ELTE | PPK
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

DEZINFORMÁCIÓ ÉS MESTERSÉGES INTELLIGENCIA KUTATÓCSOPORT

Mesterséges intelligencia és dezinformáció – utópiák és disztópiák

Krekó Péter, PhD.

Habilitált egyetemi docens, ELTE PPK Szociálpszichológia Tanszék/
Dezinformáció és Mesterséges Intelligencia Kutatócsoport

„Vissza a jövőbe!”- A közösségi média és az AI hatása a történelemre
Történelemtanárok Egylete, 2024. október 5.

Fő kérdések

- Hogyan hat a dezinformáció előállításával kapcsolatos technológiai fejlődés az információ kognitív feldolgozására, az érzelmi reakciókra?
- Milyen releváns pszichológiai vizsgálatok vannak a területen?
- Milyen gyakorlati jelentősége van mindennek?
- Milyen irányba fordíthatja ez a történelmet?

Miről lesz szó?

1) Előállítás

2) Terjesztés

3) Detekció/Prevenció

+ Következtetések

People: fearing AI takeover
AI:

‘Sapiens’ author says AI is an alien threat that could wipe us out: ‘Instead of coming from outer space, it’s coming from California’

ELTE | PPK
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

DEZINFORMÁCIÓ ÉS MESTERSÉGES INTELLIGENCIA KUTATÓCSOPORT

1) A tartalom előállítása

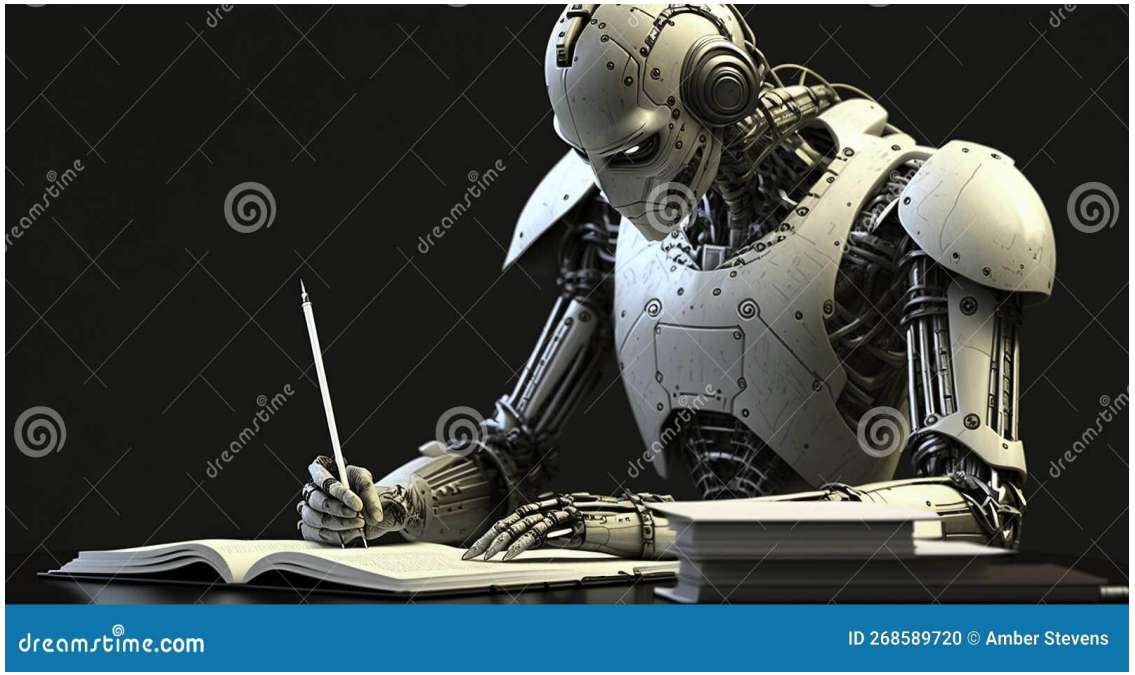
A generatív AI sokszor meggyőzőbben csap be minket

Hitelesebbnek fogadjuk el a GPT-3, mint az emberek által írt (dez) információt (Spitale et al, 2023)

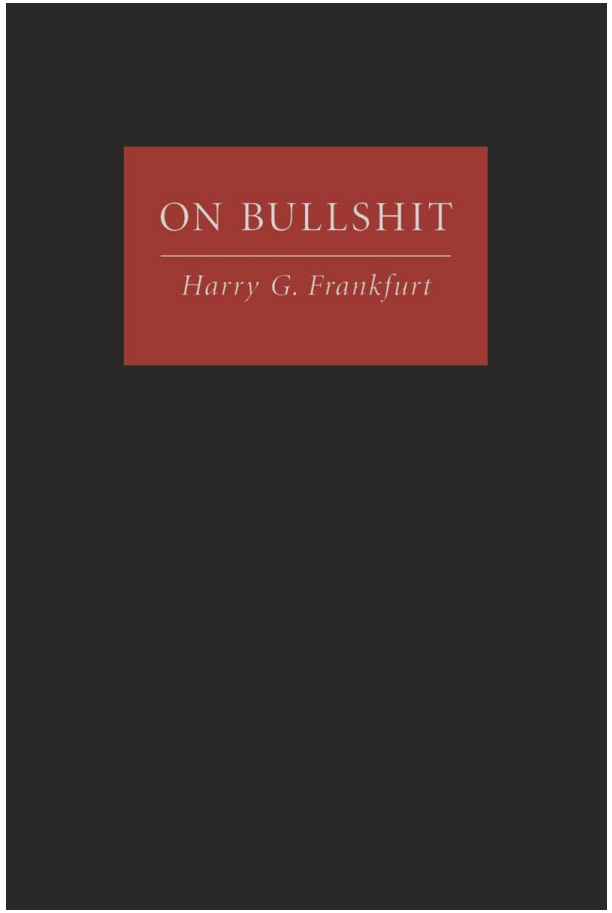
A GPT-3 által írt twitter-üzenetek az emberek számára nem megkülönböztethetőek a személyek által írtaktól (u.o.)

