Business Horizons (2020) 63, 135—146

Available online at www.sciencedirect.com

ScienceDirect

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY

www.journals.elsevier.com/business-horizons

ON THE HORIZON

Deepfakes: Trick or treat?

Jan Kietzmann [a,*], Linda W. Lee [b], Ian P. McCarthy [c,d], Tim C. Kietzmann [e,f]

[a] Gustavson School of Business, University of Victoria, 3800 Finnerty Rd., Victoria BC V8P 5C2, Canada
[b] Nottingham Trent University, Nottingham NG1 4FQ, U.K.
[c] Simon Fraser University, Vancouver, BC V6C 1W6, Canada
[d] LUISS Guido Carli University, Rome 00197, Italy
[e] Donders Institute for Brain, Cognition, & Behaviour, Radbound University, 6525 HR, Nijmegen, the Netherlands
[f] MRC Cognition & Brain Sciences Unit, Cambridge University, Cambridge CB2 7EF, U.K.

KEYWORDS
Deepfakes;
Fake news;
Artificial intelligence;
Machine learning;
Deep neural networks

Abstract Although manipulations of visual and auditory media are as old as media themselves, the recent entrance of deepfakes has marked a turning point in the creation of fake content. Powered by the latest technological advances in artificial intelligence and machine learning, deepfakes offer automated procedures to create fake content that is harder and harder for human observers to detect. The possibilities to deceive are endless—including manipulated pictures, videos, and audio—and organizations must be prepared as this will undoubtedly have a large societal impact. In this article, we provide a working definition of deepfakes together with an overview of its underlying technology. We classify different deepfake types and identify risks and opportunities to help organizations think about the future of deepfakes. Finally, we propose the R.E.A.L. framework to manage deepfake risks: **R**ecord original content to ensure deniability, **E**xpose deepfakes early, **A**dvocate for legal protection, and **L**everage trust to counter credulity. Following these principles, we hope that our society can be more prepared to counter deepfake tricks as we appreciate deepfake treats.

1. What are deepfakes?

Have you seen the 2019 video footage of President Obama swearing during a public service announcement? How about the one with Mark Zuckerberg announcing that he is deleting

* Corresponding author
E-mail addresses: jkietzma@uvic.ca (J. Kietzmann), linda.lee@ntu.ac.uk (L.W. Lee), imccarth@sfu.ca (I.P. McCarthy), t.kietzmann@donders.ru.nl (T.C. Kietzmann)

https://doi.org/10.1016/j.bushor.2019.11.006
0007-6813/© 2019 Kelley School of Business, Indiana University. Published by Elsevier Inc. All rights reserved.

Facebook, which attracted 72 million views and led to outrage among viewers who believed the content to be authentic? What about the *Willy Wonka & the Chocolate Factory* clip in which Ryan Reynolds takes real star Gene Wilder's place? If you did, and even for a moment believed their surprising content to be genuine, then you were tricked. Welcome to deepfakes. While analog and digital fakes are not new, *deepfakes* leverage powerful techniques from machine learning (ML) and artificial intelligence (AI) to manipulate or generate visual and audio content with a high potential to deceive.

The phenomenon gained its name from an anonymous user of the platform Reddit, who went by the name 'deepfakes' (deep learning + fakes) and who shared the first deepfakes by placing unknowing celebrities into adult video clips. By sharing the computer code that produced the deepfakes, widespread interest spawned in the Reddit community and led to an explosion of fake content. The first targets of deepfakes were famous people, including actors (e.g., Emma Watson, Scarlett Johansson), singers (e.g., Katy Perry) and politicians (e.g., U.S. presidents Obama and Trump), whose faces were transposed without their permission onto others. One of the early deepfakes that showcased the power of AI and deep learning was 2017's "Synthesizing Obama" (Suwajanakorn, Seitz, & Kemelmacher-Shlizerman, 2017), which featured an impressive use of lip-syncing technology based on existing audio footage. Today, we could be watching the leader of one country convincingly deliver a speech by the leader of another country, or vice versa. Such deepfake trickery is alarmingly successful for two main reasons: believability and accessibility.

1.1. Believability: Fake content is becoming more believable

The impact of deepfakes is significant because, although trust in photography has eroded over the past few decades thanks to image-editing technology (Westling, 2019), we still put a lot of stock in photographic evidence (Granot, Balcetis, Feigenson, & Tyler, 2018; Porter & Kennedy, 2012). We tend to place even more trust in the voices we know and the videos we watch (Brucato, 2015). The brain's visual system, despite being largely robust in natural settings, can be targeted for misperception. Classic examples include optical illusions and bistable figures such as the well-known Jastow rabbit-duck and Rubin vase-faces that can be viewed in two different ways (see Figure 1; Kietzmann, Geuter, & König, 2011). The surprise and disbelief upon the reveal of sleight-of-hand tricks confirm how much we trust our eyes, even if we know we are about to be fooled. If we see something with our own eyes, we believe it to exist or to be true—even if it is unlikely—as was the case with the deepfake examples at the beginning of this article.

