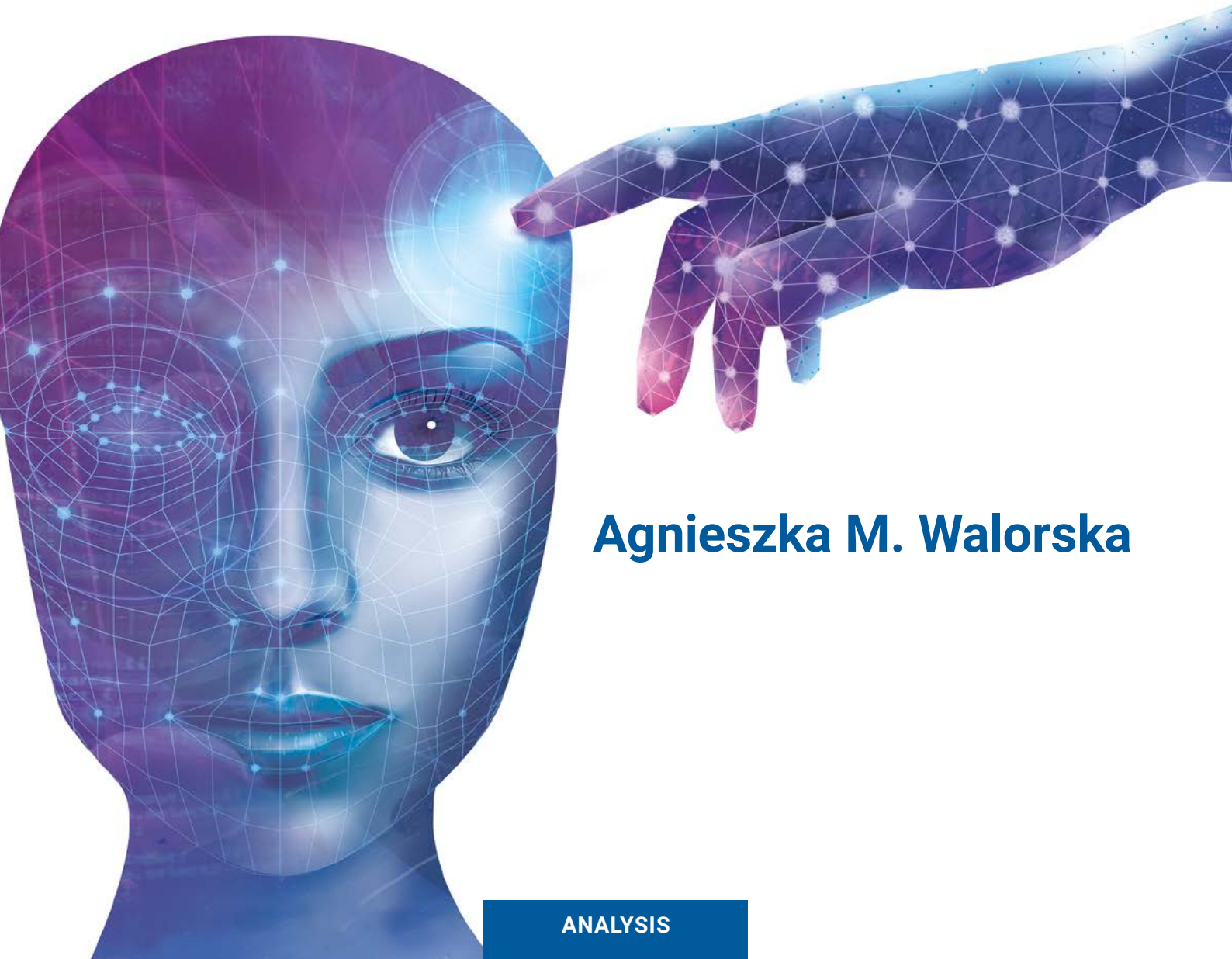
УДК 316.772.5-043.98:07]+070.16](07)

ISBN 978-617-7370-15-3

Gefördert durch die Bundesrepublik Deutschland
За підтримки Федеративної Республіки Німеччина

FRIEDRICH NAUMANN FOUNDATION For Freedom.

DEEPFAKES & DISINFORMATION



Agnieszka M. Walorska

ANALYSIS

Видавець
Фонд Фрідріха Науманна за Свободу
Карл-Маркс-Штрассе, 2
14482 Потсдам

/freiheit.org
/FriedrichNaumannStiftungFreiheit
/FNFreiheit

Авторка
Аґнєшка М. Вальорска

Редакція
Відділ «Глобальні теми»
Секція «Міжнародні справи»
Фонд Фрідріха Науманна за Свободу

Концепція та макет
TroNa GmbH

Контакт
Телефон: +49 30 22 01 26 34
Факс: +49 30 69 08 81 02
E-mail: service@freiheit.org

Стан:
Травень 2020

Список зображень
Фотомонтажі
© Unsplash.de, © freepik.de, S.30 © AdobeStock

Скріншоти
S. 16 © https://youtu.be/mSaIrz8lM1U
S. 18 © deepnude.to Аґнєшка М. Вальорска
S. 19 © thispersondoesnotexist.com
S. 19 © linkedin.com
S. 19 © talktotransformer.com
S. 25 © gltr.io
S. 26 © twitter.com

Усі інші фото
© Friedrich-Naumann-Stiftung für die Freiheit
S. 31 © Agnieszka M. Walorska

ЗМІСТ

Зміст

ОСНОВНІ ФАКТИ ТА ВИСНОВКИ

Використання штучного інтелекту (ШІ) відіграє все більш важливу роль у нашому суспільстві, але нові можливості цієї технології також тягнуть за собою нові ризики. Одним із таких ризиків є зловживання технологією задля навмисного поширення неправдивої інформації.

Політично мотивована дезінформація, безумовно, не є новим явищем, але технологічний прогрес робить створення та розповсюдження маніпульованого контенту набагато простішим та ефективнішим, ніж раніше. За допомогою алгоритмів ШІ відео тепер можна підробляти швидко та відносно дешево (діпфейки), не потребуючи для цього спеціальних знань.

Хоча дискусії точаться про можливе використання діпфейків насамперед у виборчих кампаніях, але цей тип відео складає лише частину всіх маніпуляцій: у 96 відсотках випадків діпфейки використовуються для створення порнографічних фільмів зі знаменитими жінками.

Але й жінки, які не перебувають у сфері уваги громадськості, також можуть опинитися мимовільними виконавицями головних ролей у маніпульованих відео (діпфейк як порнографія помсти). Крім того, статичні зображення за допомогою застосунків на кшталт DeepNude можуть бути перетворені в оманливо реальні фото ню. Не дивно, що ці застосунки функціонують лише із зображеннями жіночих тіл.

Але не тільки візуальним контентом можна маніпулювати або його алгоритмічно продукувати. Голоси, генеровані ШІ, вже успішно використовуються для шахрайства, пов'язаного зі значними фінансовими збитками, а GPT-2 можна використовувати для створення текстів, які вигадують будь-які факти та цитати.

Як нам найкраще обходитися з цим викликом? Компанії та науково-дослідні інститути вже вкладають великі суми в технологічні рішення для ідентифікації генерованих ШІ відео. Вигода від цих інвестицій зазвичай не є тривалою: щойно результати стають загальнодоступними, розробникам діпфейків вдається прилаштуватися до цього, прогрес завжди йде на користь обом сторонам. З цієї причини слід наполягати на більшій відповідальності платформ, які поширюють маніпульований контент.

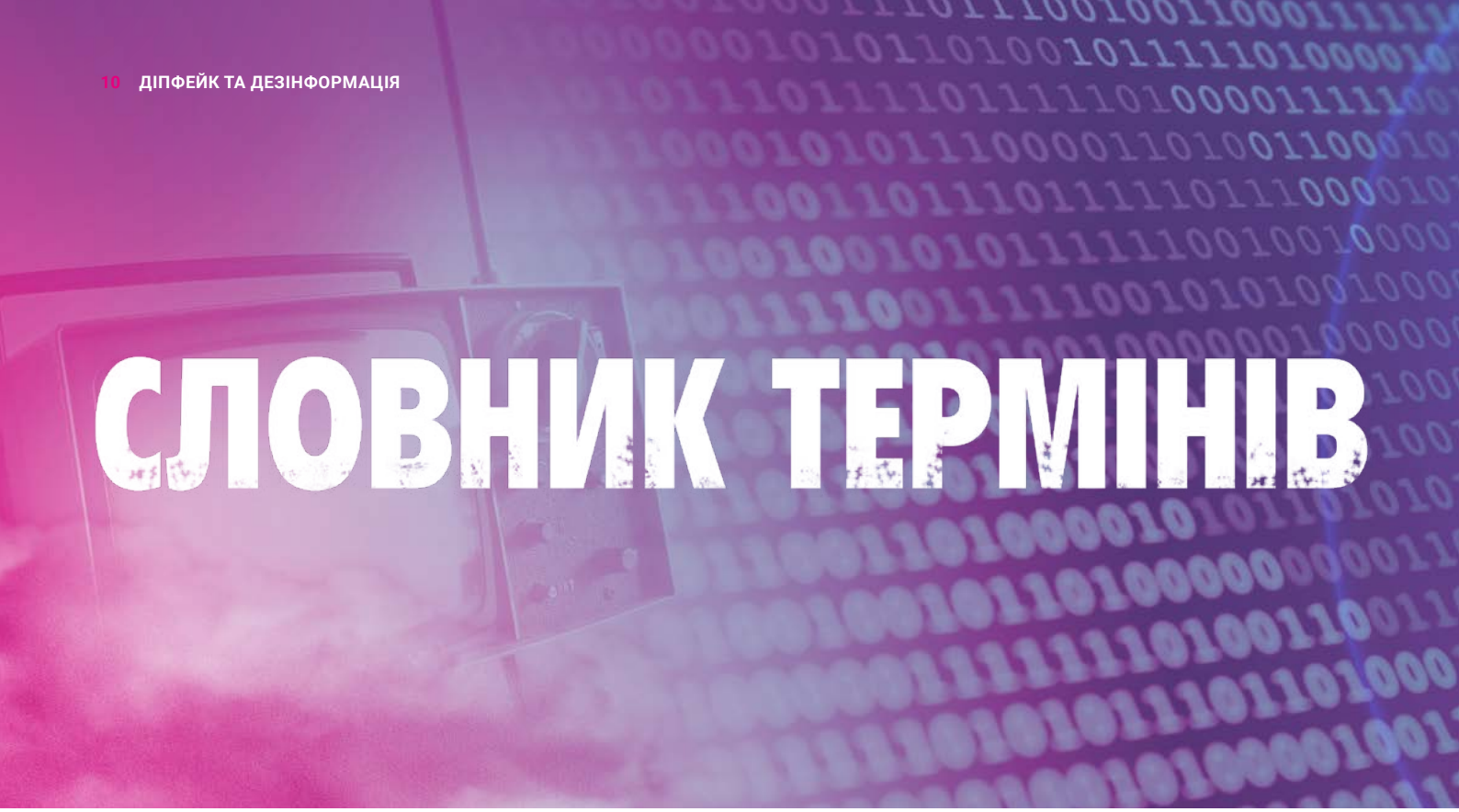
І хоча зараз Facebook та Twitter запровадили правила поводження з маніпульованим контентом, але, з одного боку, вони не є єдиними, а, з другого боку, визначення свободи вираження поглядів не бажано залишати за приватними компаніями.