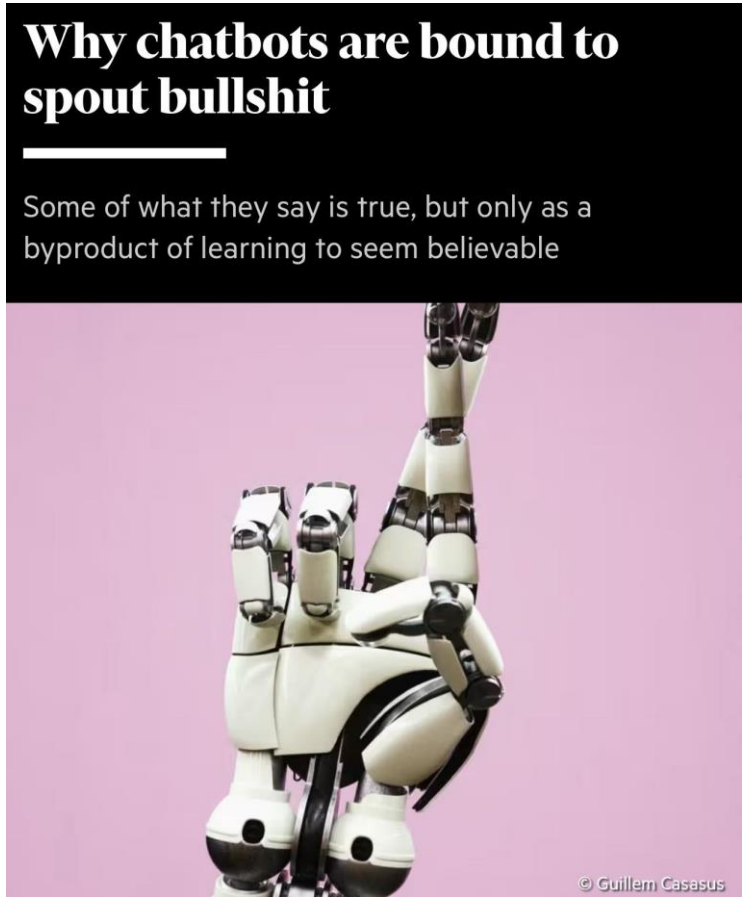
A generatív AI által létrehozott üzenetek hatékonyan meggyőzik a befogadókat sokféle politikai kérdésben (Bai et al., preprint).

Az AI által generált üzeneteket tényszerűbbnek, logikusabbnak, DE kevésbé dühösnek, egyedinek, és kevésbé „sztorizósnak” ítélték a kérdezettek (u.o.).



A bullshit-problematika (misinformation)





Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking

Gordon Pennycook[1] | **David G. Rand**[2,3]

[1]Hill/Levene Schools of Business, University of Regina, Regina, Canada
[2]Sloan School, Massachusetts Institute of Technology, Cambridge, Massachussetts
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachussetts

Correspondence
Gordon Pennycook, Hill/Levene Schools of Business, University of Regina. 3737 Wascana Pky, Regina, Saskatchewan, Canada, S4S 0A2
Email: gordon.pennycook@uregina.ca

Funding information
Social Sciences and Humanities Research Council of Canada; Miami Foundation

Abstract

Objective: Fake news represents a particularly egregious and direct avenue by which inaccurate beliefs have been propagated via social media. We investigate the psychological profile of individuals who fall prey to fake news.

Method: We recruited 1,606 participants from Amazon's Mechanical Turk for three online surveys.

Results: The tendency to ascribe profundity to randomly generated sentences—pseudo-profound bullshit receptivity—correlates positively with perceptions of fake news accuracy, and negatively with the ability to differentiate between fake and real news (media truth discernment). Relatedly, individuals who overclaim their level of knowledge also judge fake news to be more accurate. We also extend previous research indicating that analytic thinking correlates negatively with perceived accuracy by showing that this relationship is not moderated by the presence/absence of the headline's source (which has no effect on accuracy), or by familiarity with the headlines (which correlates positively with perceived accuracy of fake and real news).

Conclusion: Our results suggest that belief in fake news may be driven, to some extent, by a general tendency to be overly accepting of weak claims. This tendency, which we refer to as reflexive open-mindedness, may be partly responsible for the prevalence of epistemically suspect beliefs writ large.

KEYWORDS
analytic thinking, bullshit receptivity, fake news, news media, social media

Pseudo-profound bullshit skála

Rezgünk, rezgünk, újjászületünk. Elektromágneses rezonanciaként létezünk.

