

SITC 2024 Artificial Intelligence: From Use to Abuse

DISCOVERING THE FAKE USING AI TOOLS

DR, LAMIA FRIHA

LAMIA.FRIHA@UNIGE.CH



UNIVERSITÉ
DE GENÈVE



PLAN



UNIVERSITÉ
DE GENÈVE

- Definitions
- Types and Impacts
- Deepfake Creation
- Deepfake Detection
- Examples and interactive session
- Conclusion



Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users



Source: UBS / Yahoo Finance

 @EconomyApp

 APP ECONOMY INSIGHTS

GENERATIVE AI



UNIVERSITÉ
DE GENÈVE



Generative AI is a set of methods belonging to AI



It is based on « Next Token Prediction Systems »,



Trained on large corpus of data



Natural language interaction

SO WHAT IS A DEEPPFAKE?



UNIVERSITÉ
DE GENÈVE



Deep refers to deep learning techniques



Deepfake means manipulating or generating new image, media, text...



Two aspects:

Creation

Detection



Source <https://arxiv.org/pdf/2208.10913>

WHAT ARE THE IMPACTS IF MISUSED?



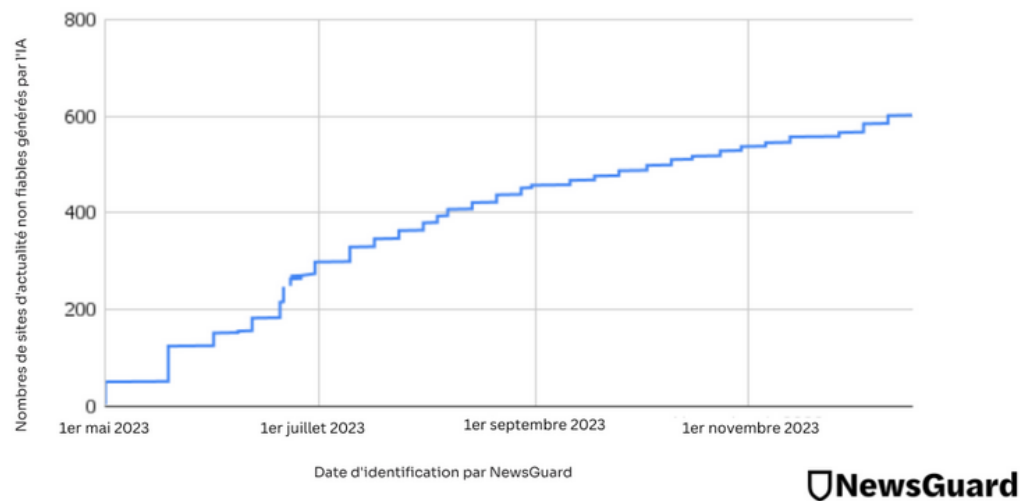
UNIVERSITÉ
DE GENÈVE



- Reputation
- Democracy
- Believe that is real and it is fake
- How to inform about real information
- How to prevent from creating violent production



Sites d'actualité non fiables générés par l'IA par date d'identification



NewsGuard a identifié plus de 600 sites d'actualité non fiables générés par l'IA en 2023. (Graphique NewsGuard)

- «Selon une étude de la société NewsGuard, spécialisée sur la lutte contre les fake news, 20% des vidéos traitant de sujets d'actualité sur TikTok contiendraient de fausses informations»
 - *Hélène Krähenbühl journaliste rts*

ADDRESSING RISKS



UNIVERSITÉ
DE GENÈVE



Regulatory frameworks



Advanced tools

3 YEARS AGO...

[HTTPS://WWW.YOUTUBE.COM/WATCH?V=GLOI9HAX9DW](https://www.youtube.com/watch?v=GLOI9HAX9DW)



2024...OPENAI SORA

[HTTPS://OPENAI.COM/INDEX/SORA/](https://openai.com/index/sora/)



ail inside a



Prompt: A young man at his 20s is sitting on a piece of cloud in the sky, reading a book.



