Forged authenticity: the case of deepfakes

Aengus Collins
Deputy Director
EPFL International Risk Governance Center

March 2020
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.
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)
Examples

Setting the scene

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=bPhUhypyj7w
Video: emerging political examples (1/2)

Video: emerging political examples (2/2)

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](http://www.talktotransformer.com)
“A train containing a shipment of mobile phones was stolen in Zurich today. Its whereabouts are unknown.”

Supported by 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
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

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

## Impact

<table>
<thead>
<tr>
<th>Severity</th>
<th>Scale</th>
<th>Resilience</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Impact</strong></td>
<td><strong>Reputational damage</strong></td>
<td><strong>Manipulation of decision-making</strong></td>
</tr>
<tr>
<td>Individual level</td>
<td>- Intimidation/abuse</td>
<td>- Attacks on politicians</td>
</tr>
<tr>
<td></td>
<td>- Defamation</td>
<td></td>
</tr>
<tr>
<td>Organizational level</td>
<td>- Brand damage</td>
<td>- Fabricated court evidence</td>
</tr>
<tr>
<td></td>
<td>- Undermining of trust in the organization</td>
<td>- Media manipulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Faked education papers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Attacks on political parties, advocacy groups, etc.</td>
</tr>
<tr>
<td>Societal level</td>
<td>- Damage to societal cohesion, norms of trust and truth, etc.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Domestic or foreign electoral manipulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Deliberate stoking of tension/panic/conflict</td>
<td></td>
</tr>
</tbody>
</table>
Setting the scene

Examples

Risks

Possible responses
## Opportunities to respond

### Risk management

<table>
<thead>
<tr>
<th>Granular assessments</th>
<th>More detailed work to assess the potential impact of deepfakes in specific domains is needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incident recording</td>
<td>We suggest a two-stage process that would build on reporting systems that are already in place for other purposes</td>
</tr>
</tbody>
</table>

### Technology

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

#### Law and regulation

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

#### Society

<table>
<thead>
<tr>
<th>Education</th>
<th>Education is not a panacea, but a stronger focus on digital responsibility (among both consumers and developers) would be welcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital governance</td>
<td>Deepfakes prompt wider questions about internet governance, including the role of prevailing incentive structures and business models</td>
</tr>
</tbody>
</table>
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