Як чітко показав Малий запит фракції ВДП у Бундестазі ФРН від грудня 2019 року, федеральний уряд не готовий до теми «Використання маніпульованого контенту для дезінформації». Немає чіткої відповідальності за цю тему і також немає конкретного законодавства, поки що застосовуються лише «загально-абстрактні положення». У відповідях федерального уряду не простежуються ані відповідна стратегія, ані інвестиційний намір.

Загалом, спроби регулювання на німецькому та європейському рівнях здаються навряд чи придатними для стримування базованої на ШІ дезінформації. При цьому є й інший шлях. У деяких американських штатах діють закони, скеровані як проти непогодженої діпфейкової порнографії, так і проти використання цієї технології для впливу на виборців.

На цьому тлі законодавець мав би створити чіткі приписи щодо уніфікованого поводження цифрових платформ з діпфейками зокрема та дезінформацією в цілому. Заходи можуть сягати від маркування маніпульованого контенту через обмеження його розповсюдження (виключення з алгоритмів рекомендацій) до його видалення. Крім того, сприяння медіаграмотності має бути пріоритетом для всіх громадян, незалежно від віку. Важливо привернути увагу громадськості до існування діпфейків та сприяти підвищенню компетенції людей критично розглядати аудіовізуальний контент, навіть якщо стає все важче ідентифікувати це як підробку. У зв'язку з цим варто поглянути на країни Північної Європи, зокрема Фінляндію, населення якої є найбільш стійким до дезінформації.

Але ми не повинні робити одне: піддатися спокусі заборонити діпфейки в принципі. Як і будь-яка технологія, вона теж, опріч пов'язаних з нею небезпек, відкриває безліч цікавих можливостей, зокрема для освіти, кінематографу та сатири.



СЛОВНИК ТЕРМІНІВ

Artificial General Intelligence (Штучний загальний інтелект) / Сильний ШІ

Сильний ШІ або AGI – це концепція, в якій комп'ютерні системи опановують широкий спектр різних завдань і тим самим досягають людського рівня інтелекту. Зараз таких прикладних програм ШІ ще не існує. Це означає, що наразі жодна система не в змозі одночасно розпізнати рак, грати в шахи та керувати автомобілем, навіть якщо для всіх трьох проблем вже існують спеціальні системи. Нині декілька науково-дослідних інститутів та компаній працюють над сильним ШІ, але немає єдиної думки щодо того, чи можна це реалізувати, і якщо так, то коли.

Big Tech

Термін «*Big Tech*» використовується в медіа як збірний термін на позначення панівних компаній у галузі інформаційних технологій. Його часто використовують по черзі з термінами «GAFA» або «Big Four» для Google, Apple, Facebook та Amazon («GAFAM», якщо також враховувати Microsoft). Стосовно китайських Big Tech – компаній абревіатура BATX використовується для Baidu, Alibaba, Tencent і Xiaomi.

Cheap Fakes / Shallow Fakes

На відміну від діпфейків, Cheap Fakes (дешеві фейки) – це маніпуляції із зображеннями, аудіо чи відео, які створюються за допомогою відносно простих технологій. Прикладами цього можуть бути зниження швидкості відтворення аудіозаписів або показ контенту в зміненому контексті.

DARPA

Defense Advanced Research Projects Agency (Агентство передових оборонних дослідницьких проєктів США) – це агентство Міністерства оборони США, завданням якого є дослідження та фінансування інноваційних технологій у військовій галузі. Проєкти, що фінансуються DARPA, забезпечили важливі технології, що також використовуються у невійськовій сфері, як-от інтернет, машинний переклад або самокеровані (безпілотні) транспортні засоби.

Deepfake

Deepfakes (контамінація англ. *deep learning* – глибинне навчання – та *fake* – підробка) є продуктом двох алгоритмів ШІ, які взаємодіють у так званій Generative Adversarial Network (укр. *генеративна змагальна мережа*), скорочено GAN. GAN найкраще можна описати як спосіб алгоритмічного генерування нових типів даних із наявних наборів даних. Так, наприклад, GAN може проаналізувати тисячі записів Дональда Трампа, а потім створити нове зображення, що нагадує оцінені записи і водночас не є точною копією одного з цих записів. Ця технологія може бути застосована до різних типів контенту – зображення, рухомого зображення, звуку та тексту. Термін діпфейк, однак, використовується насамперед для аудіо- та відеоконтенту.

Deep Learning

Deep Learning (укр. глибинне навчання) – це підгалузь машинного навчання, де використовуються штучні нейронні мережі, які навчаються з великого обсягу даних. Подібно до того, як люди навчаються на досвіді, алгоритм глибинного навчання багаторазово виконує завдання, аби поступово покращувати результат. Ми говоримо про Deep Learning, тобто про «*глибинне навчання*», оскільки нейронні мережі мають кілька шарів, які уможливлюють навчання. Глибинне навчання дозволяє машинам розв'язувати складні проблеми, навіть якщо вони використовують різноманітні, неструктуровані набори даних.

Deep Porn

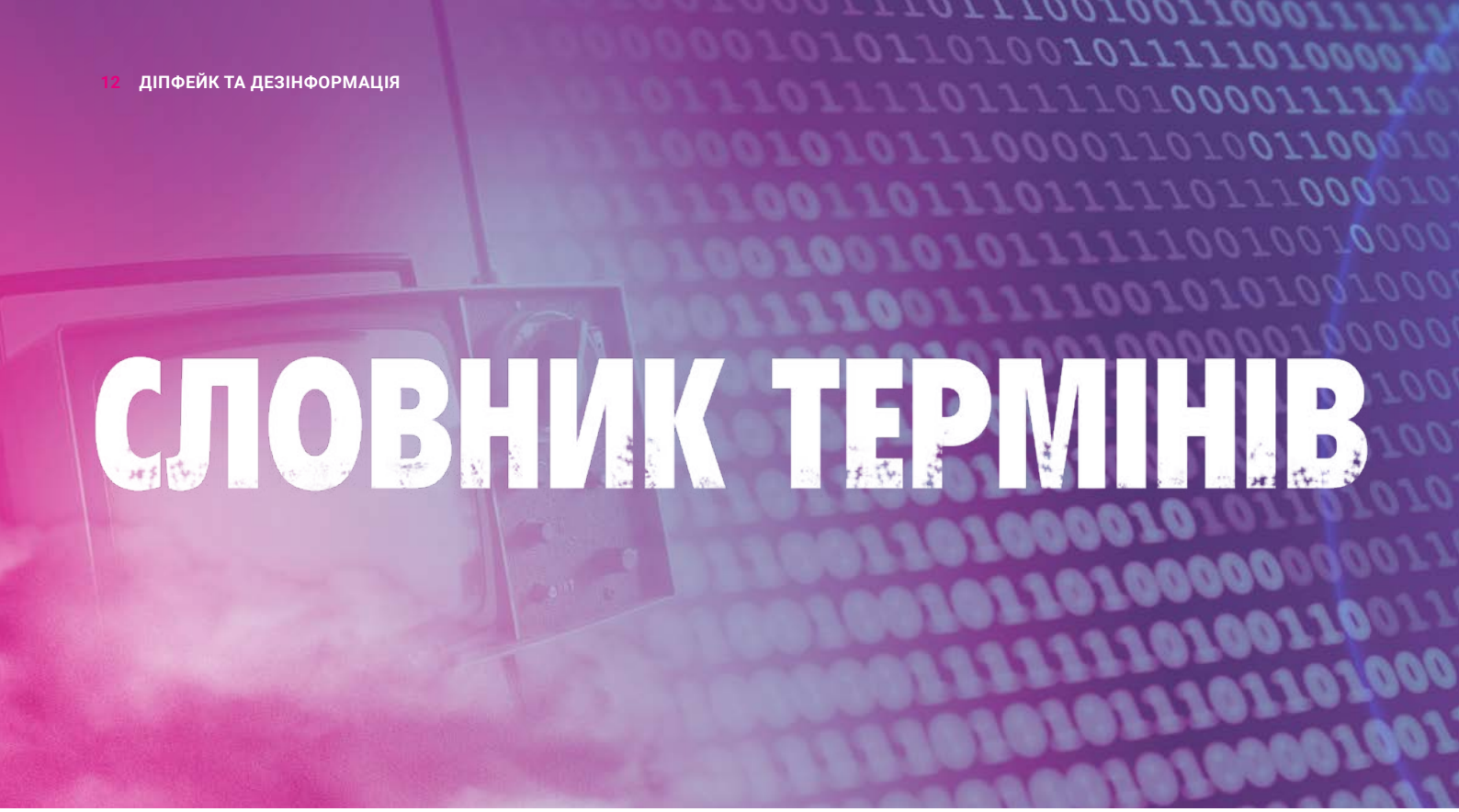
DeepPorn використовує методи глибинного навчання для генерування штучних порнографічних зображень.

Generative Adversarial Network

Генеративні змагальні мережі – це алгоритмічні архітектури, які використовують дві нейронні мережі – одну генеративну та одну дискримінаційну. Обидві конкурують одна з одною (генеративна мережа генерує дані, а дискримінаційна мережа фальсифікує їх) для генерування нових синтетичних наборів даних. Процес повторюється багаторазово для досягнення результатів, надзвичайно схожих на реальні дані. Мережі можуть працювати з різними типами даних, тому їх можна використовувати для генерування зображень, а також тексту, аудіо та відео.