Az egymásrautaltság elmúló cselekedetekben gyökerezik.
Az örök nyugalom újjászületik a végtelen emberi megfigyelésben.
A végtelen a karmikus valóság tükröződése.
A rejtett jelentés a csodák szimbolikus ábrázolásától függ.
A transzcendencia átalakítja a kiválóság kapuját.
Egy átalakulás érző ébredésének közepén vagyunk, amely összehangol minket magával a bioszférával.
A hálózat közeledik a fordulóponthoz. A kozmikus tudás ébredése lesz a jövő.
Nem engedhetjük meg többé magunknak, hogy megszakítással éljünk.
A megszakítás a belátás ellentéte.
A tudat a kvantumenergia morfikus rezonanciájából áll.
A kvantum az erő ébredését jelenti


uilt entirely with AI imagery
What if
INTERNATIONAL TENSIONS ESCALATE


Lord Bebo
@MyLordBebo
BOMBSHELL: Israel runs ads for Ukrainian soldiers to join the Israeli army!
They offer a career, to move the family and also get a citizenship!
-> Fight dudes with RPGs and AKs or sit in a trench and get shelled 24/7?
How many will follow this call ... to sunny israel!
Hey guys, recently I can
back from the front
Skip Ad
0:02 / 0:36
9:13 PM · Nov 1, 2023 · 282K Views

Home News Research Fellowships Leadership
Digital News Report

Home > Research

Types, sources, and claims of COVID-19 misinformation

A man wearing a face mask uses his phone in the Times Square subway station. REUTERS/Lucas Jackson

Dr. J. Scott Brennen · Felix Simon · Dr Philip N. Howard · Prof. Rasmus Kleis Nielsen
Tuesday 7 April 2020

Russia in South Africa 🇷🇺
@EmbassyofRussia

Today is 22nd anniversary of the day when NATO brought death & destruction to #Yugoslavia by attacking a sovereign state in total disregard for intl law. NATO's 78 days long "humanitarian intervention" took lives of 2000 innocent people, turned thousands more into refugees

Bejegyzés lefordítása

MFA Russia 🇷🇺 és 8 másik felhasználó

Israel-Hamas war: Beware of the cheap fake videos on the Internet

Two misinformation experts explain why and how you can develop the power to resist these deceptions.

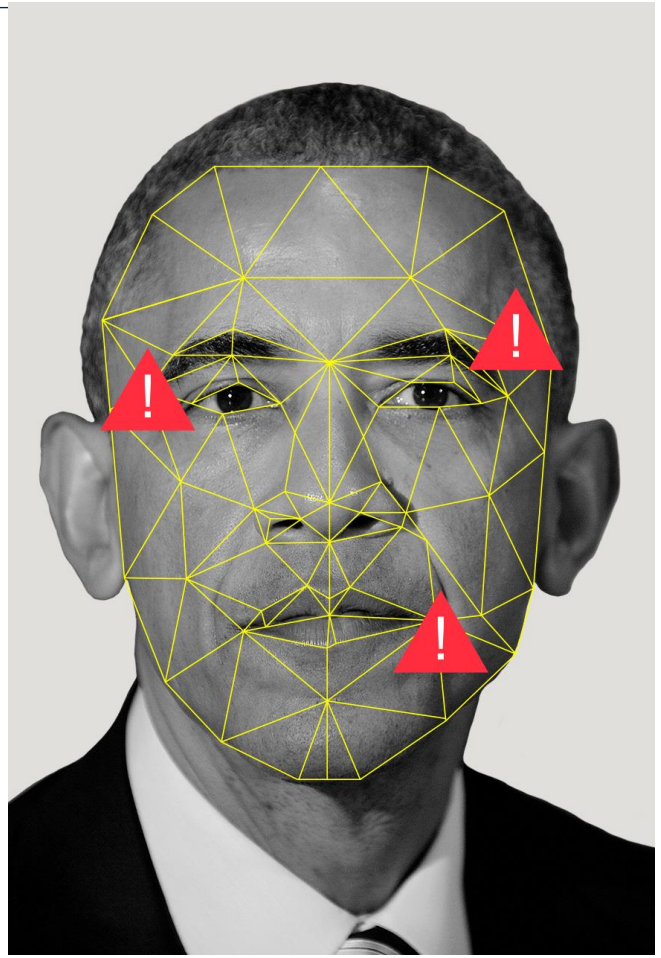
Sam Wineburg and Michael Caulfield

A makeshift operating theater area inside Al-Shifa hospital during the Israeli ground operation around the hospital, in Gaza City, on Nov 12.
PHOTO: REUTERS