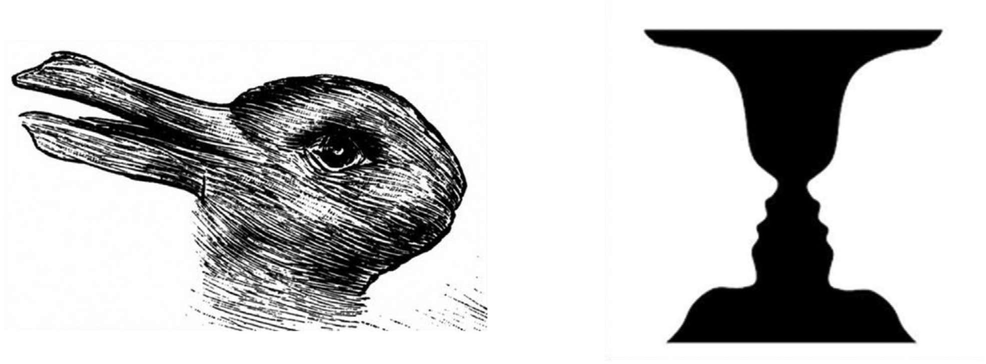
Discerning viewers of early deepfakes could often tell that the content had been altered. Today, only 2 years after the term deepfakes was coined, it is becoming harder and harder to distinguish authentic from artificially-created videos. To illustrate deepfakes, we reference a video throughout this article called "Jim Carrey GLOWS." The original shows an interview with Alison Brie, the lead actor from the Netflix series *GLOW*, which originally aired on *Late Night with Seth Meyers*. In the deepfake, comedian Jim Carrey's face imperceptibly replaces Brie's. In the ensuing discussions online, Carrey—despite having no part in the creation of this video—was even credited for his knack of impersonating others. Deepfakes are becoming much better and much more believable, fast.

1.2. Accessibility: Creating deepfakes is becoming easier

Post-production work on movies has long made fakes appear very realistic at the cinema. For example, *The Curious Case of Benjamin Button* won the Academy Award for Best Visual Effects in 2009. The movie relied on computer-generated imagery (CGI) to help tell the story of a baby born with the appearance and maladies of an elderly man who then spends 84 years growing younger, metamorphosing into an infant. Creating such fakes requires expertise, extensive training, expensive hardware, and special software. Despite what the term CGI suggests, each project is a result of labor-intensive work. However, tools of today—and certainly those of tomorrow—allow anyone to create fakes that appear real without significant investment in training, data collection, hardware, and software. Unskilled individuals will soon be able to manipulate existing media or generate new content with relative ease.

In 2018, the popular face-swapping program FakeApp required large amounts of input data to generate deepfakes. In 2019, similar applications were already less demanding and more accessible. Zao, the popular Chinese app for mobile devices, lets users place their faces into scenes from hundreds of movies and TV shows for free. All Zao required as source material was a series of selfies

Figure 1. Bistable figures: Your brain decides what you perceive



Jastow rabbit-duck

Rubin vase-faces

with specific facial expressions and head postures. A few clicks and seconds were all that was needed to put one's face into a famous movie scene. Alternatively, for those who disliked what someone said, products like Deepmind's WaveNet can be used to generate realistic speech from text input. This can later be integrated with automated video editing to change the words coming right out of someone's mouth. Stanford University researchers demonstrated text-to-speech (TTS) editing of a talking-head video by changing Apple's price per share in a video announcement simply by substituting a number in the text transcript (Fried et al., 2019).

We can expect the creation and distribution of fake content, and the resultant confusion and uproar, to only increase as the accessibility of technology to create high-quality deepfakes improves. Likewise, social media platforms that provide organizations and individuals with the tools and technology to create and post content (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011; Kietzmann, Silvestre, McCarthy, & Pit, 2012) make it very easy to distribute deepfakes; this is one of the many dark sides of social media (Baccarella, Wagner, Kietzmann, & McCarthy, 2018).

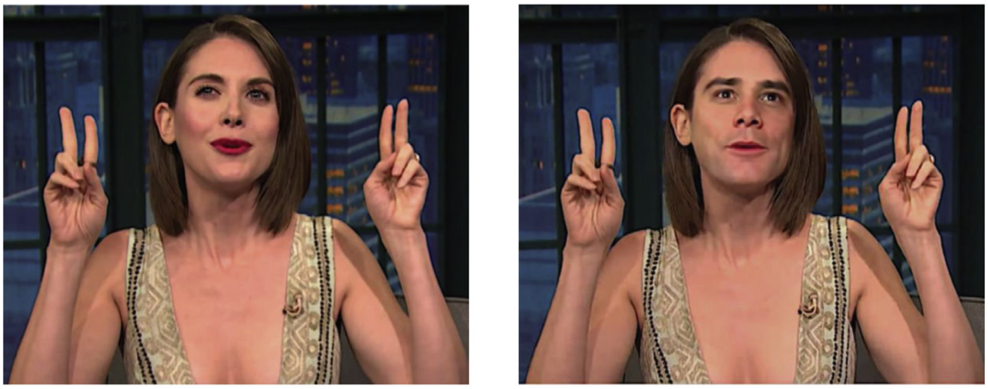
1.3. Believability and accessibility: A recipe for success

In combination, developments in believability and accessibility drive the popularity and influence of deepfakes. Soon, everyone will be able to choose to be the star of their favorite movie, possibly choosing their spouses, friends, or colleagues as their romantic partners, allies, or enemies in AI-manipulated movies. Or we can be placed, unwillingly or unwittingly, within highly undesirable movies—which continues to happen to celebrities—or we can be seen and heard saying things we never said (e.g., sharing fake news about the companies we work for). Whether we want to or not, more and more deepfake images, audio, and video will be created and shared. This is alarming, and appropriate technological and societal countermeasures will be increasingly important and necessary. Yet, deepfakes also offer potential benefits. We provide a brief tour of the most commonly used deepfake technique before highlighting different types of deepfakes and how these affect individuals, organizations, and governments. We conclude by promoting the R.E.A.L. framework to manage deepfake risks.