GPT-2

GPT-2 – це структура, розроблена дослідницькою компанією OpenAI і заснована на штучній нейронній мережі, яка здатна автоматично генерувати англомовні тексти. Базою даних для GPT-2 служать близько 45 мільйонів сторінок тексту. На відміну від звичайних генераторів тексту, GPT-2 не складає тексти з готових текстових блоків і також не обмежується певним доменом. Вона може генерувати новий контент на основі будь-якого речення або фрагмента тексту.



СЛОВНИК ТЕРМІНІВ

IBM Watson

IBM Watson – це система на основі машинного навчання, розроблена IBM для того, щоб мати можливість розуміти поставлені природною мовою запитання та відповідати на них. IBM Watson привернув до себе багато уваги медіа, коли він у 2011 році у телевізійній грі – вікторині «Jeopardy» («Ризикуй!») переміг найкращих гравців-людей. Тим часом IBM Watson позиціонує себе як «ШІ для бізнесу» і складається з цілої палітри хмарних та інформаційних продуктів для різних галузей – від медицини до кіновиробництва.

Зима ШІ

Зима ШІ – це період зменшення інтересу та скорочення фінансування досліджень у галузі штучного інтелекту. Термін був введений у вжиток за аналогією з ідеєю ядерної зими. Технологічна сфера ШІ зазнала декількох ажіотажів з 50-х років XX століття, після чого йшли слідом розчарування, критика і скорочення фінансування.

Artificial Neural Networks

Штучні нейронні мережі (скорочено ШНМ) – це комп'ютерні системи, віддалено навіяні біологічними нейронними мережами, що перебувають в мозку людей і тварин. ШНМ «*вчаться*» виконувати завдання на основі прикладів, але не запрограмовані специфічними щодо завдань правилами. Вони, наприклад, можуть навчитися ідентифікувати зображення, що містять котів, аналізуючи приклади зображень, котрі були вручну позначені як «*кіт*» або «*не кіт*», і використовуючи результати для ідентифікації котів на інших зображеннях.

Машинне навчання

Машинне навчання – це, по суті, метод, який використовує алгоритми для того, щоб проаналізувати дані, навчитися з них, а потім зробити прогноз. Отже, замість того, щоб вручну програмувати програмне забезпечення точними інструкціями для виконання певного завдання, його тренують за допомогою великого обсягу даних та алгоритмів, які дають йому здатність навчитися, як виконувати завдання.

Microtargeting

Мікротаргетинг – це метод цифрового маркетингу, за допомогою якого намагаються максимально індивідуально використовувати рекламні меседжі для потенційних клієнтів. Для цього, залежно від платформи, враховуються демографічні характеристики, інтереси, історії веб-перегляду тощо. Залежно від цих критеріїв, один і той самий відправник може адресувати свої повідомлення до різних людей абсолютно по-різному. Цей метод спочатку був розроблений для використання в політичних кампаніях, але зараз використовується також у комерційних кампаніях.

Phishing

Фішинг – це метод кібератаки, в якому електронні листи використовуються як інструмент. Метою є змусити одержувача електронної пошти вважати, що повідомлення є автентичним та актуальним для нього (наприклад, повідомлення від його банку) і таким чином мотивувати його клікнути на посилання або завантажити додаток. Таким чином гакери можуть отримати доступ до конфіденційної інформації, наприклад паролів.

Revenge Porn / Порнографія помсти

Порнографія помсти стосується обміну інтимними сексуальними зображеннями на фотографії чи відео без згоди на те зображеної особи. Часто колишні партнери хочуть таким чином помститися після припинення відносин. Три чверті жертв порнографії помсти – жінки.

Слабкий ШІ або спеціалізований ШІ

Алгоритми слабкого ШІ спеціалізуються на виконанні дуже специфічних завдань, наприклад, розпізнавання облич, розуміння мови, гра в шахи. Навіть якщо вони в цьому зазвичай набагато кращі або ефективніші, ніж люди, вони можуть розв'язувати лише ці конкретні проблеми. Усі прикладні програми, які існують сьогодні, включно з такими, що видаються складними, як-от самокеровані автомобілі або мовні асистенти, належать до категорії слабких ШІ.

Social Engineering

Соціальною інженерією називають заходи, які призводять до цілеспрямованого впливу на людей, наприклад, щоб отримати доступ до конфіденційної інформації або домогтися розблокування коштів. Ця практика є також відомою як «*social hacking*», коли метою соціальної інженерії є отримання доступу до комп'ютерних систем відповідних осіб чи організації.

Суперінтелект

Суперінтелект – це гіпотетична концепція, в якій штучний інтелект перевершує не тільки найрозумніших людей, але й колективний інтелект людства.

Vishing

Вішинг (voice phishing – голосовий фішинг) – це метод фішингу, у якому замість електронної пошти як інструмент використовується телефонний дзвінок. Використання діпфейків для генерування голосу може підвищити ефективність цього методу.

1.0 СТАН РОЗВИТКУ

ШТУЧНИЙ ІНТЕЛЕКТ ТА ЙОГО РОЛЬ У ДЕЗІНФОРМАЦІЇ

Хоча коріння штучного інтелекту сягає середини XX століття, цій технології тривалий час приділялося мало уваги. Лише на початку XXI століття, схоже, намітився кінець тривалої зими ШІ: у 2011 році комп'ютерна система IBM Watson обіграла найкращих гравців-людей у телевізійній грі – вікторині «Jeopardy»[1)], прототип самокерованого автомобіля компанії Google подолав понад 100 000 миль (160 000 кілометрів), а Apple представила «*розумну персональну помічницю*» Сірі.

З того часу неухильно зростає суспільний інтерес до штучного інтелекту, насамперед до пов'язаних з ним ризиків. Дискурс про суперінтелект, викликаний однойменною книгою Ніка Бострома, яка була опублікована у 2014 році, ще більше загострив увагу.

Відтоді видатні особистості неодноразово висловлювалися на цю тему попереджаючи, а іноді й з тривогою. Часто цитують Стівена Гокінга («*Розвиток штучного інтелекту може означати кінець людства*») та Ілона Маска («*ШІ – це фундаментальний екзистенційний ризик для людської цивілізації*»).

У той час як суперінтелект і так званий «*сильний ШІ*» (AGI, Artificial General Intelligence) – це ще далеке майбутнє, «*слабкий ШІ*» зі своїми далеко не слабкими алгоритмами вже сьогодні грає все більшу роль у бізнесі, суспільстві та політиці. Авторка переконана, що вплив на здоров'я, енергію, безпеку, мобільність та багато інших напрямів буде значною мірою позитивним. Однак ми зможемо лише тоді скористатися з позитивної сторони розвитку подій, якщо усвідомимо ризики цієї технології та успішно їм протидіятимемо.

1) У цій вікторині учасники отримують загальну інформацію у вигляді відповідей, і вони мають сформулювати свої відповіді у формі запитань. До німецьких адаптивних версій належали Riskant від RTL та Der Große Preis від ZDF.

«Ми зможемо лише тоді скористатися з позитивної сторони розвитку подій, якщо усвідомимо ризики цієї технології та успішно їм протидіятимемо»

Одним із таких ризиків є зловживання технологією задля навмисного поширення неправдивої інформації. Політично мотивована дезінформація – це, звичайно, не нове явище. Сталін і Мао – найвидатніші імена серед тих диктаторів, за яких регулярно редагувалися фотографії, щоб старі зображення відповідали актуальній «*правді*»: фотографії тих, хто потрапляв у немилість, видалялися, а нові члени партійної верхівки додавалися пізніше; контекст зображень також змінювався, наприклад, через інше тло. Маніпульований візуальний запис мав на меті створити нові факти, переписати історії та історію.

У той час такі прилаштування були довготривалими і вимагали спеціальних знань, сьогодні за допомогою потрібного застосунку на смартфоні кожен може зробити це самостійно без будь-яких проблем. І на фотографіях технології не зупиняються. Створити підроблене відео, яке виглядало б правдивим, наразі все ще є досить клопіткою справою. Однак завдяки певним методам штучного інтелекту стає значно простіше маніпулювати наявними відео. Ці відео стали нині відомими як діпфейки. Вони все ще рідко трапляються в інтернеті, але зі збільшенням використання та розповсюдження становлять все більшу проблему для нашого суспільства.

Маніпульований контент не тільки з високою швидкістю поширюється на таких платформах, як Facebook або YouTube, він також цілеспрямовано демонструється охочим до перегляду адресатам. Крім того, поширення дезінформації все більше переходить на служби обміну миттєвими повідомленням, наприклад, WhatsApp. Там зашифровані повідомлення поширюються через приватні з'єднання, що збільшує довіру до переданої інформації. Це створює щось на кшталт прихованої віральності.

Шифрування приватного спілкування в інтернеті, подібно до таємниці листування, є бажаним надбанням: у такий спосіб повідомлення не можуть переглядатися сторонніми особами. Однак шифрування також означає, що поширена там інформація не може бути перевірена на її правдивість і тому не може відповідно модеруватися.

2.0 CHEAP FAKES & DEEPFAKES

ТЕХНОЛОГІЧНІ МОЖЛИВОСТІ МАНІПУЛЮВАННЯ З ТЕКСТОМ, ЗОБРАЖЕННЯМИ, АУДІО ТА ВІДЕО

В останні два роки термін діпфейк набуває все більшої популярності. Але чим є насправді діпфейки і чим вони відрізняються від іншого маніпульованого контенту?

Хоча перші експерименти з ШІ щодо маніпулювання відео відбувалися наприкінці 90-х років, широка громадськість дізналася про цю технічну можливість лише наприкінці 2017 року. На той момент з'явилася також термінологія, коли користувач Reddit на ім'я DeepFakes та інші члени спільноти Reddit *«r/deep-fakes»* публікували створений ними контент.

Не дивно, що у багатьох випадках це були порнографічні відеоролики, в яких обличчя актрис замінювалися обличчями знаменитостей, як-от Скарлетт Йоганссон чи Тейлор Свіфт. Більш невинними прикладами були сцени з кінофільмів, у яких усі обличчя актрис та акторів замінювало обличчя Ніколаса Кейджа.

«Не дивно, що у багатьох випадках це були порнографічні відеоролики, в яких обличчя актрис замінювалися обличчями знаменитостей, як-от Скарлетт Йоганссон чи Тейлор Свіфт»

ОСЬ ЯК ПРАЦЮЮТЬ ДІПФЕЙКИ

DeepFakes (контамінація *Deep Learning* та *Fake*, англ. – фальшивка) є продуктом двох алгоритмів ШІ, які взаємодіють у так званій Generative Adversarial Network (укр. генеративній змагальній мережі), скорочено GAN.