UPDATED 2023. NOV. 18. DE. 6:04 SGT

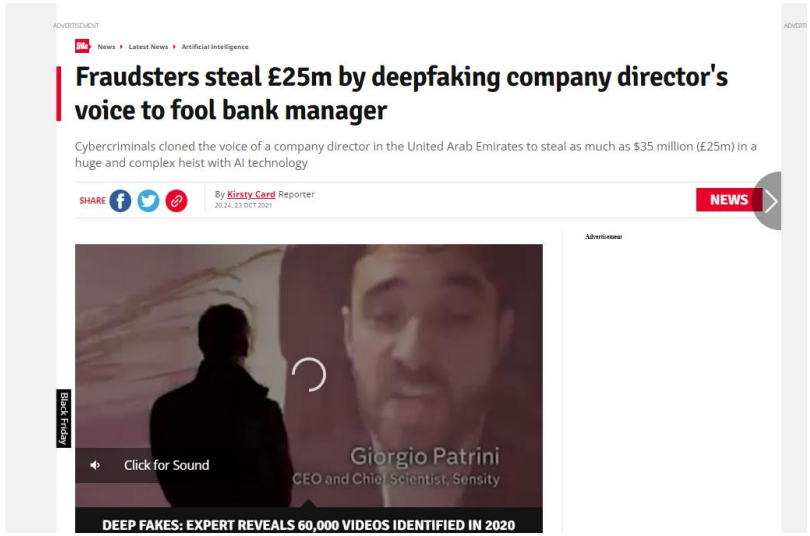
Deepfake-ek

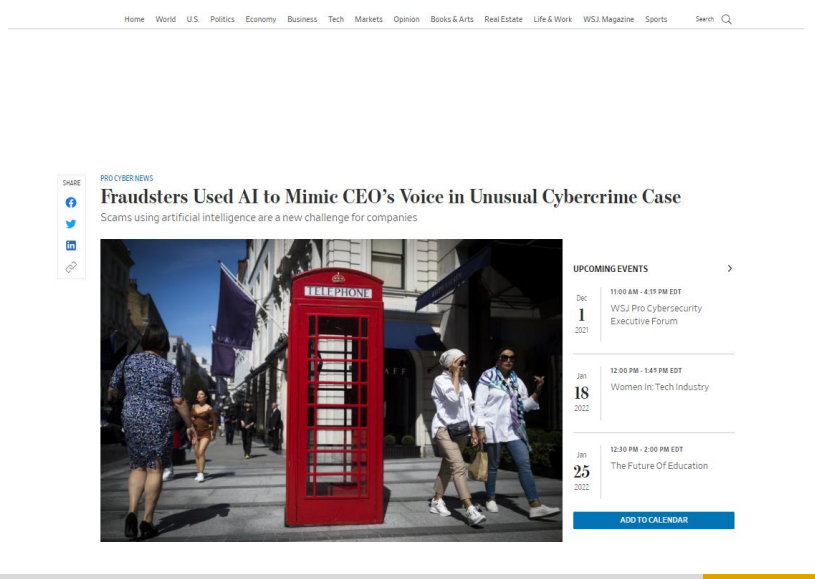
- Az álhírek „nehézfegyverzete” (Veszelszki, 2020)
- A “deep learning” és “fake” szavak kombinációja (+deep voice)
- Ultrarealisztikus, gépi tanulással, azon belül mély neurális hálózatok felhasználásával létrehozott hamis audiovizuális tartalom
- Valós képi és hangzó anyag felhasználásával készül
- Sokszor valós háttéren, csak minimális módosítással
- **Erősebb érzelmi reakció kiváltására alkalmas** (Arsenault, 2020)
- Az emberek hajlamosak silány minőségű valótlan tartalmakat is elhinni, ha azt saját pártjuktól származik (Whyte, 2020).
- DE analitikus képességek és a politikai érdeklődés segít a felismerésükben (Appel, Prietzel, 2023)



A deepfake veszélyei: a valóra vált post-truth disztópia?

- 2030-ra várhatóan szabad szemmel detektálhatatlanná, és mindenki számára előállíthatóvá válik ez a technológia
- Az emberek nem tudják megkülönböztetni a valóságostól: „csak a szememnek hiszek”
- Hatékonyabbak a hagyományos fake-eknél, akkor is hat, ha tudjuk hogy hamis (!)
- Alkalmazásai:
  - Személyes bosszú/cyberbullying; Személyes erőszak: pl. Whatsapp-lincselések Indiában
  - Politikai: választások befolyásolása, diplomáciai konfliktusok és fegyveres konfliktusok szítása,
  - Jogi: megtévesztő marketingkampányok, bizonyíték jogi eljárásban
  - Gazdasági: megtévesztő marketingkampányok, csalások, vállalatok elleni fake news kampányok





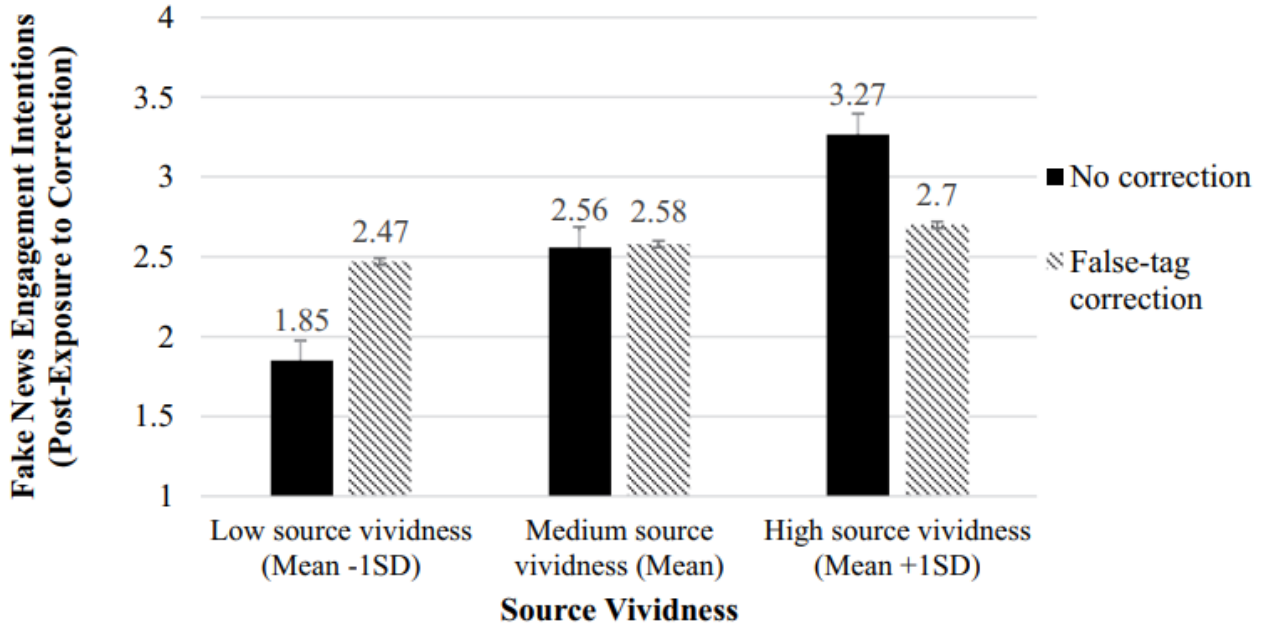
A Deepfake és az episztemikus bizonytalanság

