EPFL

Forged authenticity: the case of deepfakes

Aengus Collins

Deputy Director
EPFL International Risk
Governance Center

March 2020

irgc
international risk
governance center

Setting the scene

Examples

Risks

Possible responses

What is risk governance?

We define **risk** as an uncertain consequence of an event or activity with respect to something that humans value

We define **governance** as the totality of actions, processes, traditions and institutions by which authority is exercised and collective decisions are taken and implemented

Information unleashed

33 zettabytes

Global datasphere in 2018 (IDC/Seagate)

2.5 billion

Facebook monthly active users (Facebook)

30,000 hours

New video uploaded to YouTube every hour (Statista)

What are deepfakes?

- Deepfake refers to digital content that has been created or manipulated using machine learning

- Typically used to refer to fabricated video content, but machine learning can be used to generate images, audio and text

7,964

Deepfake videos online in December 2018 (Deeptrace)

14,678

Deepfake videos online in December 2019 (Deeptrace)

Setting the scene

Examples

Risks

Possible responses

Images: real or deepfake?

✗

✗

✓

✗

✓

✗

Sources: https://thispersondoesnotexist.com/ and http://www.whichfaceisreal.com/

Video: Bill Hader or Arnold Schwarzenegger?

Source: https://www.youtube.com/watch?v=bPhUhypV27w

Video: emerging political examples (1/2)

Source: https://www.zmescience.com/science/news-science/indias-first-political-deepfake-during-elections-is-deeply-concerning/

Video: emerging political examples (2/2)

Source: https://www.zmescience.com/science/news-science/indias-first-political-deepfake-during-elections-is-deeply-concerning/

Text: predicting the best next word

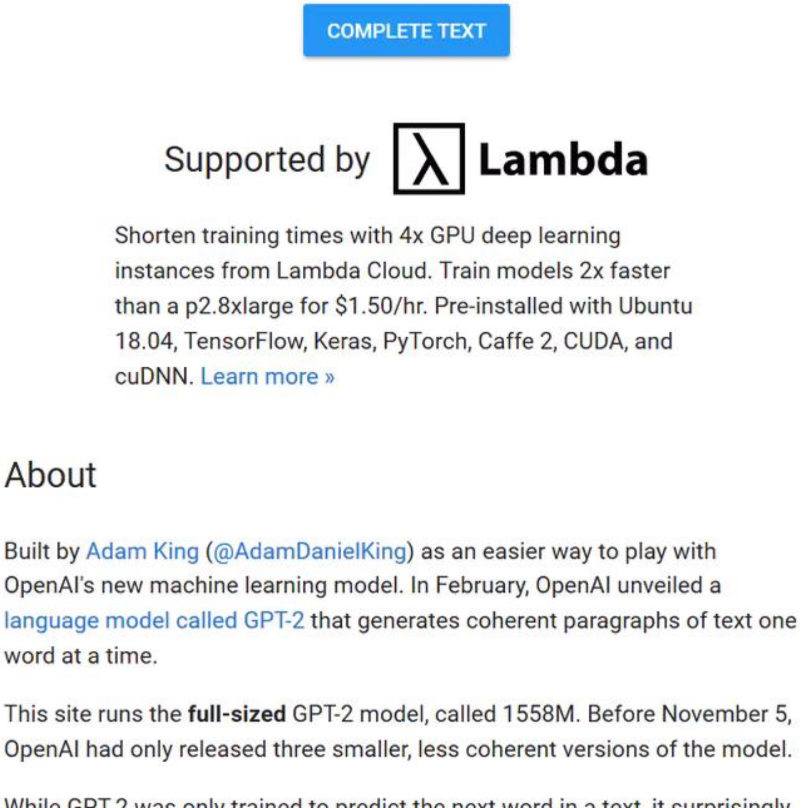
GPT-2 is a neural network, trained on 8m websites, that generates new text in response to an initial prompt

The full model was originally withheld because of fears it would be used maliciously

The results are still very patchy, but show promise:

www.talktotransformer.com

“A train containing a shipment of mobile phones was stolen in Zurich today. Its whereabouts are unknown.”