Prompt: A stylish woman walks do

A GENERATIVE AI PRODUCING A NEW CONTENT FOR THE STIC:A HAIKU



UNIVERSITÉ
DE GENÈVE



- *Bits weave in Bern 's breeze,*
- *Code whspers through Alpinepeaks,*
- *Informatique thrives.*

CREATE A FUNNY LIMERICK FOR THE SWISS INFORMATICS SOCIETY



UNIVERSITÉ
DE GENÈVE

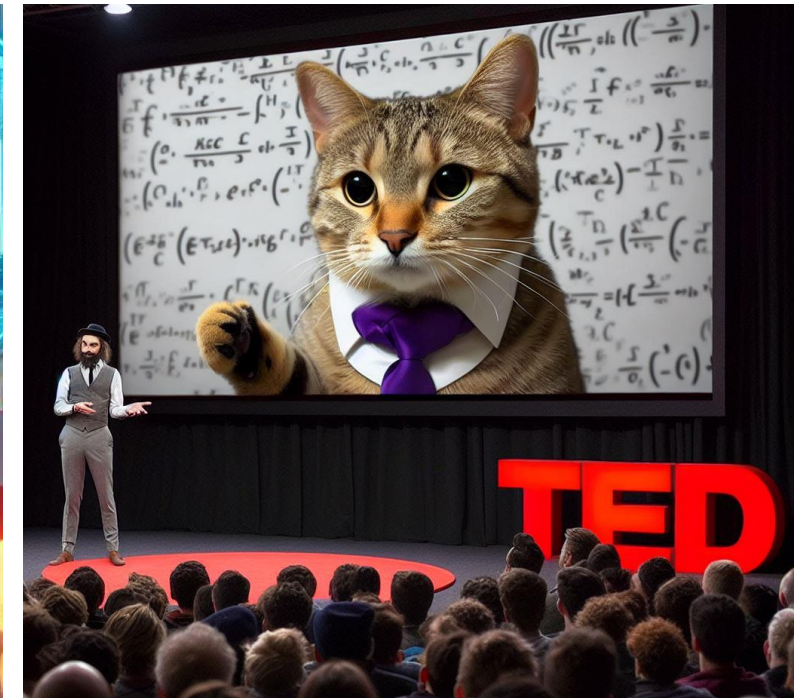
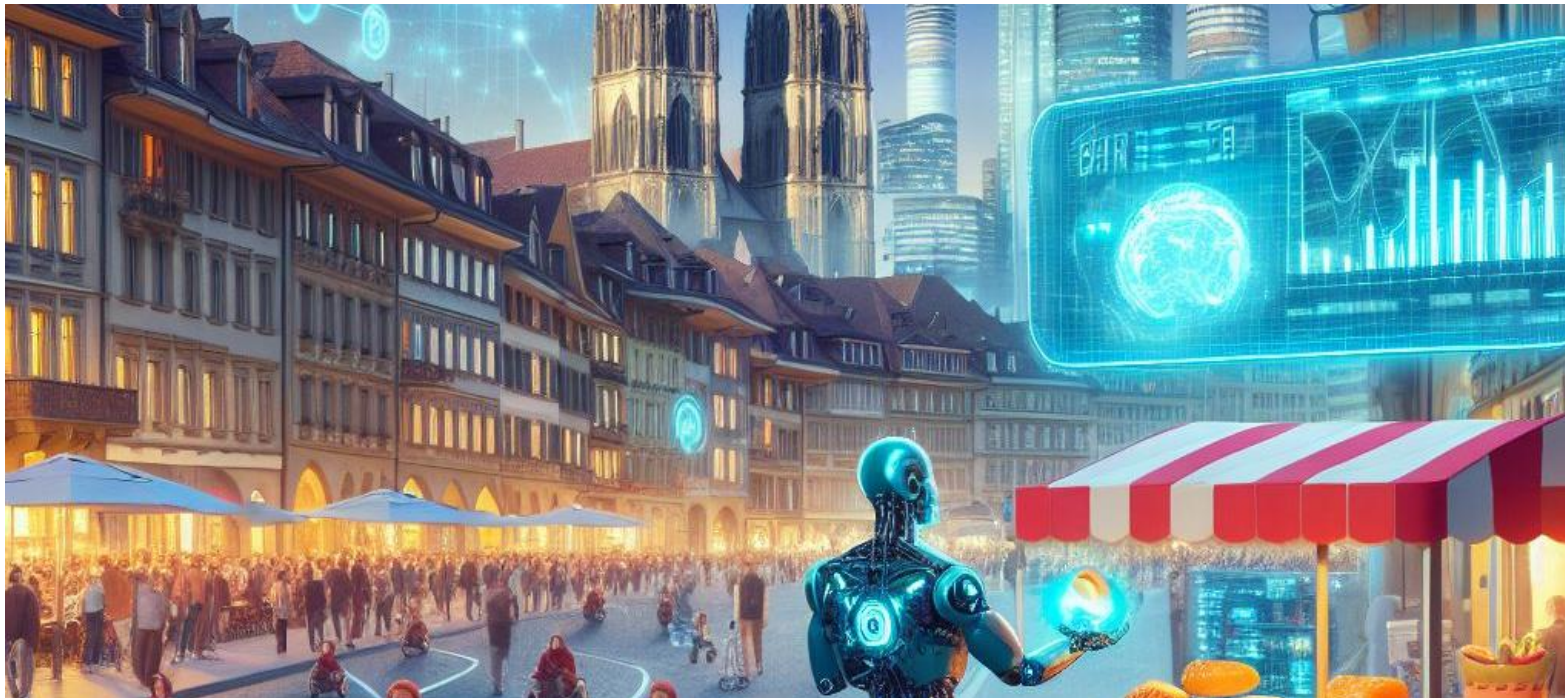