A jelenség a megvezetés új lehetőségeit kínálja: a hamis információt állítja be igaznak vagy pedig a hiteles információt hamisnak állíthatjuk be (Bontridder & Poullet, 2021).

A deepfake-ek veszélyével való riogatás nem javítja a valósi és manipulált videok elkülönítésének képességét, csak a szkepszist erősíti (Ternovski et al., 2021).

To cite this article: Jiyoung Lee & Soo Yun Shin (2021): Something that They Never Said: Multimodal Disinformation and Source Vividness in Understanding the Power of AI-Enabled Deepfake News, Media Psychology, DOI: 10.1080/15213269.2021.2007489

To link to this article: https://doi.org/10.1080/15213269.2021.2007489



igure 1. Interaction between source vividness and false-tag correction (vs. no correction) on fake news engagement intentions tudy 2; abortion). Error bars are 95% confidence intervals.

ELTE | PPK
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

DEZINFORMÁCIÓ ÉS MESTERSÉGES INTELLIGENCIA KUTATÓCSOPORT

2) Tartalom terjesztése

A dezinformáció terjesztése

Demográfiai, és pszichológiai profilírozás (pl. Cambridge Analytica; Youyou, 2015).

Szegmentálás és mikrotargetálás (pl. (Dobber et al., 2020). hatékonyabb visszhangkamrába zárás, intellektuális izoláció, naiv szociológia,

Szociális botok:

- 2016-os amerikai elnökválasztási kampány utolsó hónapjában a Twitteren megjelenő posztok egyötödét botok generálták és terjesztették el (Masood és mtsai., 2023; Sebastian, 2023).
- A hiteltelen források hatásának növelésében rendkívül hatásosak lehetnek a social bot-ok, az emberek pedig gyakran nem azonosítják őket, és megosztják/vitába szállnak velük (Shao et al., 2018).

3) Az álhírek detekciója/azonosítása/cáfolata

ELTE | PPK
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

DEZINFORMÁCIÓ ÉS MESTERSÉGES INTELLIGENCIA KUTATÓCSOPORT

3) Az álhírek detekciója/cáfolata

Stratégiák a dezinformáció elleni fellépésben (Krekó, 2020)

	Dezinformációval való találkozás előtt (Prevenció)	Dezinformációval való találkozás után (utólagos kezelés)
Kínálati oldalt célzó beavatkozások	**Megelőző csapás** (pl. eltávolítás)	**Válaszcsapás** (pl. Álhírek címkézése, a forrás hiteltelenítése)
Keresleti oldalt célzó beavatkozások	**Immunizálás** (pl. előzetes figyelmeztetés, beoltás, az analitikus készség erősítése)	**Gyógyítás** („Kognitív infiltráció” Cáfolat/fact-checking)

Leveraging ChatGPT for Efficient Fact-Checking

Emma Hoes[†] Sacha Altay[‡] and Juan Bermeo[§]

May 29, 2023

Abstract

We explore the potential of automated online content moderation through Large Language Models (LLMs) as a remedy to the overwhelming surge of online content that fact-checking organizations cannot verify. Although LLMs, such as ChatGPT, may exacerbate this problem by facilitating content production, they could also be leveraged to enhance the efficiency and expediency of fact-checking processes. We conducted a systematic evaluation to measure ChatGPT's fact-checking performance by submitting 21,152 fact-checked statements to ChatGPT as a zero-shot classification. We find that ChatGPT is able to accurately categorize statements as true or false in 69% of cases. Addressing memorization, the performance of ChatGPT is similar on claims that have not been fact-checked or are post its training data cutoff-date. These findings demonstrate the potential of ChatGPT to help label misinformation and deepen our comprehension of how LLMs could improve content moderation practices, complementing the crucial work of human fact-checking experts in upholding the accuracy of information dissemination.

Keywords—- Misinformation, Fact-checking, Artificial Intelligence, AI, Fake News, ChatGPT, LLMs, Content Moderation, PolitiFact

Word count: 3353

Pseudoscience Detection Using a Pre-trained Transformer Model with Intelligent ReLabeling

Thilina C. Rajapakse and Ruwan D. Nawarathna

Abstract Often dismissed as a harmless pastime for the gullible, pseudoscience nonetheless has devastating effects, including the loss of life. Its menace is amplified by the difficulty of differentiating pseudoscience from science, especially for the untrained eye. A novel method recognizes pseudoscience in the text by utilizing a fine-tuned Robustly Optimized Bidirectional Encoder Representation from Transformers Approach (RoBERTa) model. The dataset of 112,720 full-text articles used in this work is made publicly available to remedy the lack of datasets related to pseudoscience. A novel technique, Intelligent ReLabeling (IRL), is employed to minimize mislabeled data, enabling the rapid creation of high-quality textual datasets. IRL eliminates the need for expensive manual verification processes and minimizes domain expertise requirements in many applications. The final model trained with IRL achieves an F_1 score of 0.929 on a separate manually labeled test dataset.