2. How deepfakes work

Before we discuss how deepfakes work, we need to make clear that existing deepfake techniques and technologies are continually changing, and entirely new ones are already emerging. The majority of current deepfakes in the visual domain follow a procedure in which the real face of a person is exchanged with a fake image showing somebody else. As an example of this, consider the images from the "Jim Carrey GLOWS" deepfake video mentioned above. Figure 2 shows a screenshot of Alison Brie from the original talk show interview on the left, and on the right is a frame from the resulting deepfake video featuring Brie's body with Carrey's face. We have chosen to use this example for three reasons. First, it shows a female and a male celebrity, both likely known to many readers. Second, the deepfake actually exists, and readers can look at the original video and the deepfake output to see for themselves how convincing the deepfake is. We ask the reader to do this so that our explanation of the creation

Figure 2. A deepfake featuring Jim Carrey and Alison Brie



Original showing Alison Brie

Deepfake showing Jim Carrey instead of Brie

process is easier to follow. Third, compared to political deepfakes, this video's content is not deeply controversial and thus does not distract the reader's attention from understanding the process of creating a deepfake.

Three steps were taken to create this fake (see Figure 3):

1. The region of the image showing Brie's face was extracted from an original movie frame;

2. This image was then used as input to a *deep neural network* (DNN), a technique from the domain of ML and AI that was used to automatically generate a matching image showing Carrey instead (Step 2); and

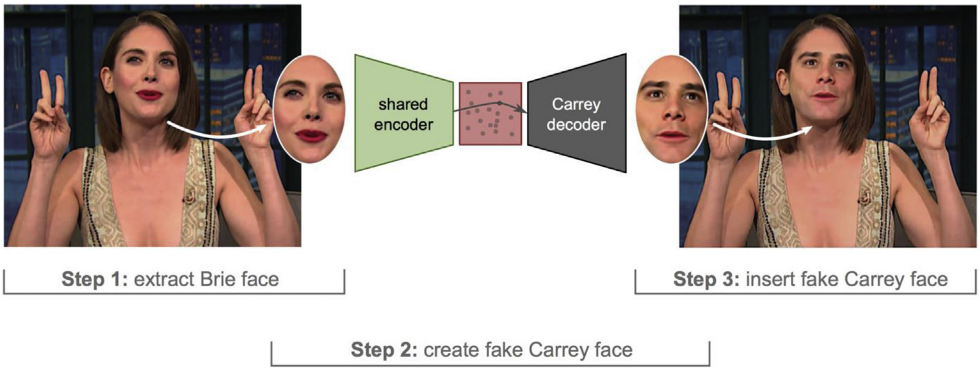
3. This generated face was then inserted into the original reference image to create the deepfake.

The central technical advance of deepfakes lies in Step 2, the automated creation of fake facial images that match the original in all elements but the identity of the person shown. We explain the process by which this is accomplished next.

2.1. Deep learning for deepfakes

As the name suggests, the main technological ingredient in creating deepfakes is *deep learning*: an ML technique from AI that can be used to train DNNs reminiscent of neurons in the brain. DNNs consist of a large set of interconnected artificial neurons, commonly referred to as *units*. Much like neurons in the brain, each unit itself performs a rather simple computation, and all units together can perform complex nonlinear operations such as recognizing a specific person from seeing pixels on a screen (Kietzmann, McClure, & Kriegeskorte, 2019).

Figure 3. Three-step procedure for creating deepfakes


shared encoder
Carrey decoder
Step 1: extract Brie face
Step 3: insert fake Carrey face
Step 2: create fake Carrey face

In the brain, information flow is regulated by the strength of the connections among neurons. To get better at a given task, the brain's learning mechanisms operate on these connections, strengthening or weakening them as required to improve our task performance over time. Likewise, the computations of DNNs are dictated by the strength of the connection of their respective units. These connections, too, need to be trained. Untrained DNNs have random connections among units, which will lead to random information flow through the network and thereby random output. For an untrained DNN operating on images of faces, all facial expressions are thereby arbitrary and indiscriminate, and correctly identifying a facial expression would only happen by chance. A trained DNN, on the other hand, will have improved the connection strength of the units and learned the underlying characteristics of a face.

The goal of deep learning is to update the connection strengths—or *weights* in DNN terminology—to optimize the information flow and output. This progressively drives the network output to minimize errors by defining how the network should ideally respond in a variety of known conditions. For example, when shown known input images, DNNs can be trained to adjust their weights to reduce detection errors so they can eventually identify and properly detect objects in the real world, estimate 3-D depth from 2-D images, and recognize digits and letters on bank checks, license plates, tax forms, letters, and so on. While the training process can lead to unprecedented task performance, it is data-hungry. Today's deep learning requires millions of connection weights to be learned, which in turn necessitates large sets of training data. That is why mainly celebrities are targeted by deepfakes: because an extensive library of images and videos already exists to train the networks.

2.2. The autoencoder

Now that the general procedure and the basic concepts of deep learning are explained, we can take a closer look at the process of creating deepfake content. To illustrate this process, we compare what a DNN does when it creates a deepfake of a facial image to what artists do when they draw a picture of a face.