*In Bern, where the code wizards dwell, They
debugged their way out of a spell.
Their keyboards would clatter, As they brewed
up code chatter,
And their laughter echoed through each byte
cell.*

FUNNY DEEPPFAKES



UNIVERSITÉ
DE GENÈVE

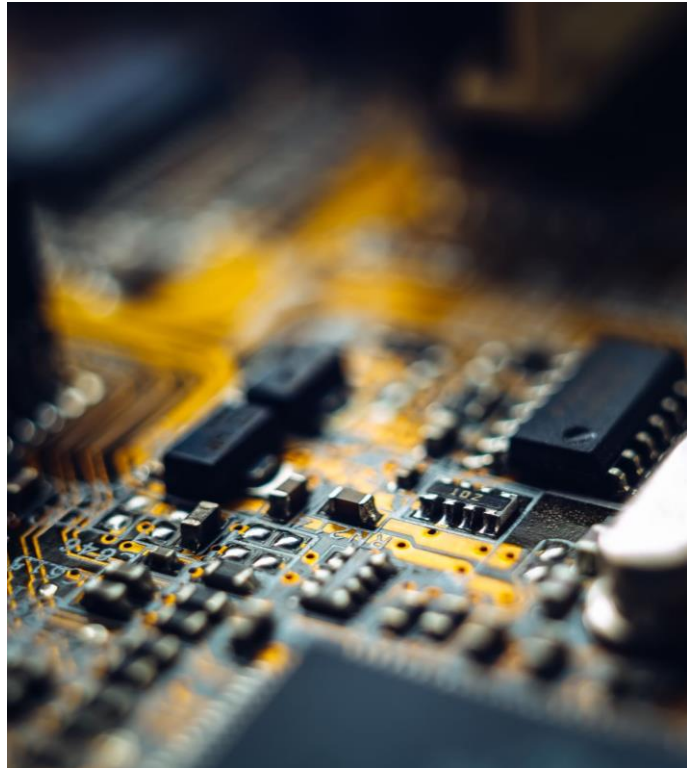


Bing Chat Copilot

HOW DEEPPFAKES WORKS



UNIVERSITÉ
DE GENÈVE



- Autoencoders
- Generation Adversarial Network (GANs)
- Face Swapping
- Face Morphing
- Lip Sync
-