Keywords Natural language processing · Classification · Data cleaning · Pseudoscience dataset

TABLE II
SUMMARY OF PROMINENT DEEPFAKE DETECTION METHODS

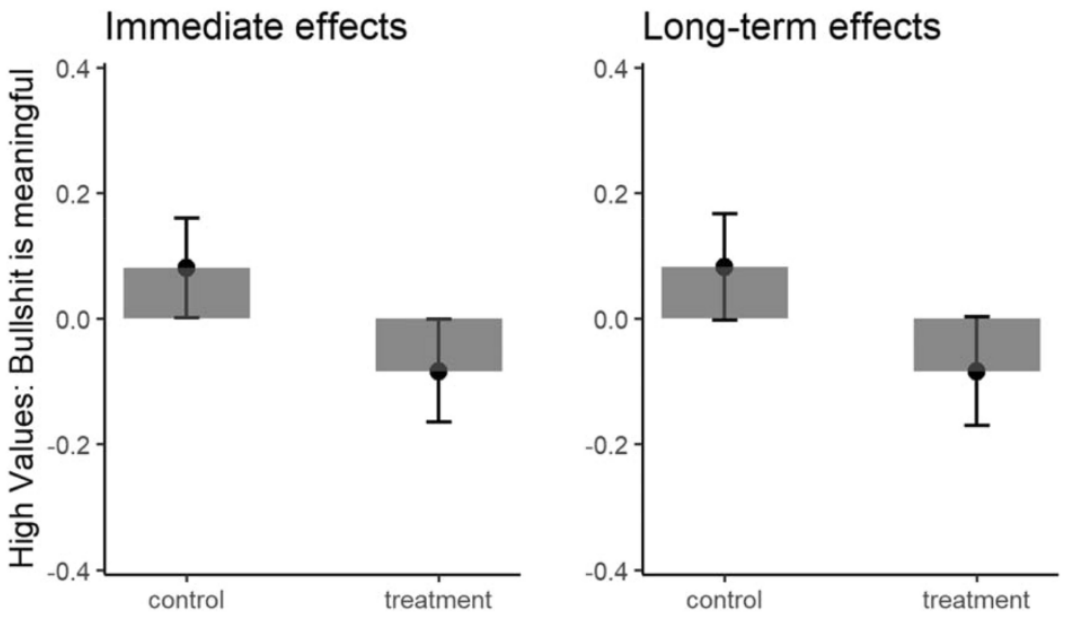
Methods	Classifiers/ Techniques	Key Features	Dealing with	Datasets Used
Eye blinking [99]	LRCN	- Use LRCN to learn the temporal patterns of eye blinking. - Based on the observation that blinking frequency of deepfakes is much smaller than normal.	Videos	Consist of 49 interview and presentation videos, and their corresponding generated deepfakes.
Intra-frame and temporal inconsistencies [98]	CNN and LSTM	CNN is employed to extract frame-level features, which are distributed to LSTM to construct sequence descriptor useful for classification.	Videos	A collection of 600 videos obtained from multiple websites.
Using face warping artifacts [103]	VGG16 [101] ResNet50, 101 or 152 [102]	Artifacts are discovered using CNN models based on resolution inconsistency between the warped face area and the surrounding context.	Videos	- UADFV [104], containing 49 real videos and 49 fake videos with 32752 frames in total. - DeepfakeTIMIT [64]
MesoNet [94]	CNN	- Two deep networks, i.e. Meso-4 and MesoInception-4 are introduced to examine deepfake videos at the mesoscopic analysis level. - Accuracy obtained on deepfake and FaceForensics datasets are 98% and 95% respectively.	Videos	Two datasets: deepfake one constituted from online videos and the FaceForensics one created by the Face2Face approach [111].
Capsule-forensics [106]	Capsule networks	- Latent features extracted by VGG-19 network [101] are fed into the capsule network for classification. - A dynamic routing algorithm [108] is used to route the outputs of three convolutional capsules to two output capsules, one for fake and another for real images, through a number of iterations.	Videos/ Images	Four datasets: the Idiap Research Institute replay-attack [109], deepfake face swapping by [94], facial reenactment FaceForensics [110], and fully computer-generated image set using [112].
Head poses [104]	SVM	- Features are extracted using 68 landmarks of the face region. - Use SVM to classify using the extracted features.	Videos/ Images	- UADFV consists of 49 deep fake videos and their respective real videos. - 241 real images and 252 deep fake images from DARPA MediFor GAN Image/Video Challenge.
Eye, teach and facial texture [114]	Logistic regression and neural network	- Exploit facial texture differences, and missing reflections and details in eye and teeth areas of deepfakes. - Logistic regression and neural network are used for classifying.	Videos	A video dataset downloaded from YouTube.
Spatio-temporal features with RCN [95]	RCN	Temporal discrepancies across frames are explored using RCN that integrates convolutional network DenseNet [79] and the gated recurrent unit cells [96]	Videos	FaceForensics++ dataset, including 1,000 videos [97].
Spatio-temporal features with LSTM [140]	Convolutional bidirectional recurrent LSTM network	- An XceptionNet CNN is used for facial feature extraction while audio embeddings are obtained by stacking multiple convolution modules. - Two loss functions, i.e. cross-entropy and Kullback-Leibler divergence, are used.	Videos	FaceForensics++ [97] and Celeb-DF (5,639 deep-fake videos) [141] datasets and the ASVSpoof 2019 Logical Access audio dataset [142].
Analysis of PRNU [115]	PRNU	- Analysis of noise patterns of light sensitive sensors of digital cameras due to their factory defects. - Explore the differences of PRNU patterns be-tween the authentic and deepfake videos because face swapping is believed to alter the local PRNU patterns.	Videos	Created by the authors, including 10 authentic and 16 deepfake videos using DeepFaceLab [34].
Phoneme-viseme mismatches [133]	CNN	- Exploit the mismatches between the dynamics of the mouth shape, i.e. visemes, with a spoken phoneme. - Focus on sounds associated with the M, B and P phonemes as they require complete mouth closure while deepfakes often incorrectly synthesize it.	Videos	Four in-the-wild lip-sync deepfakes from Instagram and YouTube (www.instagram.com/bill_posters_uk and youtu.be/VWMEDacz3L4) and others are cre-ated using synthesis techniques, i.e. Audio-to-Video (A2V) [63] and Text-to-Video (T2V) [134].