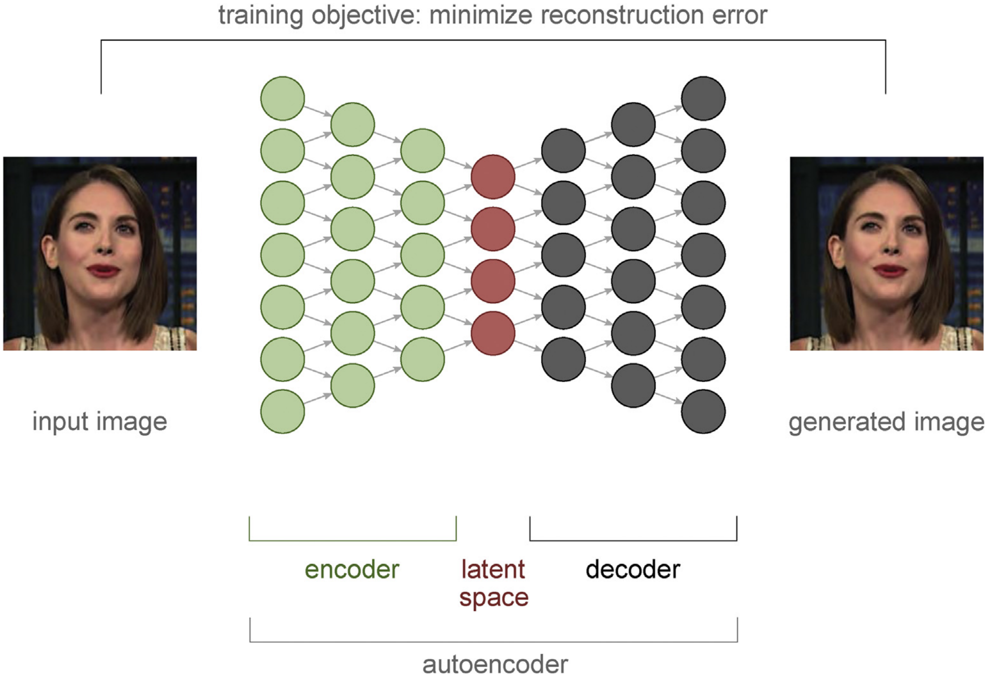
After studying a number of photographs, artists can often draw pictures of the people depicted in them, even placing them in novel scenarios. For this to succeed, artists learn to generate key characteristics of their photo reference such as the smile or the eye expression (e.g., raising of eyebrows, or lowering of the head while looking up). This compression of the image into patterns and characteristics of the input is a result of the limited capacity of our brains to store visual information and is needed for artists to create novel images beyond existing pictures.

A deep network architecture that mimics a similar process to an artist making sense of a human's face is an *autoencoder* (auto referring to the self). Based on a given, large set of input images (e.g., all depicting Alison Brie), an autoencoder is trained to recognize key characteristics of a face and subsequently recreate input images as its output. The process of first recognizing a comparably small number of facial characteristics in the input and then generating real-looking faces as output is accomplished in three subparts: an *encoder*, a *latent space*, and a *decoder* (see Figure 4).

Much like an artist drawing an image, the encoder goes through a similar compression process. It takes tens of thousands of pixels and compresses them into typically 300 measurements that relate to particular facial characteristics. They encode whether the eyes are open or closed, the head pose, the emotional expression, the eye expressions, ambient light, or skin color—all characteristics to which an artist may pay attention. The job of a successful encoder is to transfer an input image into these 300 measurements. Put differently, the encoder enables information to flow from a very detailed input image into what is known as a compressed information bottleneck. The joint activity of these units signals the presence or absence of facial features in the input image. To illustrate, let us consider an encoder compressing the input into only two measurements, one of which expresses the horizontal angle of the head while the other indicates whether a person is smiling or looks surprised. Provided with an input image, the encoder will yield two measurements that jointly encode head orientation and emotion. These can be visualized as a point in 2-D space in which the intersection of the x- and the y-axes represent the two measurements. The space of all possible combinations of measurements of facial characteristics is known as latent space.

Latent spaces are often compared to information bottlenecks. For the autoencoder, this bottleneck is needed so that the network can learn more general facial characteristics rather than memorizing all input examples of specific people. The compression achieved by encoding an input image into the latent space is remarkable. If the latent space consisted of 300 measurements, it would only require 0.1% of the memory needed to

Figure 4. Autoencoder: A DNN architecture commonly used for generating deepfakes



store the original input image. As noted previously, the latent space represents different facial aspects of the person on which it is trained. An autoencoder trained on images showing Alison Brie's face, for instance, will learn to map a given input image of her into a latent space that specifically represents her.

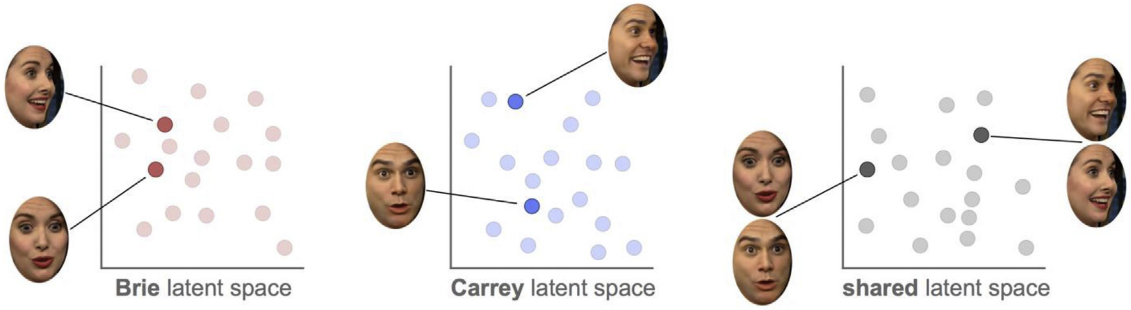
The path from the information bottleneck to the output must recreate an image from the latent space. It is known as the decoder. While the encoder's job is to compress an input image into a set of only 300 measurements (i.e., a specific point in the latent space), the purpose of the decoder is to decompress this information in order to reconstruct an image as truthfully as possible. In our example, its job is to reconstruct the input image of Alison Brie from its representation in the latent space. The performance of the whole autoencoder network is measured by how much the input and generated (output) images resemble each other.