WIKIPEDIA DEFINITION OF GANS



UNIVERSITÉ
DE GENÈVE

- A **generative adversarial network (GAN)** is a class of machine learning frameworks and a prominent framework for approaching generative AI.^{[1][2]} The concept was initially developed by Ian Goodfellow and his colleagues in June 2014.^[3] In a GAN, two neural networks contest with each other in the form of a zero-sum game, where one agent's gain is another agent's loss.
- Given a training set, this technique learns to generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic to human observers, having many realistic characteristics. Though originally proposed as a form of generative model for unsupervised learning, GANs have also proved useful for semi-supervised learning,^[4] fully supervised learning,^[5] and reinforcement learning.^[6]



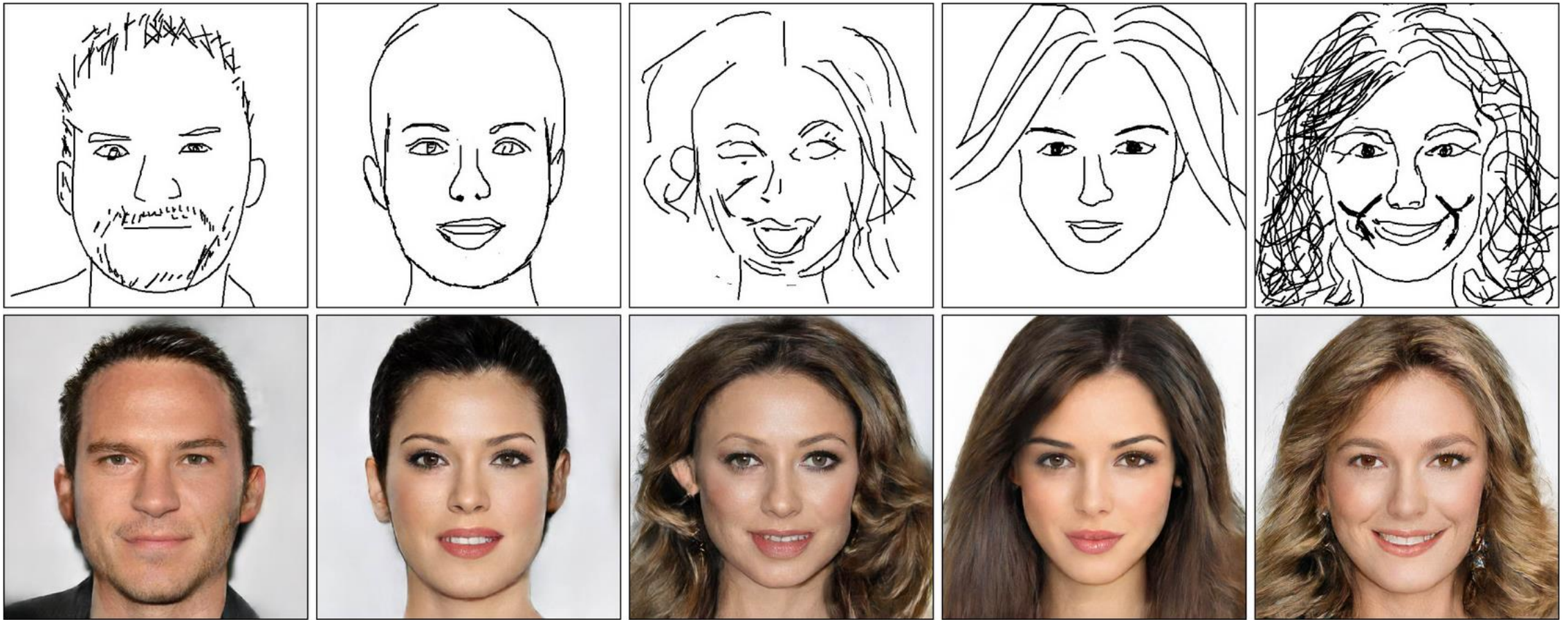


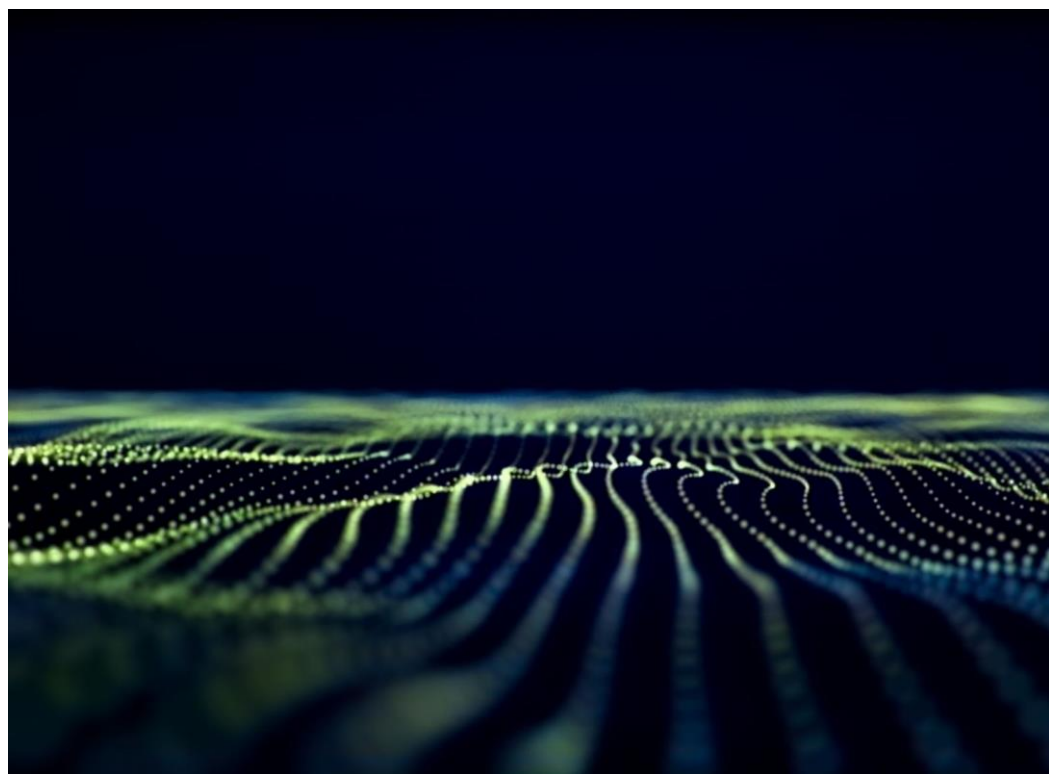
Figure: Our DeepFaceDrawing system allows users with little training in drawing to produce high-quality face images (Bottom) from rough or even incomplete freehand sketches (Top). Note that our method faithfully respects user intentions in input strokes, which serve more like soft constraints to guide image synthesis.