TABLE I
SUMMARY OF NOTABLE DEEPFAKE TOOLS

Tools	Links	Key Features
Faceswap	https://github.com/deepfakes/faceswap	- Using two encoder-decoder pairs. - Parameters of the encoder are shared.
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN	Adversarial loss and perceptual loss (VGGface) are added to an auto-encoder architecture.
Few-Shot Face Translation	https://github.com/shaoanlu/fewshot-face-translation-GAN	- Use a pre-trained face recognition model to extract latent embeddings for GAN processing. - Incorporate semantic priors obtained by modules from FUNIT [42] and SPADE [43].
DeepFaceLab	https://github.com/iperov/DeepFaceLab	- Expand from the Faceswap method with new models, e.g. H64, H128, LIAEF128, SAE [44]. - Support multiple face extraction modes, e.g. S3FD, MTCNN, dlib, or manual [44].
DFaker	https://github.com/dfaker/df	- DSSIM loss function [45] is used to reconstruct face. - Implemented based on Keras library.
DeepFake_tf	https://github.com/StromWine/DeepFake_tf	Similar to DFaker but implemented based on tensorflow.
AvatarMe	https://github.com/lattas/AvatarMe	- Reconstruct 3D faces from arbitrary "in-the-wild" images. - Can reconstruct authentic 4K by 6K-resolution 3D faces from a single low-resolution image [46].
MarioNETte	https://hyperconnect.github.io/MarioNETte	- A few-shot face reenactment framework that preserves the target identity. - No additional fine-tuning phase is needed for identity adaptation [47].
DiscoFaceGAN	https://github.com/microsoft/DiscoFaceGAN	- Generate face images of virtual people with independent latent variables of identity, expression, pose, and illumination. - Embed 3D priors into adversarial learning [48].
StyleRig	https://gvv.mpi-inf.mpg.de/projects/StyleRig	- Create portrait images of faces with a rig-like control over a pretrained and fixed StyleGAN via 3D morphable face models. - Self-supervised without manual annotations [49].
FaceShifter	https://lingzhili.com/FaceShifterPage	- Face swapping in high-fidelity by exploiting and integrating the target attributes. - Can be applied to any new face pairs without requiring subject specific training [50].
FSGAN	https://github.com/YuvalNirkin/fsgan	- A face swapping and reenactment model that can be applied to pairs of faces without requiring training on those faces. - Adjust to both pose and expression variations [51].
Transformable Bottleneck Networks	https://github.com/kyleolsz/TB-Networks	- A method for fine-grained 3D manipulation of image content. - Apply spatial transformations in CNN models using a transformable bottleneck framework [52].
"Do as I Do" Motion Transfer	github.com/carolineec/EverybodyDanceNow	- Automatically transfer the motion from a source to a target person by learning a video-to-video translation. - Can create a motion-synchronized dancing video with multiple subjects [53].
Neural Voice Puppetry	https://justusthies.github.io/posts/neural-voice-puppetry	- A method for audio-driven facial video synthesis. - Synthesize videos of a talking head from an audio sequence of another person using 3D face representation. [54].

Figure 7

From: A prosocial fake news intervention with durable effects



Bullshit receptivity immediately after the intervention (left panel) and 4 weeks later (right panel). Error bars represent standard errors. The y-axis represents standardized scores in terms of bullshit receptivity ratings controlled for pre-intervention bullshit receptivity ratings. Higher scores indicate beliefs in bullshit are meaningful such as "interdependence is rooted in ephemeral actions".

A prosocial fake news intervention with durable effects

Gábor Orosz, Benedek Paskuj, Laura Faragó & Péter Krekó

Scientific Reports **13**, Article number: 3958 (2023) | Cite this article

Abstract

The present online intervention promoted family-based prosocial values—in terms of helping family members—among young adults to build resistance against fake news. This preregistered randomized controlled trial study is among the first psychological fake news interventions in Eastern Europe, where the free press is weak and state-sponsored misinformation runs riot in mainstream media. In this intervention, participants were endowed with an expert role and requested to write a letter to their digitally less competent relatives explaining six strategies that help fake news recognition. Compared to the active control group there was an immediate effect (d = 0.32) that persisted until the follow-up four weeks later (d = 0.22) on fake news accuracy ratings of the young, advice-giving participants. The intervention also reduced the bullshit receptivity of participants both immediately after the intervention and in the long run. The present work demonstrates the power of using relevant social bonds for motivating behavior change among Eastern European participants. Our prosocial approach with its robust grounding in human psychology might complement prior interventions in the fight against misinformation.