In summary, the autoencoder moves beyond existing image material and learns a generative model of a person's face. As mentioned above, every point in the latent space corresponds to an image of a given person. An autoencoder trained on Brie includes a decoder that can generate fake but eerily real-looking Brie images. The trouble, however, is that, while the autoencoder can generate different faces from select points in latent space, we cannot simply instruct this Brie image generator to create a smiling Brie in the same way that we could instruct an artist to draw one. While all faces are points in the latent space (see Figure 5 for a 2-D illustration), we do not actually know which point in this vast space of nearly infinite possibilities will correspond to the image we desire. Solving this problem is the trick that makes deepfakes seem like works of magic.

2.3. The deepfake trick

To identify specific images, we need a way to find the corresponding points in the latent space. The trick for creating deepfakes is to set up the structure of the autoencoder in a way that an image of another person can act as a guide to help find the specific combination of measurements that yields the desired image. If the trick works, one can use an image of Alison Brie as a guide, and subsequently generate a previously nonexistent picture of Jim Carrey showing the same facial expression and head pose. The input image acts as a reference point, which is similar to telling an artist to draw a picture of you but with the asymmetrical grin of actor Andy Samberg from sitcom *Brooklyn Nine-Nine*, or with Elvis Presley's low-riding eyelids and his slightly quizzical raised eyebrows.

Figure 5. An illustrative example of an autoencoder latent space trained on faces



The trick to making this possible is using the same shared encoder for both people. In the encoding process, the DNN selects 300 measurements it deems meaningful based on the training images for each person. If images of two people are compressed on separate encoders (see Figure 5), different features would be seen as meaningful and we could not combine them in a valuable way (i.e., the expressions on Brie and Carrey would not be aligned).

The autoencoder trick is to train two autoencoders, each with a person-specific decoder, using the exact same encoder. This encoder will learn to use general features that the faces of both people have in common (see Figure 5, right panel). This allows for similar pictures of two different people to be positioned in a similar location of the latent space. For example, pictures showing either a smiling Carrey or Brie will lead to very similar measurements, or unit activations, in the latent space.

Similar measurements resulting from images of two separate people are key to understanding deepfakes. They allow us to transform a picture showing the face of one person into one showing somebody else. The resulting image will be 100% fake, but the generated face will exhibit the same emotional expression, head posture, etc. as shown in the original input image. This new image can then be doctored back into the original image to create a fake scene.

3. Types of deepfakes

AI-based tools to create fake content, like all technologies, are progressing sharply from their early incubation stage to a period of rapid growth and increased performance. Table 1 presents the different types of deepfakes and the potential business applications that will emerge as their underlying techniques approach maturity.

4. The impact of deepfakes on individuals, organizations, and governments

Most of the examples of deepfakes in this article present a gloomy look at society, showing how we are currently using deepfakes to fool and potentially exploit others. Unfortunately, that is how technology is often first used. Technological progress, it seems, promotes the good and bad in people, moving us forward and backward at the same time. Deepfakes are no exception and, like other technologies, will have a bright side and a dark side (see Baccarella et al., 2018). Thus, we want to balance our discussion by outlining several bright and dark opportunities deepfakes offer to individuals, organizations, and governments. Then, we present a framework for how decision makers, technologists, leaders of social media platforms, and policymakers can deal with the challenges presented by the dark side of deepfakes.

As individuals, we may soon enjoy injecting ourselves into Hollywood movies and becoming the hero(ine) in the games we play on our phones or game consoles. Instead of going to the store, we might 'deepfake ourselves' by sharing our photos—and eventually, our personal decoders—in order to create virtual mannequins that model different outfits on us. It is the ultimate personalization (Dietmar, 2019). We may appreciate the entertaining side of deepfakes as well, as videos like the Brie/Carrey face-swap or the many deepfakes featuring Nicolas Cage in various Hollywood scenes are likely to continue. In these early days of deepfakes, their quality and believability—as well as their strangeness and newness—make these videos engaging.

While these examples of deepfakes are not inherently malicious or created with the intention of causing harm, they are also not victimless

Table 1. Types and examples of deepfakes

Type	Description	Current example	Business application
Photo deepfakes	**Face and body-swapping** Making changes to a face, replacing or blending the face (or body) with someone else's face (or body)	FaceApp's aging filter alters your photo to show how you might look decades from now (Kaushal, 2019)	Consumers can try on cosmetics, eyeglasses, hairstyles, or clothes virtually
Audio deepfakes	**Voice-swapping** Changing a voice or imitating someone else's voice	Fraudsters used AI to mimic a CEO's voice and then tricked a manager into transferring $243,000 (Suwajanakorn et al., 2017)	The voice of an audiobook narration can sound younger, older, male, or female and with different dialects or accents to take on different characters
	Text-to-Speech Changing audio in a recording by typing in new text	Users made controversial Jordan B. Peterson, a famous professor of psychology and author, say anything they wanted until his threat of legal action shut the site NotJordanPeterson down (Cole, 2019)	Misspoken words or a script change in a voiceover can be replaced without making a new recording
Video deepfakes	**Face-swapping** Replacing the face of someone in a video with the face of someone else	Jim Carrey's face replaces Alison Brie's in a *Late Night with Seth Meyers* interview.	Face-swapped video can be used to put the leading actor's face onto the body of a stunt double for more realistic-looking action shots in movies.
	Face-morphing A face changes into another face through a seamless transition	Former *Saturday Night Live* star Bill Hader imperceptibly morphs in and out of Arnold Schwarzenegger on the talk show *Conan*	Video game players can insert their faces onto their favorite characters
	Full-body puppetry Transposing the movement from one person's body to that of another	"Everybody Dance Now" shows how anyone can look like a professional dancer	Business leaders and athletes can hide physical ailments during a video presentation.
Audio & video deepfakes	**Lip-syncing** Changing the mouth movements and words spoken in a talking head video	In "You Won't Believe What Obama Says In This Video!" Jordan Peele edits Obama to use profanity in a public service announcement	Ads and instructional videos can be 'translated' into other languages using the same voice used in the original recording