Source: <http://geometrylearning.com/DeepFaceDrawing/>

DEEPPFAKE DETECTIONS: TOOLS



UNIVERSITÉ
DE GENÈVE



Sentinel

Intel's Real-Time Deepfake Detector

WeVerify

Microsoft Video authenticator tool

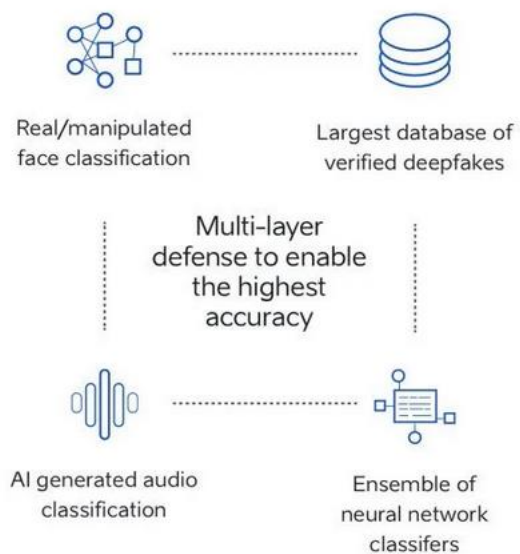
FakeCatcher

Synchro labial

ZeroGpt

AI Image Detector

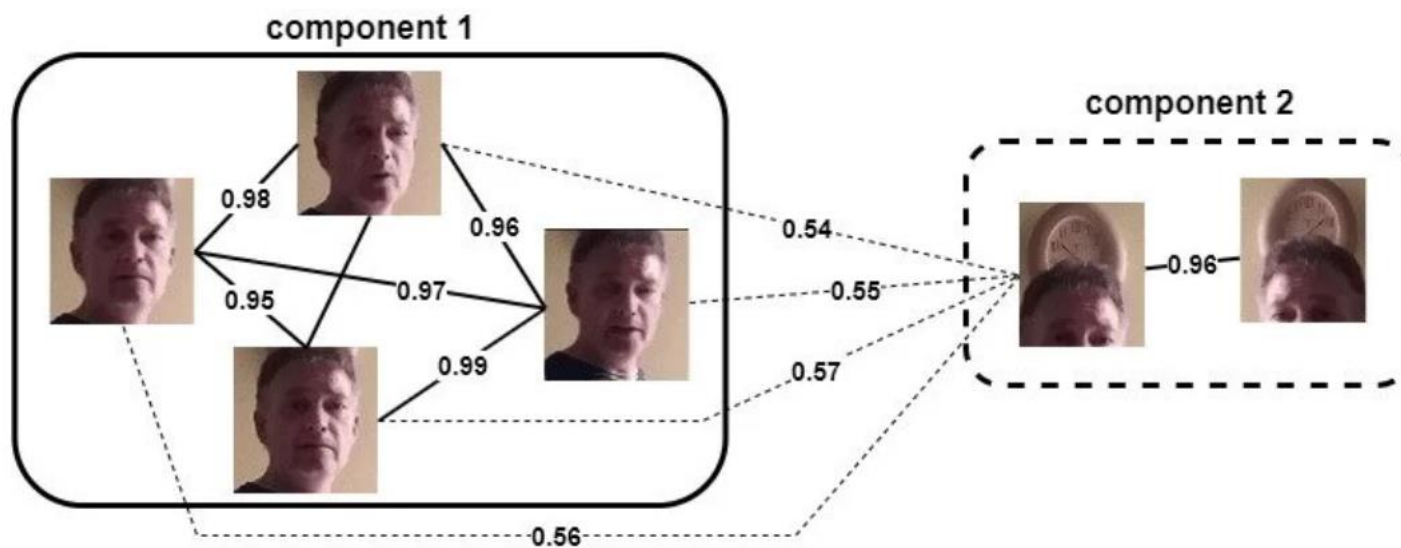
...



Our technology

Detection modelled
after cybersecurity's
standard of Defense in
Depth (DiD)

(Image: Sentinel)



(Image: WeVerify)

intelligent human-in-the-loop content verification and disinformation analysis methods and tools.