Supported by λ Lambda

Shorten training times with 4x GPU deep learning instances from Lambda Cloud. Train models 2x faster than a p2.8xlarge for $1.50/hr. Pre-installed with Ubuntu 18.04, TensorFlow, Keras, PyTorch, Caffe 2, CUDA, and cuDNN. Learn more »

About

Built by Adam King (@AdamDanielKing) as an easier way to play with OpenAI's new machine learning model. In February, OpenAI unveiled a language model called GPT-2 that generates coherent paragraphs of text one word at a time.

This site runs the **full-sized** GPT-2 model, called 1558M. Before November 5, OpenAI had only released three smaller, less coherent versions of the model.

While GPT-2 was only trained to predict the next word in a text, it surprisingly

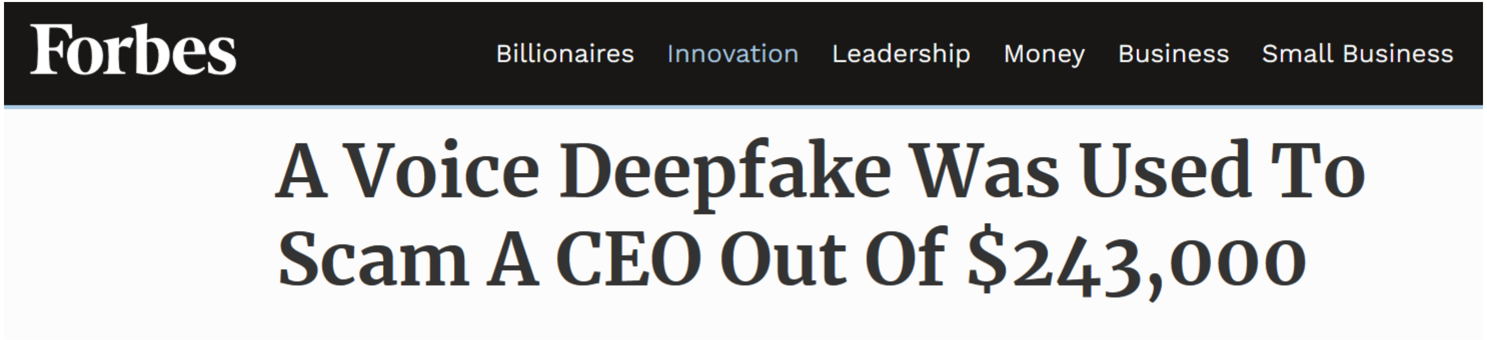
Audio: well behind video, but making progress

Source: https://www.buzzfeednews.com/article/charliewarzel/i-used-ai-to-clone-my-voice-and-trick-my-mom-into-thinking

Audio: worries about misuse

Realistic audio deepfakes could be a particularly powerful mode of social engineering

There is already at least one example of significant fraud that relied on deepfake audio



A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

Source: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/

Setting the scene

Examples

Risks

Possible responses

Where do the main risks lie?

- With one major exception, there are few public examples of deepfakes being used to cause harm

- The exception is the use of deepfake technology to swap women’s faces into pornographic images and videos

- One worry is that deepfakes are evolving more rapidly than our understanding of the risks that they **may** pose