GAN можна найкраще описати як можливість алгоритмічного генерування нових типів даних з наявних наборів даних. Так, наприклад, GAN може проаналізувати тисячі знімків Дональда Трампа, а потім створити нове зображення, що нагадує розшифровані знімки, але не є точною копією одного з цих знімків. Ця технологія може бути застосована до різних типів контенту – зображення, рухомого зображення, звуку та тексту. Але позначення діпфейк застосовується насамперед до аудіо- та відеоконтенту.

Наразі для отримання правдивого результату потрібні навчальні дані лише кількох сотень фотографій чи аудіозаписів. Уже за якихось $3 кожен може замовити підроблене відео будь-якої людини за умови, що є щонайменше 250 фотографій цієї людини, утім це навряд чи буде проблемою для більшості тих, хто користується Instagram або Facebook. Синтетичні записи голосу також можна генерувати всього за 10 доларів за кожні 50 слів.

2.1 DeepFakes vs Cheap Fakes

Попри те, що маніпулювання порнографією, безумовно, є одним із найпоширеніших прикладів діпфейків, воно не є основною причиною нинішніх дискусій у суспільстві. Цікаво, що відео, яке викликало цю дискусію, взагалі було не діпфейком, а Cheap Fake (укр. дешевий фейк) (іноді його називають також Shallow Fake – укр. поверховий фейк): підроблений дуже простими технічними засобами відеоролик зі спікером Палати представників Конгресу США Ненсі Пелосі. Оригінальна швидкість запису була знижена приблизно до рівня 75 відсотків, а висота звуку була підвищена для підтримки природного тембру голосу.

Результат: у того, хто переглянув відео, могло скластися правдоподібне враження, що Ненсі Пелосі була п’яною. У соціальних мережах воно було поширено мільйони разів. Це показує, як навіть найпростіші підробки спотворюють реальність і можуть бути використані в політичних цілях. З усім тим поки що було дуже важко сфальсифікувати запис таким чином, щоб відповідна особа робила зовсім інші рухи або вимовляла зовсім інші слова, ніж в оригінальному відео. Поки що.

ПРИКЛАДИ ЗАСТОСУВАННЯ

Фото: https://youtu.be/mSaIrz8lM1U

2.2 Маніпулювання зразками руху

У 2018 році велику увагу привернула прикладна програма чотирьох дослідників Університету Берклі, яка використовувала штучний інтелект для перенесення танцювальних кроків людини-джерела (наприклад, професійної танцюристки) на цільову особу[2)].

На основі вихідного відео рухи переносяться на «*схематичну фігурку*». Наступний крок – нейронна мережа синтезує цільове відео відповідно до «*рухів схематичної фігурки*». Результатом є «*фейкове*» відео, в якому третя людина танцює як професіонал.

Звичайно, такий алгоритм можна використовувати не тільки для імітації танцювальних рухів, але й потенційно будь-якої іншої форми руху. Це відкриває шлях для зображення політичних опонентів у компрометувальних ситуаціях. Яким би був ефект, наприклад, відеозапису, де політичний діяч робить нацистське вітання або просто показує середній палець?

2) https://arxiv.org/pdf/1808.07371.pdf

ЩО ТАКЕ ШТУЧНІ НЕЙРОННІ МЕРЕЖІ?

Штучні нейронні мережі (укр. ШНМ, Artificial Neural Networks, скорочено ANN) – це комп'ютерні системи, віддалено навіяні біологічними нейронними мережами, що перебувають у мозку людей і тварин.

ШНМ *«вчаться»* виконувати завдання на основі прикладів, але не є запрограмованими специфічними щодо завдань правилами. Вони, наприклад, можуть навчитися ідентифікувати зображення, що містять котів, аналізуючи приклади зображень, котрі були вручну позначені як *«кіт»* або *«не кіт»*, і використовуючи результати для ідентифікації котів на інших зображеннях.

Голос і міміка

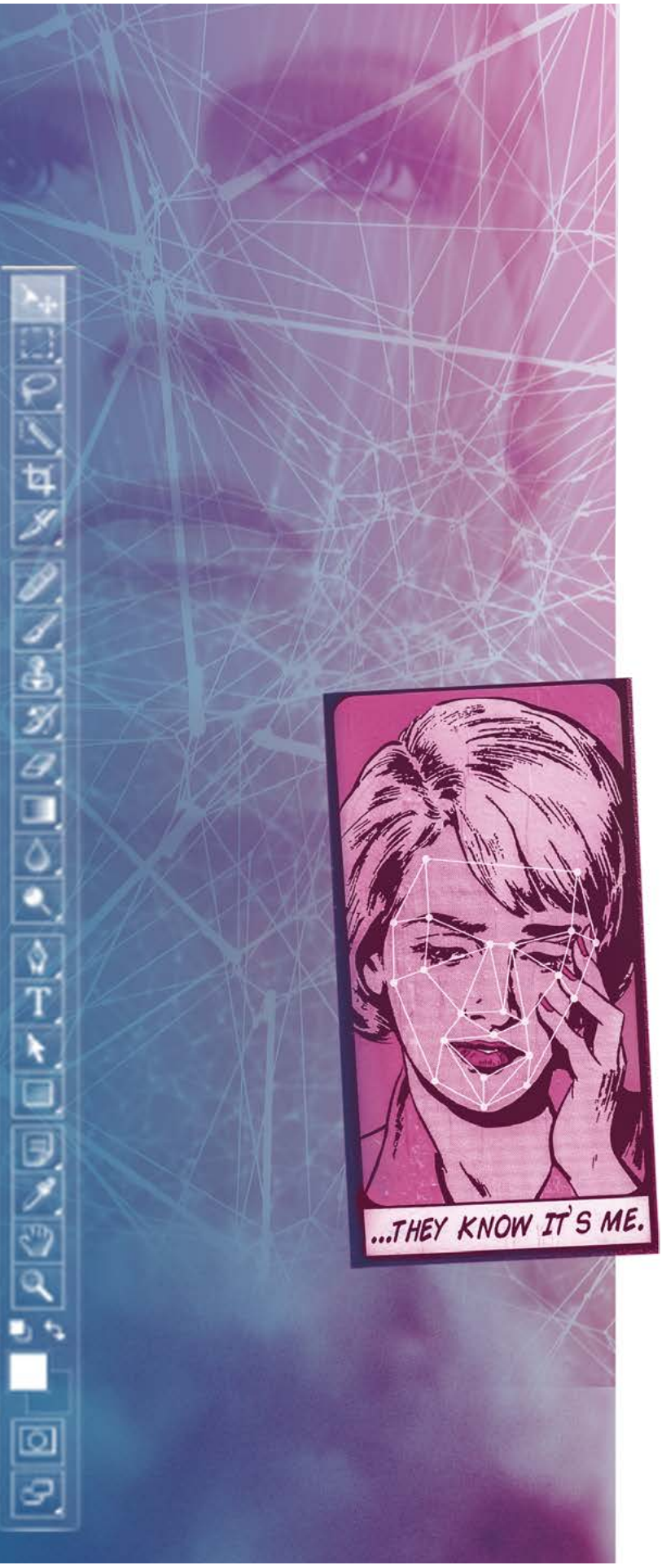
Підробки можуть мати ще більш далекосяжні наслідки, коли людям вкладають у вуста слова, яких вони ніколи не говорили, але в яких жести, міміка та голос здаються напрочуд справжніми. Були створені кілька таких відеороликів, зокрема з Бараком Обамою та Марком Цукербергом, однак не для того, щоб обдурити аудиторію, а щоб продемонструвати можливості технології та її небезпеку. Тим часом діпфейк був створений та розповсюджений також бельгійською політичною партією, що називається «Соціалістична партія – інші» (SP.A).

У травні 2018 року вона запостила у Facebook відео, в якому Трамп висміював Бельгію за те, що та залишається вірною Паризькій кліматичній угоді[3)].

Попри очевидно низьку якість і досить неприродний рух губ, що мало б негайно викликати підозру в уважного глядача, воно викликало сотні коментарів, у яких багато хто висловлював своє обурення тим, що американський президент наважився втрутитися в бельгійську політику щодо змін клімату. У випадку з цим відео автори також переймалися просвітою.

Відео було цілеспрямованою провокацією для привернення уваги людей до онлайн-петиції, яка закликає уряд Бельгії вжити термінових заходів на захист клімату. Але що було б, якби хтось зробив відео, в якому Трамп говорить не про бельгійську кліматичну політику, а, наприклад, що він планує ядерну атаку на Іран?

3) https://www.facebook.com/watch/?v=10155618434657151



Маніпулювання із зображенням: DeepNude та штучні обличчя

Контент, який часто не зараховується до діпфейків, хоча він генерується за дуже схожою технологією, – це контент зображення та тексту. Причина цього проста: як зображеннями, так і текстами можна без використання складної технології настільки легко маніпулювати, що «*додана вартість*» (або недолік, залежно від погляду) у порівнянні з аудіо- та відеоконтентом є невеликою. До того ж відеозаписи проти текстових та статичних зображень є набагато ефективнішими для того, щоб викликати такі емоції, як страх, лють чи ненависть.

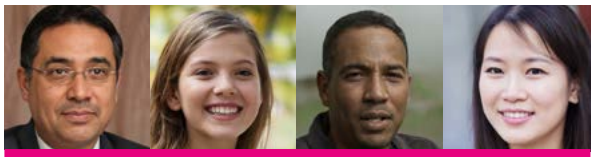
Однак деякі приклади маніпулювання таким контентом на основі ШІ привернули увагу. Як і у відео, так само і у випадку зображень алгоритми в основному використовуються для створення сфальшованого порнографічного контенту. Такі програми, як DeepNude, можуть за лічені секунди перетворити фотографію з бікіні в дуже реалістичне зображення оголеної людини.

Нікого не здивує, що застосунок працює лише щодо жінок (при спробі використати зображення чоловіка просто генеруються жіночі геніталії), роблячи таким чином кожну жінку потенційною жертвою «*порнопомсти*» (Revenge Porn), навіть якщо не існує жодного реального зображення її оголеною.

Зображення оголеної жінки, створене за допомогою програми deepnude.to

Зрештою нейронні мережі можуть не тільки використовуватися для маніпулювання із зображеннями таких людей, що існують, вони також «*створюють*» абсолютно нових людей – або принаймні абсолютно нові обличчя.