ZEROGPT



UNIVERSITÉ
DE GENÈVE



ONLINE TOOL TO DETECT IF A
TEXT WAS WRITTEN BY A HUMAN
OR AN AI SUCH CHATGPT



TRAINED ON A HUGE CORPUS OF
TEXT



SUPPORT SEVERAL LANGUAGES

<https://www.zerogpt.com/>

AI IMAGE DETECTOR



UNIVERSITÉ
DE GENÈVE



Artistic image detector



Developped by Huggingface.com



It is a proof of concept demonstration
using vision model transformer

<https://huggingface.co/spaces/umm-maybe/AI-image-detector>

<https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>

EXAMPLES AND INTERACTIVE SESSION



UNIVERSITÉ
DE GENÈVE



Let's try together and use :

- ZeroGpt: <https://www.zerogpt.com/>
- AI Image Detector:
<https://huggingface.co/spaces/umm-maybe/AI-image-detector>

LET'S TRY TOGETHER AND USE



UNIVERSITÉ
DE GENÈVE



- ZeroGpt: <https://www.zerogpt.com/>
- AI Image Detector: <https://huggingface.co/spaces/umm-maybe/AI-image-detector>



la qualité d'un homme se calcule à sa démesure; tentez, essayez, échouez même, ce sera votre réussite

Detect Text

 Upload File

101/15 000 Characters
(Get up to 100,000 [here](#))

****Please input more text for a more accurate result***

Your Text is Human written

0%

AI GPT*

open(window.clickTag))

IMAGE FROM THE WEBSITE OF SI



UNIVERSITÉ
DE GENÈVE

Furthermore the intended scope of this tool is artistic images; that is to say, it is not a deepfake photo detector, and general computer imagery (webcams, screenshots, etc.) may throw it off.

In general, this tool can only serve as one of many potential indicators that an image was AI-generated. Images scoring as very probably artificial (e.g. 90% or higher) could be referred to a human expert for further investigation, if needed.

For more information please see the blog post describing this project at: <https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>



Nettoyer

Soumettre

PLANETSOLAR: REAL IMAGE



UNIVERSITÉ
DE GENÈVE

Furthermore the intended scope of this tool is artistic images; that is to say, it is not a deepfake photo detector, and general computer imagery (webcams, screenshots, etc.) may throw it off.

In general, this tool can only serve as one of many potential indicators that an image was AI-generated. Images scoring as very probably artificial (e.g. 90% or higher) could be referred to a human expert for further investigation, if needed.

For more information please see the blog post describing this project at: <https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>



Nettoyer

Soumettre

output

artificial



PLANETSOLAR FROM A REAL IMAGE BY COPILOT



UNIVERSITÉ
DE GENÈVE

Furthermore the intended scope of this tool is artistic images; that is to say, it is not a deepfake photo detector, and general computer imagery (webcams, screenshots, etc.) may throw it off.

In general, this tool can only serve as one of many potential indicators that an image was AI-generated. Images scoring as very probably artificial (e.g. 90% or higher) could be referred to a human expert for further investigation, if needed.

For more information please see the blog post describing this project at: <https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>



Nettoyer

Soumettre




PRODUCED USING COPILOT FROM A REAL IMAGE



UNIVERSITÉ
DE GENÈVE

For more information please see the blog post describing this project at: <https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>

image



Nettoyer

Soumettre

output

human

human

artificial




Furthermore the intended scope of this tool is artistic images; that is to say, it is not a deepfake photo detector, and general computer imagery (webcams, screenshots, etc.) may throw it off.

In general, this tool can only serve as one of many potential indicators that an image was AI-generated. Images scoring as very probably artificial (e.g. 90% or higher) could be referred to a human expert for further investigation, if needed.

For more information please see the blog post describing this project at: <https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>

image



Nettoyer

Soumettre

output

human

human 96%

artificial 4%

ORIGINAL IMAGE



UNIVERSITÉ
DE GENÈVE





GENERATED FROM A TEXT PROMPT

Furthermore the intended scope of this tool is artistic images; that is to say, it is not a deepfake photo detector, and general computer imagery (webcams, screenshots, etc.) may throw it off.

In general, this tool can only serve as one of many potential indicators that an image was AI-generated. Images scoring as very probably artificial (e.g. 90% or higher) could be referred to a human expert for further investigation, if needed.