Risk assessment

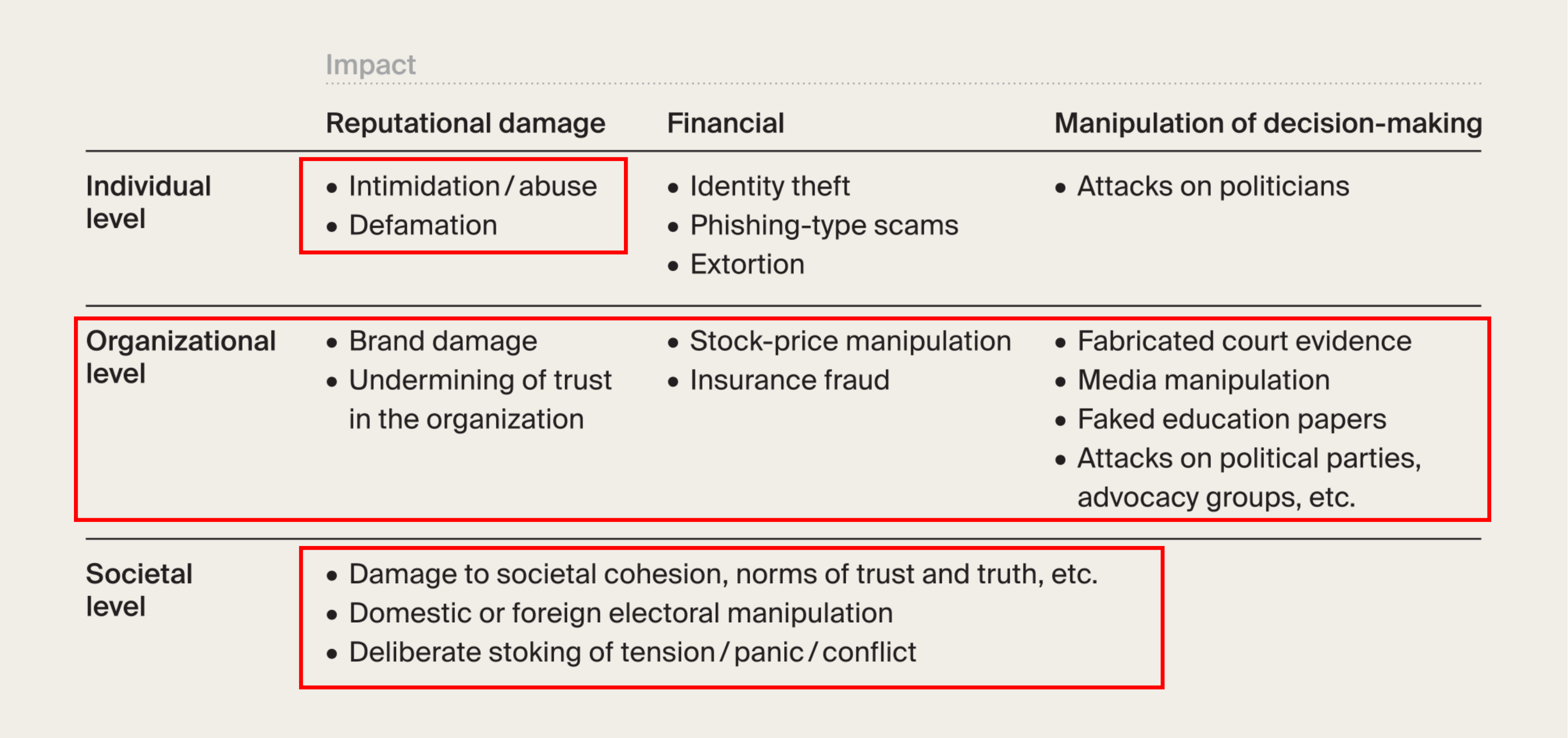
Incident reporting

Impact, motivation and scale

Severity

Scale

Resilience



	Impact		
	Reputational damage	Financial	Manipulation of decision-making
Individual level	• Intimidation / abuse • Defamation	• Identity theft • Phishing-type scams • Extortion	• Attacks on politicians
Organizational level	• Brand damage • Undermining of trust in the organization	• Stock-price manipulation • Insurance fraud	• Fabricated court evidence • Media manipulation • Faked education papers • Attacks on political parties, advocacy groups, etc.
Societal level	• Damage to societal cohesion, norms of trust and truth, etc. • Domestic or foreign electoral manipulation • Deliberate stoking of tension / panic / conflict		