Комерційне застосування для цієї технології є очевидним: бази даних зображень можуть бути забезпечені за допомогою ШІ набагато рентабельніше, ніж з використанням реальних людей. Однак це також означає, що створення фальшивих профілів у соціальних мережах, які можна використовувати, наприклад, для поширення певного політичного контенту, значно полегшується.



Обличчя, що генеруються випадковим чином за допомогою thispersondoesnotexist.com

Існують також підозри щодо спроб шпигувати за допомогою генерованих на комп'ютері фото профілів, наприклад, у профілі *«Katie Jones»* на LinkedIn нібито дослідниці американського аналітичного центру.

Профіль на LinkedIn: генероване комп'ютером зображення профілю Katie Jones

Перш ніж експертний аналіз виявив декілька візуальних аномалій, які вказували на те, що зображення є синтетичним, профілеві вдалося з'єднатися з 52 політичними діячами у Вашингтоні, включаючи заступника помічника державного секретаря, високопоставленого радника одного із сенаторів та одного авторитетного економіста[4)].

LinkedIn швидко видалив обліковий запис, але він, очевидно, належить до мережі фантомних профілів, деякі з котрих все ще можуть існувати і використовуватися, наприклад, для фішинг-атак.

Тексти, генеровані ШІ

Описана програма може розвиватися особливо тоді, коли вона сполучена із засобами, які пропонують генерування тексту, керованого ШІ.

Багато хто чув про таку можливість у контексті створеного дослідницькою компанією OpenAI текстового генератора GPT-2, який через свій потенціал зловживань попервах вважався занадто небезпечним, щоб зробити його доступним для публіки[5)]. Пізніше компанія вирішила оприлюднити GPT-2 у кілька етапів, оскільки виробники поки що не змогли знайти явних доказів зловживання[6)].

Хоч це досі так і було, вони також визнають, що люди значною мірою вважають створені GPT-2 тексти правдивими, що генератор може бути чітко налаштованим на екстремістський контент – розпізнавання генерованих текстів буде становити проблему. За допомогою програми «*Talk To Transformer*» кожен може випробувати принцип дії GPT-2.

AI-generated fake content could unleash a virtual arms race of misinformation online, experts say.

"Once you get the person to click on something, you've gotten them to put themselves in a position to think a certain way, if they haven't already done so," said Katherine Jellison, a professor at Georgia Tech's School of Interactive Computing and author of the book "Cyberbullying in the Age of the Internet."

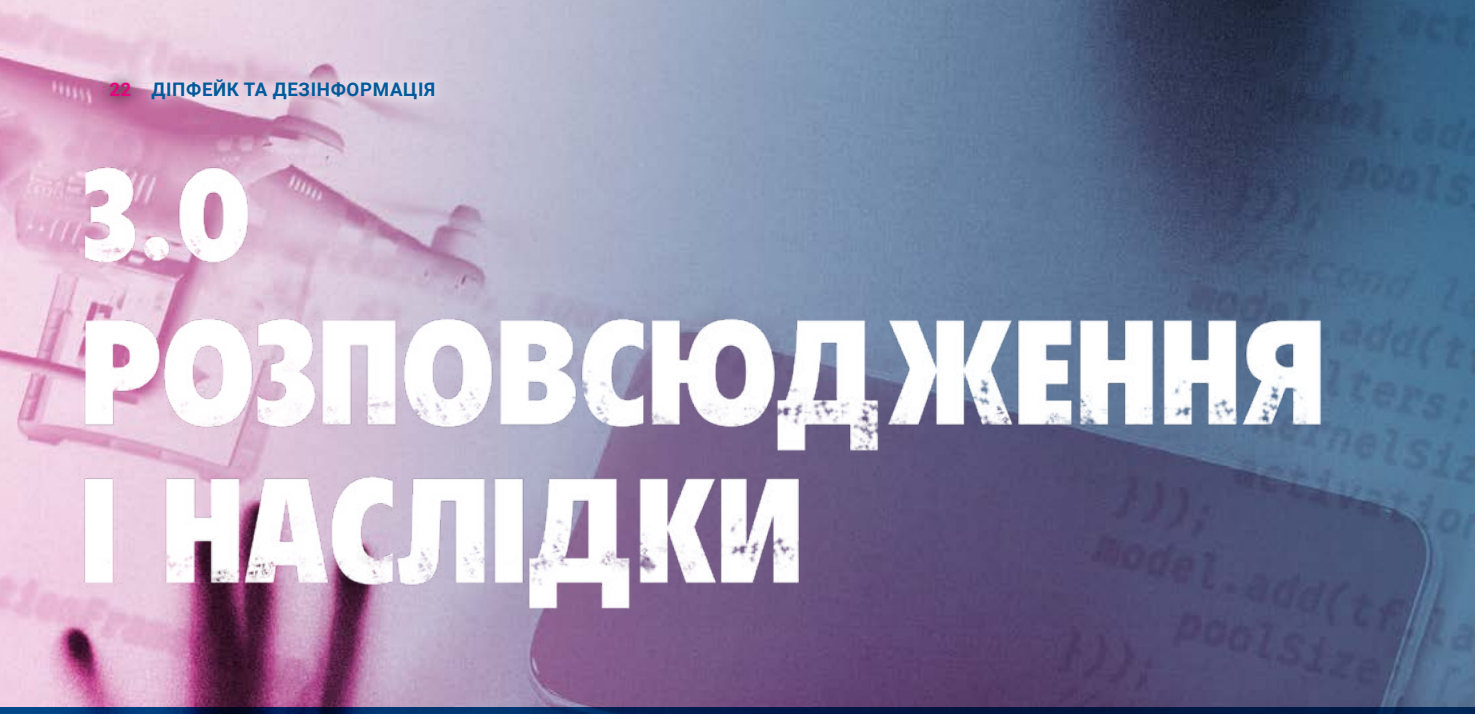
Виділений текст – за замовчуванням; решта тексту – генерований Ші за допомогою talktotransformer.com

Якщо в генератор ввести одне або більше речень, він генерує текст, який приймає дані введення як вихідну точку. Результати є часто – не завжди – напрочуд когерентними. Вони вгадують тон, що відповідає заданому за замовчуванням, та симулюють правдивість за допомогою вигаданих експертів, статистики та цитат.

4) https://www.cnet.com/news/spy-reportedly-used-ai-generated-photo-to-connect-with-targets-on-linkedin/

5) https://openai.com/blog/better-language-models/

6) https://openai.com/blog/gpt-2-1-5b-release/



3.0 РОЗПОВСЮДЖЕННЯ І НАСЛІДКИ

НАСКІЛЬКИ НЕБЕЗПЕЧНИМИ Є НАСПРАВДІ ДІПФЕЙКИ?

3.1 Розповсюдження

Точно кількісно оцінити розповсюдження діпфейків непросто, до того ж можна припустити, що їх кількість постійно зростає.

Компанія Deeptrace, яка пропонує технологічне рішення для виявлення діпфейків, спробувала дати точну оцінку у своєму звіті «The State of DeepFakes: Landscape, Threats and Impact»[7]. Відповідно до звіту, опублікованого у вересні 2019 року, кількість діпфейків за сім місяців майже подвоїлася із 7 964 у грудні 2018 року до 14 678 у липні 2019 року. 96 відсотків цих діпфейків – це непогоджений порнографічний контент, який показує лише жіноче тіло. У першу чергу вони скеровані на відомих жінок, чиї підроблені фотографії тисячами є доступними в інтернеті. Відповідно до звіту Deeptrace, лише чотири найпопулярніші веб-сайти DeepPorn мають тимчасом понад 134 мільйони переглядів підроблених відеороликів про жінок-знаменитостей.

Але багато приватних осіб також постраждали від уже згаданої порнографії помсти. Збільшення стає можливим насамперед завдяки кращій доступності як інструментів, так і послуг, що дозволяють створювати діпфейки навіть без навичок програмування.

У 2019 році також уже повідомлялося про випадки, коли генеровані ШІ мовні клони використовувались для соціальної інженерії. У серпні The Wall Street Journal[8] повідомив про перший випадок голосового шахрайства на основі ШІ, також відомого як вішинг (скорочення від voice phishing), яке німецькій компанії-жертві коштувало 220 000 євро.

Програмне забезпечення настільки успішно імітувало голос німецького менеджера, включаючи мелодику та легкий німецький акцент, що його британський колега негайно виконав нагальне прохання додзвонювача переказати названу суму. Поки це був поодинокий випадок, але можна припустити, що подібні спроби в майбутньому будуть частішими.

Значна частина медійного висвітлення діпфейків зосереджена на їхньому потенціалі для дискредитації політичних опонентів та підриву демократичних процесів.

Поки що цей потенціал не розвинувся.

Хоч відео таких політиків, як Барак Обама, Дональд Трамп чи Маттео Ренці, технічно маніпулювалися, але поки що це робилося в основному з метою сатири чи демонстрації, і все швидко вияснялося.

«Хоч відео таких політиків, як Барак Обама, Дональд Трамп чи Маттео Ренці, технічно маніпулювалися, але поки що це робилося в основному з метою сатири чи демонстрації, і все швидко вияснялося»

3.2 Наслідки

Водночас той факт, що досі політики ніколи не використовували діпфейки для дезінформації, аж ніяк не означає, що вони не мали впливу на політичний дискурс. Приклад, якому західні медіа приділили мало уваги, показує, як саме знання про існування діпфейків може впливати на політичний клімат.

Президент Габону Алі Бонго після інсульту місяцями не з'являвся на публіці. Ясна річ, ширилися різні чутки, і вже лунали голоси, що президент помер. Для того, щоб покласти край спекуляціям, у грудні 2018 року було оприлюднено відео, в якому він виголосив свою звичайну новорічну промову. Однак запис мав протилежний ефект. Багато хто вважав, що Бонго виглядав дивно, і відразу виникли підозри, що відео підроблене. Незабаром після цього військові розпочали невдалий державний переворот і назвали цей нібито діпфейк як частину мотивації[9].

Однак наступний криміналістичний аналіз підтвердив справжність запису. Алі Бонго тимчасом одужав після інсульту і далі перебуває на посаді.

Це свідчить про те, що найбільшою загрозою з боку діпфейків не конче мають бути самі діпфейки. Уже сама технологічна можливість створення таких відеозаписів викликає питання: чи можна довіряти автентичності рухомих зображень?