crimes. After all, celebrities did not consent to their portrayal in the deepfakes and might object to them strongly. The same technology that made the Carrey/Brie face-swap entertaining, for instance, was used time and again to transplant the face of Scarlett Johansson and many of her famous contemporaries onto the bodies of actors in adult videos. The harm this can do to us all becomes even clearer in the case of a then-18-year-old woman—an ordinary, nonfamous citizen—who one day discovered hundreds of explicit deepfake images and videos with her face on the bodies of porn actresses (Melville, 2019). These deepfakes not only put her reputation at risk, but also her emotional well-being, her career prospects as an aspiring lawyer, and her physical safety. With such a powerful technology and the increasing number of images and videos of all of us on social media, anyone can become a target for online harassment, defamation, revenge porn, identity theft, and bullying—all through the use of deepfakes.

For organizations, deepfakes have pros and cons, too. The upsides can be found in the entertainment and fashion industries in which celebrities can simply make their personal deep network models available so that deepfake footage can be created without the need for travel to a video shoot, for example. Hollywood will be an early adopter. Certainly, face-swapping (aka face-

Table 2. Links to the videos referenced in this article

Video Title	URL
Original: Alison Brie Snagged Her GLOW Role by Freestyling about Lady Parts	https://www.youtube.com/watch?v=QBmYDzLhWoY
Deepfake: Jim Carrey GLOWS	https://www.youtube.com/watch?v=b5AWhh6MYCg
A world without Facebook	https://www.facebook.com/watch/?v=343812022777503
Bill Hader impersonates Arnold Schwarzenegger [DeepFake]	https://www.youtube.com/watch?v=bPhUhypV27w&feature=youtu.be
David Beckham speaks nine languages to launch Malaria Must Die Voice Petition	https://youtu.be/QiiSAvKJIHo
Everybody Dance Now	https://youtu.be/PCBTZh41Ris
Ryan Reynolds & the Chocolate Factory	https://www.youtube.com/watch?v=3qTXIwjAUZM
Synthesizing Obama: Learning Lip Sync from Audio	https://youtu.be/9Yq67CjDqvw
Text-based Editing of Talking-head Video (SIGGRAPH 2019)	https://www.youtube.com/watch?v=0ybLCfVeFL4&feature=youtu.be
You Won't Believe What Obama Says In This Video!	https://youtu.be/cQ54GDm1eL0

leasing) and voice dubbing will be popular so that movie or advertising producers can fix misspoken lines or make script changes without rerecording footage and create seamless dubs of actors speaking different languages. More realistic stunt doubles can be created and actors can look older or younger with the use of deepfakes instead of time-consuming make-up.

In terms of the negative effect of deepfakes on organizations, technological advancements often make incumbents redundant. For example, the entire dubbing and re-voicing industry, which has long translated movies so that the new words match the original lip movement of the actor, is endangered and at risk of becoming extinct now that languages and lips can be changed. Such industry developments are evolutionary. In terms of the dark side of deepfakes, we predict that in the early days, many unsuspecting firms will fall victim to trickery. There will likely also be many organizations that will suffer from deepfake news releases. Videos that deliberately state false earnings estimates will hurt stock prices, and deepfake videos of CEOs in compromising situations will impact their firms' reputation and put stakeholder agreements at risk, to name just a few examples. Then, of course, there are lots of opportunities for *algorithmic blackmail*, in which managers are offered a choice to either pay a fee to stop a deepfake from being shared or suffer the very public consequences.

For governments, the positive potential of deepfakes lies in the ability to communicate with various stakeholders in a way that is accessible to them. For instance, a public service announcement can be broadcast in a number of different languages, much like a consensual deepfake in which football celebrity David Beckham advocates in nine different languages and voices to end malaria (see Table 2). At the same time, the dark side of deepfakes is undeniably powerful, with the potential to give the average person the ability to create and distribute well-timed acts of sabotage. A government leader could be shown covering up a misdeed or making racist remarks just before an election or a major decision. Further, deepfake technology "will be irresistible for nation-states to use in disinformation campaigns to manipulate public opinion, deceive populations and undermine confidence in [...] institutions" (Riechmann, 2018). Deepfake propaganda and election meddling, and the disinformation they seed threaten efficient governance for all democracies if not democracy itself.

5. The R.E.A.L. framework for managing deepfake risks

As this article shows, the potential for dark, malicious, deceptive and destructive potential of deepfakes for individuals, organizations (and brands) and governments outweighs the bright opportunities for thoughtful, sincere, and constructive applications today. We hope we motivated managers to think about how deepfakes can transform their businesses in positive ways.

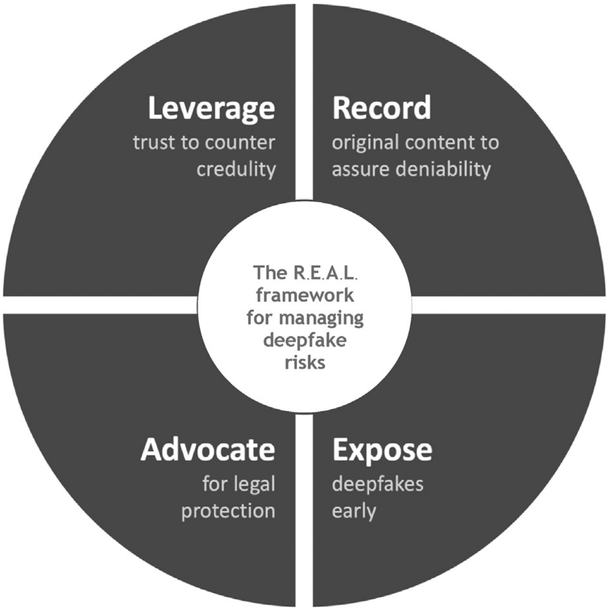
More importantly, we urge all decision makers, including technologists, leaders of social media platforms, and policymakers to help organizations and, in turn, society, prevent and mitigate the dark side of content manipulation. With this goal in mind, we propose a R.E.A.L. framework for managing deepfake risks:

- Record original content to assure deniability;
- Expose deepfakes early;
- Advocate for legal protection; and
- Leverage trust (see Figure 6).