For more information please see the blog post describing this project at: <https://medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226>

image



✎ ✕

Nettoyer

Soumettre

output

artificial

artificial	81%
human	19%

QUIZ



UNIVERSITÉ
DE GENÈVE



- <https://sensity.ai/blog/deepfake-detection/deepfake-forensic-reports-in-courts/>

CONCLUSION



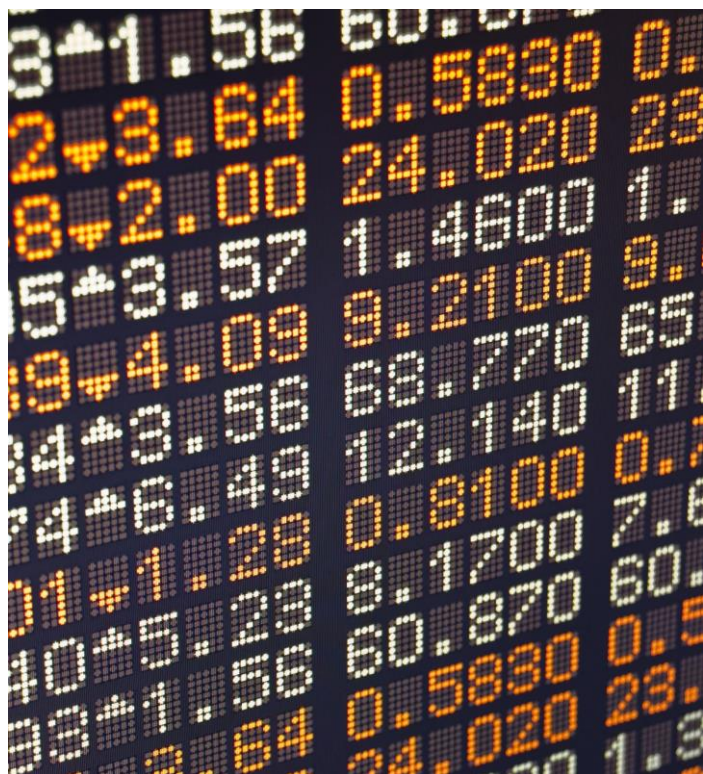
UNIVERSITÉ
DE GENÈVE

- Deepfake detection is a challenging issue technologically
- Requires strong collaboration between
 - reserachers,
 - policy makers
 - and tech compagnies
- Education and awarness for individuals and society is also important

REFERENCES



UNIVERSITÉ
DE GENÈVE



- <https://www.nature.com/articles/nature.2017.22784>
- <https://www.unite.ai/fr/g%C3%A9n%C3%A9rateurs-de-visages-ai%C3%A9atoires/>
- <https://www.unite.ai/best-deepfake-detector-tools-and-techniques/>
- <https://www.unite.ai/fr/meilleurs-outils-et-techniques-de-d%C3%A9tection-de-deepfake/>
- https://www.nature.com/articles/d41586-023-03479-4?utm_source=Live+Audience&utm_campaign=9389b5cdb7-briefing-ai-20231121&utm_medium=email&utm_term=0_b27a691814-9389b5cdb7-50201148&mc_cid=9389b5cdb7&mc_eid=0ff5adf963

THE TEAM



UNIVERSITÉ
DE GENÈVE



**Prof. Giovanna Di Marzo
Serugendo**
Director
Centre Universitaire d'Informatique
Digital Innovation Hub



Dr Lamia Friha
DISTIC
Manager Accelerator Digital Science
and Services
Manager R&D unit
Digital Forge



Dr Pierre-Yves Burgi
Director – DISTIC
Accelerator Digital Science and
Services
R&D unit



Alain Hugentobler
DISTIC
Expert Cybersecurity
Accelerator Digital Science and
Services, Digital Forge
R&D unit

THANK YOU



UNIVERSITÉ
DE GENÈVE



Bing Chat Copilot

QUESTIONS



UNIVERSITÉ
DE GENÈVE