Це питання кидає тінь на президентські вибори у США 2020 року. Уже у виборчій кампанії 2016 року відігравали роль підтримувані ШІ дезінформація та маніпуляції, передусім, у формі мікротаргетингу та ботів. Діпфейки є лише ще одним інструментом з арсеналу дезінформації. Навіть якщо в передвиборчій кампанії діпфейки не будуть використані зовсім або використані лише декілька з них, то політики швидше за все з удячністю скористаються можливістю дистанціюватися від реальних, але невигідних для них записів як нібито таких.

7) The State of DeepFakes: Landscape, Threats, and Impact: Henry Ajder, Giorgio Patrini, Francesco Cavalli and Laurence Cullen, – September 2019. (англ. Держава діпфейків: ландшафт, загрози та вплив. – Прим. перекл.).

8) https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

9) https://www.technologyreview.com/s/614526/the-biggest-threat-of-Deepfakes-isnt-the-Deepfakes-themselves/

3.3 Чи є також позитивні приклади використання діпфейків?

«Технологія дає нам [...] способи заподіяти шкоду і чинити добро; вона збільшує і те, і інше. [...] Але те, що ми кожного разу маємо новий вибір – це нове благо»[10], – каже Кевін Келлі, багатолітній головний редактор і член групи засновників технологічного журналу «Wired». Чи може сказане стосуватися і діпфейків?

Ця технологія є особливо цікавою для кіноіндустрії, насамперед для постпродукції та дубляжу. Чому? Наразі кіностудії мають докладати чимало зусиль, щоб згодом адаптувати діалог.

У такому випадку доводиться ще раз проводити по бухгалтерії залучених акторів, необхідний персонал та місце розташування. Завдяки технології, що лежить в основі діпфейків, такі зміни можна робити у найкоротший час і зі значно меншими витратами. Дублювання фільмів також можна значно покращити.

Можна було б адаптувати рухи губ акторів-виконавців до слів акторів дубляжу або безпосередньо синтезувати голоси та адаптувати їх до відповідної мови так, що дубляж буде не потрібним.

10) Цитата (у перекладі) за посиланням https://www.edge.org/conversation/kevin_kelly-the-technium/

Прикладом такого застосування є відео Девіда Бекгема, який рекламує кампанію проти малярії[11]. У ньому він *«розмовляє»* кількома мовами і щоразу його губи здаються ідеально синхронізованими зі словами.

Освіта також є цікавою областю застосування. Так, наприклад, можна створити відеоролики з історичними діячами, які розповідають свою історію або відповідають на запитання. Проєкт *«Dimensions of History»*[12] (укр. Виміри історії) Фонду «Шоа» Університету Південної Каліфорнії викликав значний резонанс у медіа, в рамках якого було проведено інтерв'ю з 15 особами, котрі пережили Голокост, та зроблено з них голографічні знімки. Пересувну виставку можна було побачити в різних музеях США та нещодавно також у Шведському історичному музеї. Після перегляду відвідувачі виставки мали можливість поставити свої запитання голограмам. Програмне забезпечення для розпізнавання мови співвідносило це запитання з фрагментом інтерв'ю. З використанням технології діпфейків те саме можна було б зробити в більших масштабах та кількома мовами.

«Ця технологія є особливо цікавою для кіноіндустрії, насамперед, для постпродукції та дубляжу»

11) https://www.malariamustdie.com/

12) https://sfi.usc.edu/dit

4.0 БОРОТЬБА З ДІПФЕЙКАМИ

ЩО МИ МОЖЕМО ПРОТИСТАВИТИ ПОВ'ЯЗАНИМ З ДІПФЕЙКАМИ ВИКЛИКАМ?

Звичайно, ці позитивні приклади не повинні применшувати потенційну небезпеку, яку несуть у собі діпфейки. Ці небезпеки є безперечними і вимагають рішучих контрзаходів – щодо цього існує широка згода. Менше згоди існує щодо того, як саме мають виглядати ці контрзаходи. До того ж виникає питання, як гарантувати право індивіда на свободу висловлення думки, не ставлячи одночасно під загрозу потреби суспільства у надійній інформаційній системі.

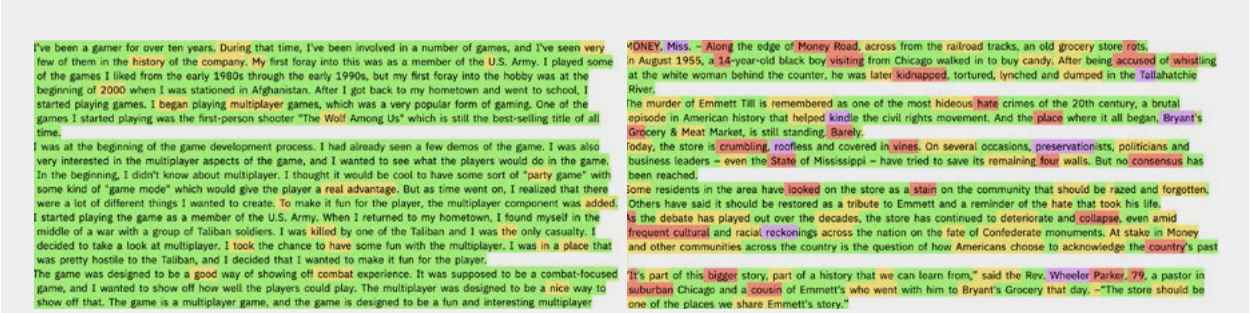
4.1 Технічні рішення для виявлення діпфейків та боротьби з ними

Одним зі способів боротьби з підробкою є розробка технологій, які можуть відрізняти підробку від реального контенту. Для цього використовуються алгоритми, подібні до тих, які були розроблені для створення омани. Використовуючи GLTR – модель, засновану на вищезгаданому генераторі тексту, – дослідники з MIT-IBM Watson AI і HarvardNLP проаналізували, чи може та сама технологія, яка пише самостійно винайдені статті, також використовуватися для розпізнавання генерованих ШІ уривків. Якщо ввести текстовий уривок у тестову програму, вона виділить слова зеленим, жовтим, червоним або фіолетовим кольорами, щоб вказати на зменшення передбачуваності.

Чим вищою є частка слів з низькою передбачуваністю, тобто позначених червоним та фіолетовим кольорами в реченнях, тим більшою є ймовірність того, що уривок – це текст автора-людини; натомість чим більш передбачуваними є слова (і «*зеленішими*» уривки), тим більшою є ймовірність, що ми маємо справу з текстовим генератором.

Подібні методи можна використовувати для розкриття маніпульованих відео. У 2018 році дослідники виявили, що очі на обличчях у відеороликах із діпфейків не блимають. Це було пов'язано з тим, що для генерування знімків використовувалися статичні зображення, а вони здебільшого показували людей з розплющеними очима. Але користь від цього відкриття була нетривалою. Щойно ця інформація стала загальнодоступною, з'явилися перші відеоролики з очима, що блимали. У майбутньому ситуація буде схожою з іншими механізмами виявлення. В одну і ту ж саму гру – кота і мишу – в галузі кібербезпеки грають десятиліттями, прогрес завжди є корисним для обох сторін.

Однак цей факт, безумовно, не є підставою зупиняти зусилля щодо виявлення діпфейків. У вересні 2019 року Facebook у співпраці з ініціативою PAI[13], Microsoft та декількома університетами оголосив конкурс «*DeepFake Detection Challenge*»[14] на суму 10 мільйонів доларів.


I've been a gamer for over ten years. During that time, I've been involved in a number of games, and I've seen very few of them in the history of the company. My first foray into this was as a member of the U.S. Army. I played some of the games I liked from the early 1980s through the early 1990s, but my first foray into the hobby was at the beginning of 2000 when I was stationed in Afghanistan. After I got back to my hometown and went to school, I started playing games. I began playing multiplayer games, which was a very popular form of gaming. One of the games I started playing was the first-person shooter "The Wolf Among Us" which is still the best-selling title of all time.

I was at the beginning of the game development process. I had already seen a few demos of the game. I was also very interested in the multiplayer aspects of the game, and I wanted to see what the players would do in the game. In the beginning, I didn't know about multiplayer. I thought it would be cool to have some sort of "party game" with some kind of "game mode" which would give the player a real advantage. But as time went on, I realized that there were a lot of different things I wanted to create. To make it fun for the player, the multiplayer component was added. I started playing the game as a member of the U.S. Army. When I returned to my hometown, I found myself in the middle of a war with a group of Taliban soldiers. I was killed by one of the Taliban and I was the only casualty. I decided to take a look at multiplayer. I took the chance to have some fun with the multiplayer. I was in a place that was pretty hostile to the Taliban, and I decided that I wanted to make it fun for the player.

The game was designed to be a good way of showing off combat experience. It was supposed to be a combat-focused game, and I wanted to show off how well the players could play. The multiplayer was designed to be a nice way to show off that. The game is a multiplayer game, and the game is designed to be a fun and interesting multiplayer

MONEY, Miss. – Along the edge of Money Road, across from the railroad tracks, an old grocery store rots.

In August 1955, a 14-year-old black boy visiting from Chicago walked in to buy candy. After being accused of whistling at the white woman behind the counter, he was later kidnapped, tortured, lynched and dumped in the Tallahatchie River.

The murder of Emmett Till is remembered as one of the most hideous hate crimes of the 20th century, a brutal episode in American history that helped kindle the civil rights movement. And the place where it all began, Bryant's Grocery & Meat Market, is still standing. Barely.

Today, the store is crumbling, roofless and covered in vines. On several occasions, preservationists, politicians and business leaders – even the State of Mississippi – have tried to save its remaining four walls. But no consensus has been reached.

Some residents in the area have looked on the store as a stain on the community that should be razed and forgotten. Others have said it should be restored as a tribute to Emmett and a reminder of the hate that took his life.

As the debate has played out over the decades, the store has continued to deteriorate and collapse, even amid frequent cultural and racial reckonings across the nation on the fate of Confederate monuments. At stake in Money and other communities across the country is the question of how Americans choose to acknowledge the country's past

"It's part of this bigger story, part of a history that we can learn from," said the Rev. Wheeler Parker, 79, a pastor in suburban Chicago and a cousin of Emmett's who went with him to Bryant's Grocery that day. –"The store should be one of the places we share Emmett's story."