5.1. Record original content to assure deniability

As dark deepfakes often seek to falsely portray somebody doing or saying something and being somewhere, the exposure of such fakes would require evidence to the contrary. Providing this data is referred to as an *alibi service* or a *life log* (Chesney & Citron, 2019). It involves a form of technology tracking and logging a person's life in terms of location, communications, and activities. Despite the potentially negative impact on privacy, from a technology perspective, the availability of mobile, wearable, and smart Internet-of-Things devices makes collecting such data possible to some extent. The data could then be encrypted, stored, and used to help identify and expose the posting of dark deepfakes.

Figure 6. The R.E.A.L. framework for managing deepfake risks


Leverage
trust to counter credulity
Record
original content to assure deniability
The R.E.A.L. framework for managing deepfake risks
Advocate
for legal protection
Expose
deepfakes early

A related technological approach to managing and limiting dark deepfakes is to develop ways and practices for authenticating genuine content. Consider, for example, a technology called Amber Authenticate, which works on devices that produce genuine photographic, audio, and video content in real time as the content is recorded. It creates a truth layer that is original content, cryptographically stamped with numerous digital fingerprints, and then archived on a public blockchain. This fingerprinting of digital content is used to track its provenance as it is distributed and to help detect and respond to attempts to produce unwanted manipulations of the original content.

5.2. Expose dark deepfakes early

The international professional services firm KPMG advised that "establishing a governance framework that embraces disruptive technologies and encourages innovation while ensuring risks are identified and managed is essential to an organization's ability to survive and thrive in a digital world" (Lageschulte, 2019). Thus, just as we adopt and develop the technological innovations that gave us deepfakes, there are technological innovations being developed to detect and classify them. These include using AI techniques to identify resolution inconsistencies; the scaling, rotation, and splicing of content that is often central to the creation of a deepfake; and the eye-blinking patterns of the human images. Such detection innovation is helped by national institutions such as the U.S.'s Defense Advanced Research Projects Agency (DARPA), which has a media forensics program, as well as fake-spotter services like Truepic (Hatmaker, 2018). Facebook, too, is investing significant resources into deepfake identification and detection (O'Brien, 2018). Yet, despite such initiatives, it is important to recognize that this a game of cat and mouse, with improvements in detection technology having to keep pace with improvements in deepfake production technology.

5.3. Advocate for legal protection

Deepfake instigators could include ex-partners or bullies, disgruntled employees or competitors, or politically-motivated actors and even nation-state attackers. Social media networks and their involvement in deepfakes need to be revisited in this light, too. Are Facebook, YouTube and the like merely technology platforms, or are they in fact publishers that should be held liable for the content on their sites, including deepfakes? Are they willingly supporting deepfakes? As informed distributors, are they (or should they be) seen as guilty themselves? Underlying legislation (e.g., Section 230 of the U.S. Communications Decency Act) currently does not offer such provisions for distributor liability for technology platforms even in cases for which platforms possess direct knowledge of the illegal comments and fail to act once they are made aware. In response to a 2018 doctored video that made House Speaker Nancy Pelosi appear to be slurring her speech (not a deepfake), Facebook said: "We don't have a policy that stipulates that the information you post on Facebook must be true" (Chiu, 2019). In contrast, when bookstores are credibly informed that a book they sell includes libelous content but fail to act, they can be held legally responsible (Candeub, 2019). In the era of deepfakes, such statements by social media executives and the existing laissez-faire approach of today's legal frameworks should concern us all. Victims should have legal recourse in instances of defamation, malice, breach of privacy, or emotional distress caused by deepfakes, as well as in cases of copyright infringement, impersonation, and fraud involving deepfakes. However, there are few legal tools that address deepfake threats today, and we hope that this article motivates many to advocate and lobby for legal changes that reflect the most recent technological threats.

5.4. Leverage trust

For managers who want to be proactive, the best way forward might be to strengthen brands and the relationships between brands and their customers. This means ensuring products perform well and are consistent with what their brands promise. While this advice may sound simple, how many brands do we know that promise more than they deliver? In the chase for market share and visibility, some brands have forgotten these fundamentals. Likewise, brands that provide superior value build trust and commitment in their customer relationships, and customers establish lasting emotional bonds with them in turn (Morgan & Hunt, 1994; Sashi, 2012). Strong brands will be better positioned to weather deepfake assaults, as their stakeholders will defend the brand (Pongsakornrungsilp & Schroeder, 2011; Punjaisri & Wilson, 2017) or at least place more trust in the brand than what they see or hear from a suspect video. When brands that are built on strong ethics are portrayed in an unfavorable light in deepfakes, the hope is that stakeholders will not simply believe their eyes and ears but be more critical and think for themselves.

6. The time is right for deepfakes

Analog and digital content manipulations are not new, and the act of doctoring content is as old as the media industry itself. However, recent developments in the use of deep learning mark the beginning of the next phase of content doctoring. Novel and publicly available tools now enable the 'semiautomated creation of much-improved and more convincing fakes. In this article, we explained what deepfakes are and how they are currently produced using a deep network structure called an autoencoder which, much like an artificial artist, learns to concentrate on key characteristics of a person's face to generate previously nonexistent images.