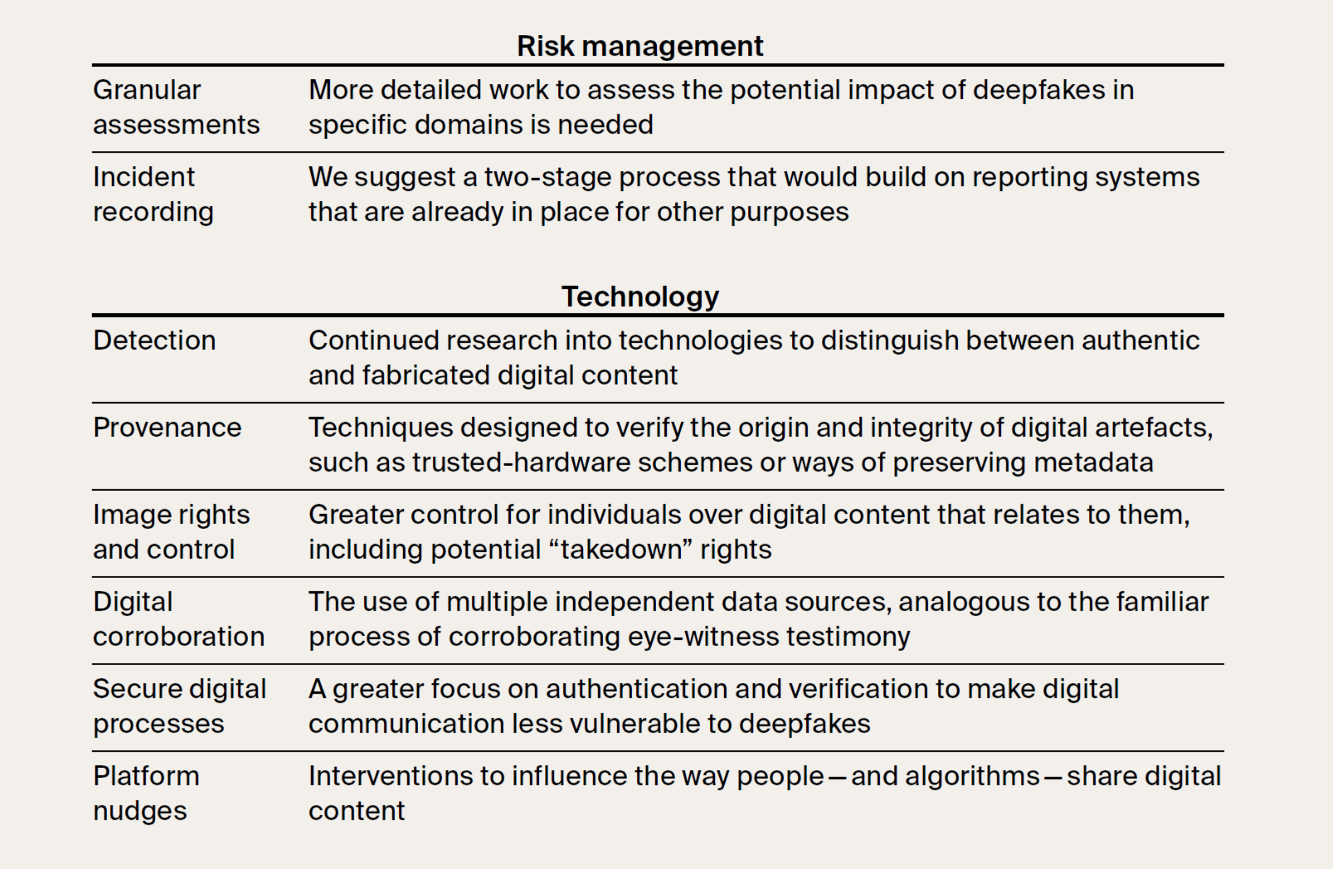
Setting the scene

Examples

Risks

Possible responses

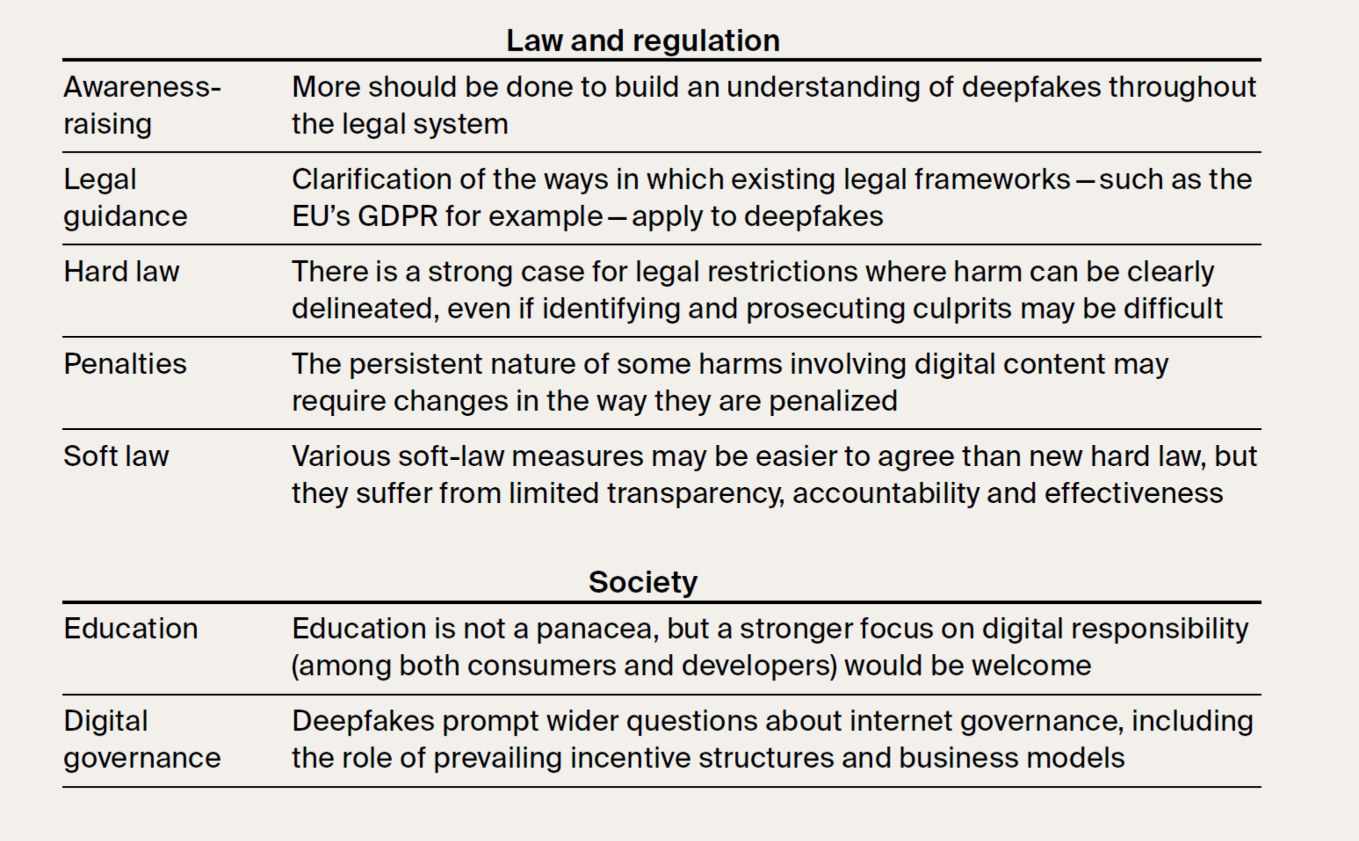
Opportunities to respond



Risk management	
Granular assessments	More detailed work to assess the potential impact of deepfakes in specific domains is needed
Incident recording	We suggest a two-stage process that would build on reporting systems that are already in place for other purposes

Technology	
Detection	Continued research into technologies to distinguish between authentic and fabricated digital content
Provenance	Techniques designed to verify the origin and integrity of digital artefacts, such as trusted-hardware schemes or ways of preserving metadata
Image rights and control	Greater control for individuals over digital content that relates to them, including potential “takedown” rights
Digital corroboration	The use of multiple independent data sources, analogous to the familiar process of corroborating eye-witness testimony
Secure digital processes	A greater focus on authentication and verification to make digital communication less vulnerable to deepfakes
Platform nudges	Interventions to influence the way people — and algorithms — share digital content

Opportunities to respond



Law and regulation	
Awareness-raising	More should be done to build an understanding of deepfakes throughout the legal system
Legal guidance	Clarification of the ways in which existing legal frameworks – such as the EU's GDPR for example – apply to deepfakes
Hard law	There is a strong case for legal restrictions where harm can be clearly delineated, even if identifying and prosecuting culprits may be difficult
Penalties	The persistent nature of some harms involving digital content may require changes in the way they are penalized
Soft law	Various soft-law measures may be easier to agree than new hard law, but they suffer from limited transparency, accountability and effectiveness

Society	
Education	Education is not a panacea, but a stronger focus on digital responsibility (among both consumers and developers) would be welcome
Digital governance	Deepfakes prompt wider questions about internet governance, including the role of prevailing incentive structures and business models

Conclusion

- Deepfakes are expanding rapidly, in terms of (i) quantity, (ii) quality, and (iii) variety
- Even if there are still relatively few public examples of harm, now is the time to assess vulnerabilities
- Deepfakes highlight the importance of fostering trust in an increasingly digitalized world

Thank you

Aengus Collins

international risk
governance center