Результат аналізу: автор-людина vs текстовий генератор, джерело: gltr.io

Facebook також доручив створити набір даних із зображеннями та відеозаписами запрошених з цією метою акторів, щоб створити достатню базу даних для челенджу. За кілька тижнів Google також з тією ж метою опублікував набір даних з 3000 маніпульованих відео. Американське агентство фінансування досліджень DARPA при Пентагоні також в рамках програми MediFor (скорочено від Media Forensics) з 2016 року працює над тим, щоб виявити маніпульований контент, і протягом двох років інвестувало в це 68 мільйонів доларів[15].Над якими технічними рішеннями для боротьби з діпфейками працюють у Німеччині та Європі і чи працюють взагалі, про це мало що відомо.

Здебільшого тут ідеться про окремі компанії, такі як уже згадана вище Deeptrace, та дослідницькі проєкти, як-от Face2Face[16], професора Мюнхенського технічного університету Маттіаса Ніснера.

Згідно з відповіддю федерального уряду на Малий запит фракції ВДП цією темою займаються, зокрема, Національний науково-дослідний центр прикладної кібербезпеки (CRISP/ ATHE-NE), а також Мюнхенський технічний університет або Інститут імені Фраунгофера.

Крім того, німецька державна радіостанція та телеканал, що мовить на закордоння, «Німецька хвиля» (Deutsche Welle, DW), Інститут цифрових медіатехнологій імені Фраунгофера (IDMT) та Афінський технологічний центр (ATC) розпочали спільний дослідницький проєкт «*Digger*». Мета проєкту – розширити веб-платформу перевірки «Truly Media» від DW та ATC завдяки аудіокриміналістичним технологіям від Fraunhofer IDMT й у такий спосіб допомогти журналістам[17]. Однак тут не простежуються ні конкретна стратегія, ні інвестиційний намір з боку федерального уряду.

13) The Partnership on AI (PAI) – це організація, яка об'єднує університети, дослідників, громадські організації та компанії з метою кращого розуміння ефекту ШІ та його впливу на суспільство: www.partnershiponai.org

14) https://ai.facebook.com/blog/deepfake-detection-challenge/

15) https://futurism.com/darpa-68-million-technology-Deepfakes

16) https://niessnerlab.org/projects/thies2016face.html

17) https://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf

«Чим вищою є частка слів з низькою передбачуваністю, тобто позначених червоним та фіолетовим кольорами в реченні, тим більшою є ймовірність того, що уривок – це текст автора-людини; натомість чим більш передбачуваними є слова (і «зеленішими» уривки), тим більшою є ймовірність, що ми маємо справу з текстовим генератором»

4.2 Спроби саморегуляції соціальних мереж

У той час, коли технологічні гіганти хочуть використовувати дані та фінансові засоби, щоб сприяти технологічному розв'язанню проблеми, лунає все більше голосів, які вимагають також подальших кроків від Facebook та Co., оскільки саме їхні платформи сприяють поширенню дезінформації. На цьому тлі Twitter і Facebook наприкінці 2019 року та на початку 2020 року відповідно прокоментували свої плани щодо поводження з діпфейками. У листопаді 2019 року Twitter попросив своїх користувачів про зворотний зв'язок щодо «*Пропозиції стосовно правил для синтетичних та маніпульованих медіа*». Відповідні правила були оголошені на початку лютого 2020 року:

«Будь-яке фото, аудіо чи відео, яке було *"значно змінено або підроблено"* для введення в оману людей, буде видалено, якщо вважатиметься, що це може завдати серйозної шкоди – наприклад, якщо це загрожуватиме фізичній безпеці людей або спричинятиме *"масові заворушення серед громадян"*. В іншому випадку Twitter все ж може позначати твіти як маніпульовані медіа, висловлювати попередження при спробі поширити контент та виключати контент з-поміж пріоритетних у каналах користувачів». Зміни набули чинності 5 березня 2020 року[18].

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

Twitter: поводження із синтетичними та маніпульованими медіа: https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-andmanipulated-media.html

«Над якими технічними рішеннями для боротьби з діпфейками працюють у Німеччині та Європі і чи працюють взагалі, про це мало що відомо»

Facebook йде на крок далі. 6 січня 2020 року Моніка Бікерт, віцепрезидент Facebook з управління глобальною політикою, повідомила в дописі у своєму блозі, що в майбутньому діпфейки, які відповідають певним критеріям, мають бути видалені з платформи[19]. Відповідно має бути видалений контент, який був оброблений або синтезований за допомогою ШІ таким чином, щоб він здавався автентичним для звичайної людини. Однак з цього припису слід виключити контент, у якому йдеться про сатиру, що залишає значну можливість для інтерпретації.

Цікаво, що цей припис стосується не Cheap Fakes, а конкретно лише контенту, створеного ШІ. Відповідно, згадане вище підроблене відео Ненсі Пелосі все ще є доступним у Facebook[20]. Хоча Facebook визнав, що його фактчекери класифікували відео як неправдиве, але відмовився видалити відео, оскільки компанія «*не має припису, який би вимагав, щоб розміщена у Facebook інформація була правдивою*»[21].

Цей підхід також відповідає розумінню мережею Facebook свободи вираження думки і виходить за рамки теми діпфейка. У контексті дебатів про політичну рекламу Роб Лезерн (Rob Leathern), директор з управління продуктом у Facebook, у січні 2020 року написав у своєму блог-пості, що такі рішення не повинні приймати приватні компанії, «*саме тому ми виступаємо за таке регулювання, яке б стосувалося всієї галузі. За відсутності регулювання Facebook та інші компанії можуть вільно формувати свою власну політику*».

Звичайно, можна дискутувати про те, чи тлумачення мережею Facebook свободи вираження є правильним з етичної точки зору. Однак висловлювання Роба Лезерна привертає увагу до важливої теми – відсутності або принаймні неповного регулювання.

18) https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

19) https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

20) Знову ж таки YouTube – інша платформа, яка своїми алгоритмами рекомендацій сприяє віральності неправдивої інформації, хоча й видалила згадане відео, але відмовляється дати чітке формулювання стосовно поводження з діпфейками у майбутньому.

21) https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413

4.3 Способи законодавчого регулювання

Згідно з відповіддю федерального уряду на вищезгаданий Малий запит від фракції ВДП у Німеччині до діпфейків застосовуються «*загально-абстрактні правила*». «*На федеральному рівні не існує спеціальних норм, які б регулювали лише діпфейк-програми або були створені для них. Федеральний уряд постійно перевіряє законодавчу базу на федеральному рівні, щоб визначити, чи є потреба в адаптації через технологічні чи суспільні проблеми*».

Це означає, що деякі аспекти проблем діпфейків, наприклад, порнографія помсти, нібито імпліцитно регулюються чинними законами, але немає експліцитного поводження з маніпульованим контентом. Це стосується не лише спеціальної теми «*Діпфейки*», а й усього спектру дезінформації в цифровому просторі. Як показує автор дослідження «*Регуляторні види реакції на дезінформацію*»[22] Фонду «Нова відповідальність» (Stiftung Neue Verantwortung), «*попередні спроби регулювання та політичні підходи до рішень [у Німеччині та Європі] є навряд чи придатними для стримування дезінформації*».

Детальний аналіз стану регулювання діпфейків у США показаний у дослідженні юридичної компанії WilmerHale «DeepFake Legislation: A Nationwide Survey»[23]. У Сполучених Штатах експліцитні закони про діпфейки вже включені до кримінального права, наприклад, у Вірджинії, яка криміналізує непогоджену діпфейкову порнографію, або в Техасі, де караються діпфейки, які мають на меті впливати на виборців. Подібні закони у вересні 2019 року були прийняті також у Каліфорнії.

Але, мабуть, найбільш масштабне регулювання діпфейків наприкінці 2019 року здійснив китайський законодавець. Китайські закони вимагають, щоб провайдери та користувачі аудіоінформаційних служб та відеоновин в інтернеті чітко позначали весь контент, створений або модифікований за допомогою нових технологій, таких як штучний інтелект.

І хоча варто подумати про те, чи не слід було б подібне регулювання перейняти й іншим країнам, саме у випадку Китаю воно залишає по собі поганий посмак: китайський уряд, зі свого боку, використовує дезінформацію на основі технологій для вжиття заходів проти протестувальників у Гонконзі, і слід вважати, що це нове регулювання буде використане як привід для подальших зусиль щодо цензури.

Ефективне регулювання нових технологічних явищ є, безумовно, непростим. У минулому з цим також постійно виникали труднощі. Керування автомобілем в Англії XIX століття, наприклад, згідно з «Locomotive Act» 1865 року вимагало, щоб друга людина йшла перед транспортним засобом і махала червоним прапором.[24]
І все ж таки є заходи, які законодавці вже тепер можуть вжити для протидії такому явищу, як діпфейк. Оскільки 96 відсотків діпфейків зараз є непогодженою порнографією, було б непогано для початку запровадити їх експліцитне покарання, як у Вірджинії чи Каліфорнії. Регулювання щодо наклепу, шахрайства та особистих прав має йти в тому ж напрямі. Крім того, законодавці мають створити чіткі приписи щодо уніфікованого поводження цифрових платформ з діпфейками зокрема та дезінформацією в цілому.

Ці заходи можуть варіюватися від маркування через обмеження розповсюдження (виключення з алгоритмів рекомендацій) аж до видалення діпфейків. Крім того, сприяння медіаграмотності має бути пріоритетом для всіх громадян, незалежно від віку. Адекватна інформація про те, як створюються та розповсюджуються діпфейки, має давати можливість громадянам розпізнавати дезінформацію як таку і не дати себе ввести в оману.

22) https://www.stiftung-nv.de/sites/default/files/regulatorische_reaktionen_auf_desinformation.pdf

23) Matthew Ferraro, WilmerHale | Deepfake Legislation: A Nationwide Survey – State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media.