We put together a framework of types of deepfakes—including image, audio, video, and audio/video—and described possible business applications for each. Deepfakes have a bright and dark side and we identified their implications for individuals, organizations, and governments. Finally, we presented the R.E.A.L. framework to help decision makers, technologists, leaders of social media platforms, and policymakers understand how to counter the dark side of deepfakes. This is an important contribution, for, as Amara's law states, we tend to overestimate the effect of a technology in the short run and underestimate its effect in the long run. Certainly, in the short term, we are likely going to see a wave of deepfake videos, movies, and apps. The timing is perfect. At a time of much-touted fake news, deepfakes will add a powerful tool to fool voters, buyers, and competitors, among others. Some deepfakes might be intended for entertainment purposes, while others might impact the outcome of an election or the stock market. As more of our lives are constantly being captured and shared (e.g., through social media), we provide more and more data about ourselves that will also be used to train DNNs with or without our explicit permission. With this greater understanding of deepfakes, our hope is that we will all be more prepared to counter deepfake tricks as we appreciate deepfake treats.

References

Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, I. P. (2018). Social media? It's serious!

Understanding the dark side of social media. *European Management Journal, 36*(4), 431—438.

Brucato, B. (2015). Policing made visible: Mobile technologies and the importance of point of view. *Surveillance and Society, 13*(3/4), 455—473.

Candeub, A. (2019, June 21). Social media platforms or publishers? Rethinking section 230. *The American Conservative*. Available at https://www.theamericanconservative.com/articles/social-media-platforms-or-publishers-rethinking-section-230/

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*. Available at https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war

Chiu, A. (2019, June 12). Facebook wouldn't delete an altered video of Nancy Pelosi. What about one of Mark Zuckerberg? *Washington Post*. Available at https://www.washingtonpost.com/nation/2019/06/12/mark-zuckerberg-deepfake-facebook-instagram-nancy-pelosi/?noredirect=on

Cole, S. (2019, August 26). A site faking Jordan Peterson's voice shuts down after Peterson decries deepfakes. *Vice*. Available at https://www.vice.com/en_us/article/43kwgb/not-jordan-peterson-voice-generator-shut-down-deepfakes

Dietmar, J. (2019). GANs and deepfakes could revolutionize the fashion industry. *Forbes*. Available at https://www.forbes.com/sites/forbestechcouncil/2019/05/21/gans-and-deepfakes-could-revolutionize-the-fashion-industry/#53cb7f713d17

Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., et al. (2019). Text-based editing of talking-head video. *Cornell University*. Available at https://arxiv.org/abs/1906.01524

Granot, Y., Balcetis, E., Feigenson, N., & Tyler, T. (2018). In the eyes of the law: Perception versus reality in appraisals of video evidence. *Psychology, Public Policy, and Law, 24*(1), 93—104.

Hatmaker, T. (2018, April 30). DARPA is funding new tech that can identify manipulated videos and 'deepfakes.' *TechCrunch*. Available at https://techcrunch.com/2018/04/30/deepfakes-fake-videos-darpa-sri-international-media-forensics/

Kaushal. (2019, July 16). 9 best face swap apps for Android and iOS (2019). *TechWiser*. Available at https://techwiser.com/best-face-swap-apps/

Kietzmann, T. C., Geuter, S., & König, P. (2011b). Overt visual attention as a causal factor of perceptual awareness. *PLoS One, 6*(7), e22614.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011a). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons, 54*(3), 241—251.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In *Oxford research encyclopedia of neuroscience*. Oxford, UK: Oxford University Press.

Kietzmann, J. H., Silvestre, B. S., McCarthy, I. P., & Pitt, L. F. (2012). Unpacking the social media phenomenon: Towards a research agenda. *Journal of Public Affairs, 12*(2), 109—119.

Lageschulte, P. (2019). Four steps to emerging technology governance. *KPMG*. Available at https://advisory.kpmg.us/articles/2018/emerging-technology-governance.html

Melville, K. (2019, August 29). The insidious rise of deepfake porn videos — and one woman who won't be silenced. *ABC News Australia*. Available at https://www.abc.net.au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-image-based-abuse/11437774

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing, 58*(3), 20—38.

O'Brien, S. A. (2018, August 8). Deepfakes are coming. Is big tech ready? *CNN Business*. Available at https://money.cnn.com/2018/08/08/technology/deepfakes-countermeasures-facebook-twitter-youtube/index.html

Pongsakornrungsilp, S., & Schroeder, J. E. (2011). Understanding value co-creation in a co-consuming brand community. *Marketing Theory, 11*(3), 303—324.

Porter, G., & Kennedy, M. (2012). Photographic truth and evidence. *Australian Journal of Forensic Sciences, 44*(2), 183—192.

Punjaisri, K., & Wilson, A. (2017). The role of internal branding in the delivery of employee brand promise. In J. M. T. Balmer, S. M. Powell, J. Kernstock, & T. O. Brexendorf (Eds.), *Advances in corporate branding* (pp. 91—108). Cham, Switzerland: Springer.

Riechmann, D. (2018). I never said that! High-tech deception of 'deepfake' videos. *WCJB*. Available at https://www.wcjb.com/content/news/I-never-said-that-High-tech-deception-of-deepfake-videos-487147011.html

Sashi, C. (2012). Customer engagement, buyer-seller relationships, and social media. *Management Decision, 50*(2), 253—272.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics, 36*(4). Article 95.

Westling, J. (2019, January 30). Deep fakes: Let's not go off the deep end. *The Tech*. Available at https://www.techdirt.com/articles/20190128/13215341478/deep-fakes-lets-not-go-off-deep-end.shtml