This survey is part of a MIT scientific research project. Your decision to complete this survey is voluntary. Please do not provide any sensitive identifiable information that would allow your identity to readily be ascertained, directly or through identifiers provided in your responses. This study is investigating how humans and artificial intelligence algorithms interact. In the study, you will answer questions and have a back and forth discussion with an artificial intelligence algorithm. If you give us permission by saying yes below, we plan to discuss/publish the results in an academic forum. In any publication, information will be provided in such a way that you cannot be identified. Only members of the research team will have access to the original data set. Before the data is shared outside the research team, any potentially identifying information will be removed. Once identifying data has been removed, the data may be used by the research team, or shared with other researchers, for both related and unrelated research purposes in the future. Your anonymized data may also be made available in online data repositories such as the Open Science Framework, which allow other researchers and interested parties to use the data for further analysis. **Clicking on the button below indicates that you are at least 18 years of age and agree to complete this survey voluntarily.**

DETEKTO TANANYAGOK

ÁLHÍREK

Mik azok az álhírek és miért vagyunk rájuk fogékonyak...

1

Mik azok az “álhírek”?

2

Miért vagyunk fogékonyak az álhírekre?

3

Hogyan készülnek és terjednek a kamu képek és videók?

PORNÓTÓL POLITIKÁIG: A DEEPFAKE ITT VAN VELÜNK

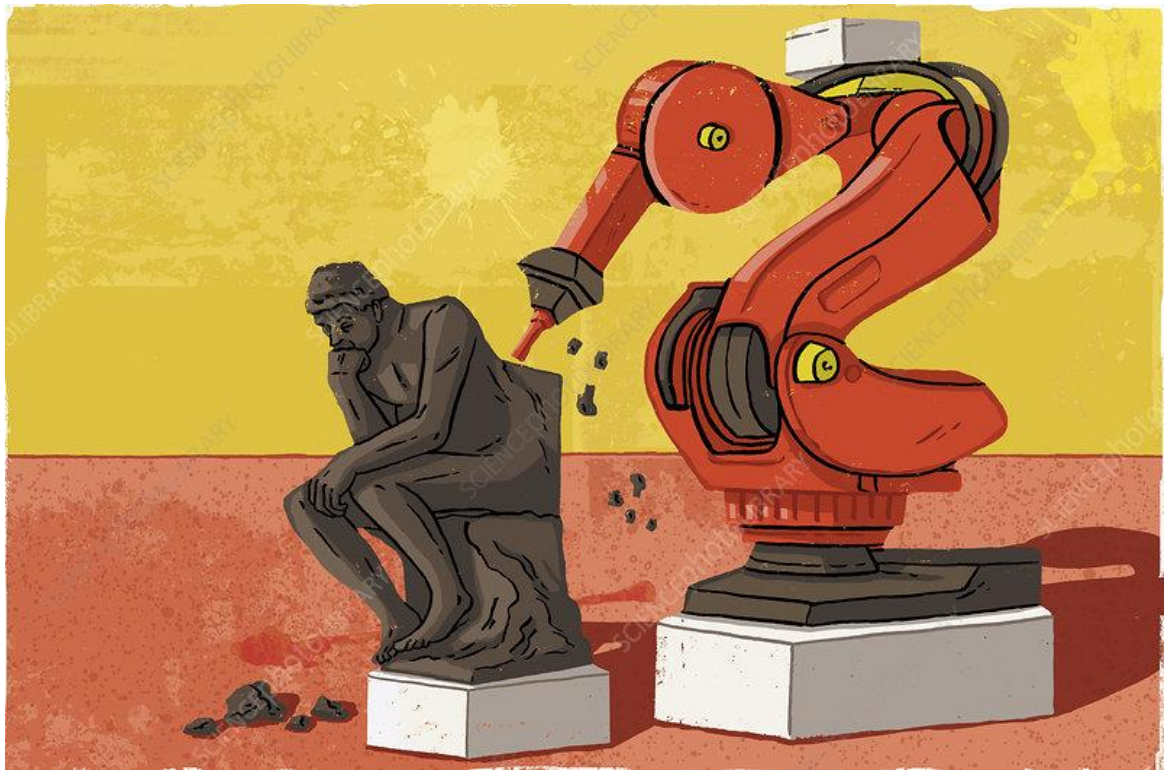
2022/06/30 - Lakmusz

Diószegi-Horváth Nóra, Német Szilvi

Mi a közös Scarlett Johanssonban, Leia hercegnőben és Vitalij Klicskóban? Mára ez nem egy közepesen rossz vicc kezdőkérdése, hanem a valóság: mindhármukól készült már élethű hamisított videó. Hol tart a deepfake, és hogyan szúrhatod ki a hamisítványt?

Következtetések: főbb dilemmák, kihívások

- Demokratikus kihívás: pluralizmus helyett fragmentáció a poitikai közösségben
- Információs tanult tehetetlenség (Nisbet&Kamenchuk, 2021); információs apokalipszis, valóság-apátia (Gregor & Mlejnková, 2021).
- Megerősítési torzítás és a morális érzelmek felkorbácsolása
- Emberkép – homo algorithmus. Szabad akarat, szabad választás?
- A kevesebb több? (digitális lábnyom, információfeldolgozás)
- Sem utópia, sem disztópia?



ELTE | PPK
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

DEZINFORMÁCIÓ ÉS MESTERSÉGES INTELLIGENCIA KUTATÓCSOPORT

Köszönöm a figyelmet!

Kreko.peter@ppk.elte.hu