24) https://sites.google.com/site/motormiscellany/motoring/law-and-the-motorist/locomotive-act-1865/

4.4 Індивідуальна відповідальність: критичне мислення та медіаграмотність

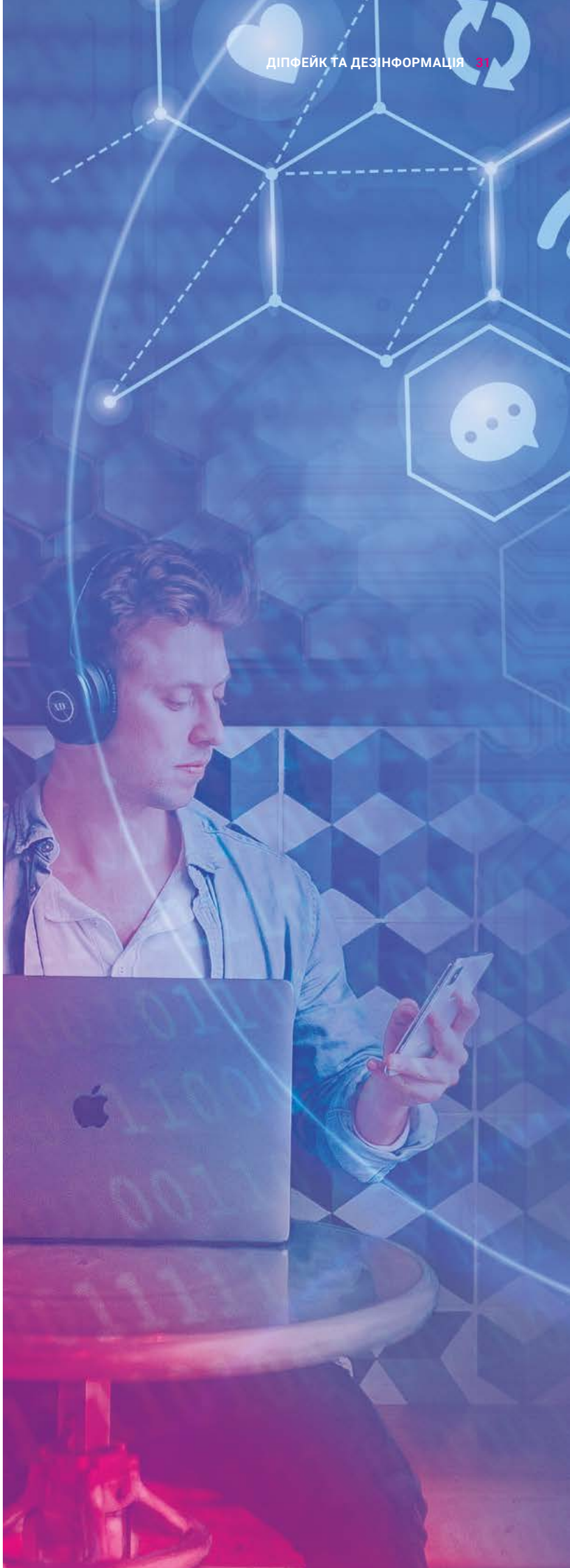
Критичне мислення та медіаграмотність є основою для диференційованого поводження з дезінформацією. Було б, звичайно, неможливо і, мабуть, не бажано вимагати від кожної окремої людини ставити під сумнів усе побачене.

І все ж сьогодні, як ніколи, людям рекомендується дуже обережно насолоджуватися тим, що вони споживають в інтернеті. Найпростіше, що кожен може зробити, якщо фото, відео чи також текст здається дивним – це пошукати в Google: багато підробленого контенту в такий спосіб швидко розкривається, деталі підробок циркулюють так само швидко.

Цей крок є особливо важливим, якщо ви маєте намір поділитися цим контентом, позначити його як «подобається» або прокоментувати. Крім того, ми можемо більше зважати на те, чи блимання, вираз обличчя або мова справляють враження неприродних, чи є частини фото розмитими, чи здаються об'єкти недоречними, навіть якщо ці особливості з прогресом технології діпфейка все більше зникатимуть.

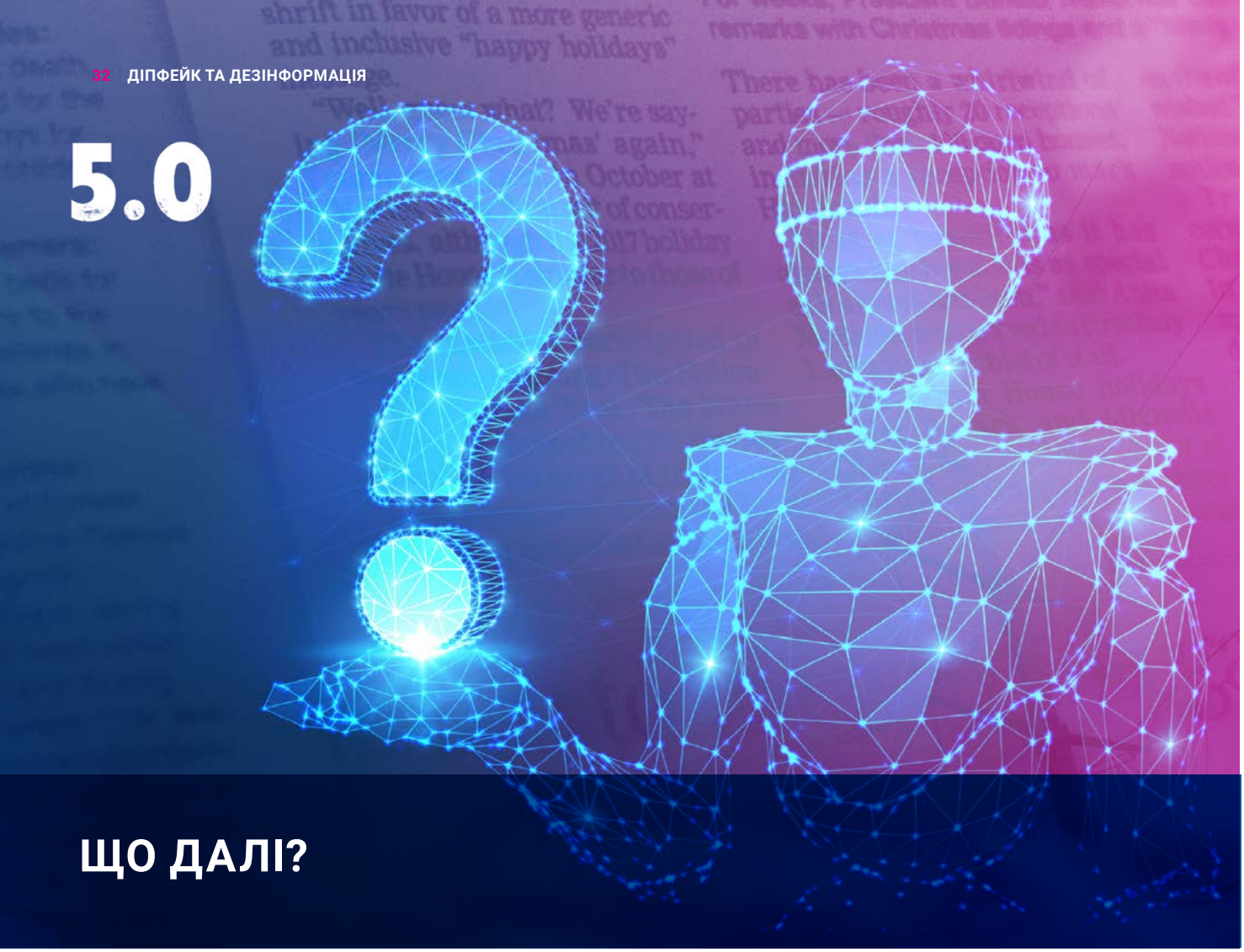
У майбутньому, можливо, з'являться додатки браузера, які подібно до блокатора реклами ідентифікуватимуть автоматизовано маніпульований контент і звертатимуть на нього увагу користувачів. Однак ці кроки передбачають, що ми взагалі усвідомлюємо можливості маніпуляцій.

Щоб досягти цього усвідомлення серед своїх громадян, Фінляндія – країна, яка посідає перше місце у дослідженні щодо вимірювання стійкості[25)] до дезінформації, робить ставку на освітні пропозиції для всього населення – від дитячого садка до пенсійного віку.



25) https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019_-ENG.pdf

5.0



ЩО ДАЛІ?

Поки що неможливо в деталях передбачити, наскільки сильним буде конкретний вплив діпфейків на політику та суспільство. Однак це не має бути причиною бездіяльності. Як уже не раз наголошувалося, ні підроблені відео, ні дезінформація як такі не є новим явищем – новим є все більша простота їх створення, підвищення їхньої якості та можливості їх розповсюдження.

Добрим лакмусовим папірцем, безумовно, стануть президентські вибори у США, які відбудуться восени 2020 року. Утім не слід рекомендувати просто «*терпляче вичікувати*».

Дослідники, технологічні компанії, журналісти, уряди та самі користувачі мають докласти всіх зусиль для нейтралізації негативних наслідків підробок.

На першому етапі необхідні чітке регулювання та послідовна боротьба з діпфейк-порнографією, оскільки вона вже широко розповсюджена і завдає значної шкоди жінкам, яких це зачіпає.

Крім того, необхідне єдине законодавче регулювання поводження з маніпульованим контентом у медіа та в соціальних мережах. Вирішувати, в якому контенті йдеться про вільне висловлювання, а який виходить за межі свободи вираження думки, не має бути в компетенції таких мереж, як Facebook, Twitter, YouTube.

Це є прерогативою законодавця та правової держави. Однак перші не повинні піддаватися спокусі заборонити діпфейки у принципі.

Опріч загроз, ця технологія також відкриває цікаві можливості, зокрема, для освіти, кінематографу та сатири. Сама по собі технологія є нейтральною, то саме люди використовують її на користь чи шкоду суспільству.

Авторка

Аґнєшка М. Вальорска

Аґнєшка М. Вальорска є експерткою з питань діджиталізації та виконавчою директоркою консалтингової компанії в галузі менеджменту і технології Capco.

Вона керувала декількома проєктами з трансформації та інновацій, у тому числі для банків, страхових, фармацевтичних та автомобільних компаній.

Заснована нею консалтингова компанія в галузі цифрових стратегій CREATIVE CONSTRUCTION була придбана компанією Capco у 2020 році.

Її особливо цікавить штучний інтелект та його вплив на взаємодію людини та машини, а отже, на бізнес-моделі та суспільство.

Своїм Digital Innovation Breakfast вона започаткувала серію заходів на ці теми з відомими учасниками, а також опублікувала велику кількість досліджень та статей, і її можна побачити як доповідачку на конференціях та в компаніях.

Її допис у книзі про алгоритмічне суспільство, де вона розглядає етичні питання штучного інтелекту, був опублікований науковим видавництвом Шпрінгер у березні 2020 року.

Вона вивчала суспільні науки та політологію у Варшавському університеті та Університеті імені Гумбольдта в Берліні, була стипендіаткою Фонду «Герті» та Німецького національного академічного фонду.

Практичний посібник

Аґнєшка М. Вальорска

ДІПФЕЙК
ТА ДЕЗІНФОРМАЦІЯ

Переклад з німецької
Володимир Олійник

Коректура
Тетяна Заіченко

Макетування
Олена Грозна