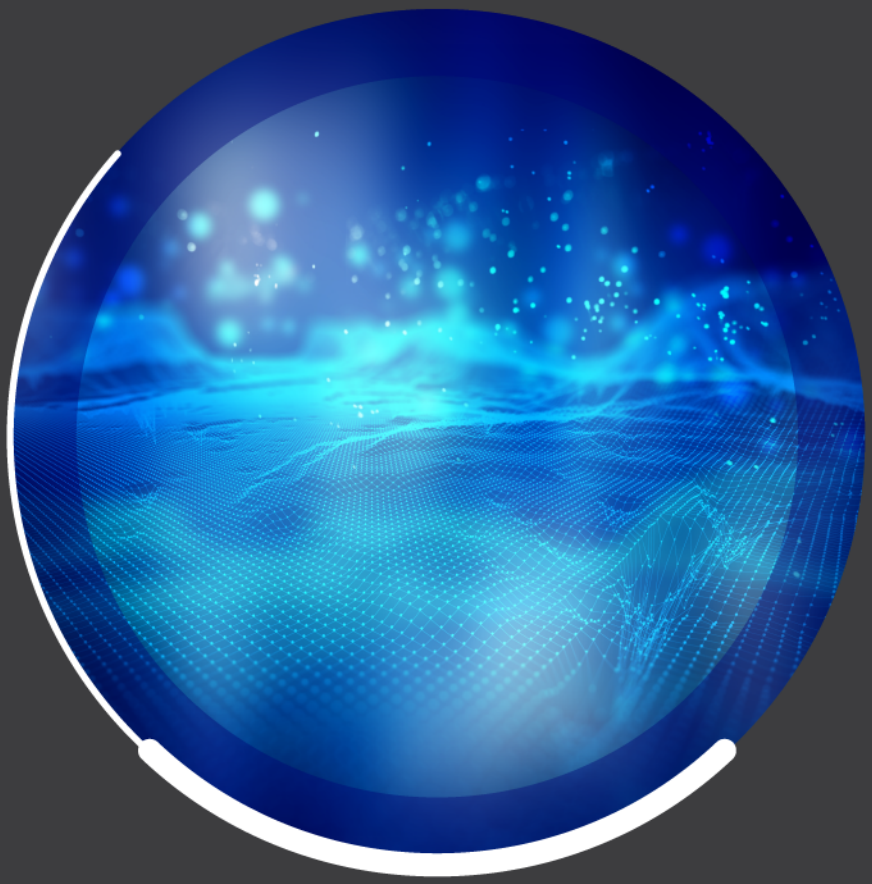
CRC | CYFLUENCE RESEARCH CENTER



22-08-2024 Research

The Rise of AI and What It Means in the Realm of Digital Influence.

Author: Florian Frank

www.cyfluence-research.com

CRC | CYFLUENCE RESEARCH CENTER

The rise of AI and what it means in the realm of digital influence.

Author: **Florian Frank**

In recent years, AIs have achieved a level of sophistication that was previously unimaginable within such a short time frame. AI is now utilized in virtually every field, from building webpages and creating texts to intelligence gathering, design, and medicine. The impact of AI on the economy and society is profound and far-reaching.

Although it›s crucial to highlight the numerous benefits AIs bring to the table, this article will focus on its darker aspects. Much of the AI discourse is plagued by hysteria, fears of sudden massive unemployment, or scenarios in which an AI decides the world would be better off without humans. While some of these concerns are legitimate, they often overshadow the genuine dangers that AIs are confronting us with today. This article will explore some of the most recent developments in AI and their implementation in hostile influence campaigns.

Throughout the article, I have included quotes from ChatGPT, as it seems befitting to let an AI speak for itself on the issues concerning its societal impact. But before we delve into the dangers and potential solutions, let›s examine how we arrived at this point.

The Journey so far

In the digital age, the terms "disinformation" and "fake news" have become prevalent, but they are often used interchangeably andinaccurately. Understanding the precise definitions of these terms is crucial for identifying and combating the spread of false information.

"Long before technological advances made self-moving devices possible, ideas about creating artificial life and robots were explored in ancient myths." *Adrienne Mayor*[1]

The idea of artificial intelligence can be traced back to as early as 2,700 years ago, long before there was any technical capability to create anything akin to modern-day machinery. As the Stanford University researcher Adrienne Mayor wrote in her work "Gods and Robots: Myths, Machines, and Ancient Dreams of Technology," the poet Hesiod (ca. 700 BC) already describes several examples of what could be viewed as early concepts of artificial intelligence[2]. There is Talos, a "man" made of bronze, whose task was to protect Europa, the daughter of Zeus, by circling the shores of Crete 3 times daily to fend off any incoming ships. Alternatively, in Hesiod's earlier work, Pandora was an artificial woman tasked with punishing humanity for discovering fire. As Mayor expressed, "It could be argued that Pandora was a kind of AI agent. Her only mission was to infiltrate the human world and release her jar of miseries[3].

[Source: Talos and Pandora high-fiving by Dalle 3 (29-05-2024)]

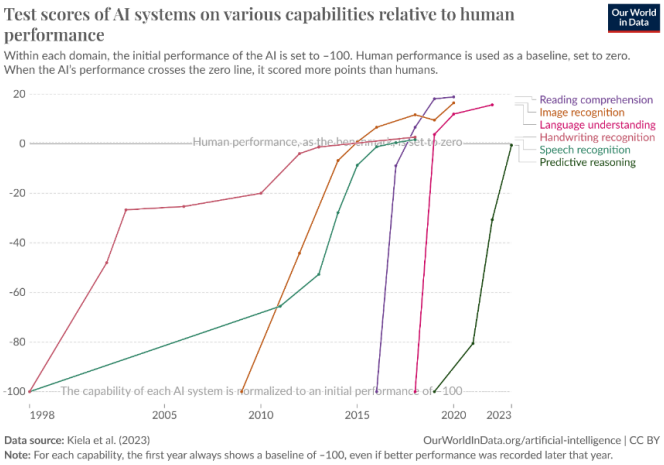
It took more than two and half thousand years after these first conceptualizations of artificial intelligence until the scientific groundwork and technology caught up to these mythical figures.

Alan Turing and Claude Shannon are the two most prominent names associated with the early stages of artificial intelligence. In Turing's 1948 report "Intelligent Machines," many of the central concepts of AI were initially laid out[4]. In 1950, Shannon built Theseus, arguably one of the first examples of machine learning in action. Named after the ancient Greek hero who escaped the labyrinth imprisoning the minotaur, Theseus was a robotic, maze-solving mouse that was able to "remember" which route it had previously taken and was thus able to escape labyrinths[5].

Although the AI field didn't stand still in the decades following, it wasn't until the 1990s and 2000s that AI experienced a resurgence driven by technological advances in computational power, algorithm development, and data availability. Machine learning, particularly neural networks, gained prominence, leading to the following milestones.

In 1997, IBM's Deep Blue defeated the world chess champion Garry Kasparov, demonstrating AI's potential in dealing with complex tasks[6]. The late 1990s and early 2000s saw the rise of support vector machines and ensemble methods. The development of deep learning in the mid-2000s, with techniques such as backpropagation in combination with the availability of large datasets, revolutionized diverse fields such as computer vision and natural language processing, setting the stage for modern AI applications.

The following chart shows the fast development in recent years, with AIs surpassing human abilities in language and image recognition tests. A meager ten years ago, this was unthinkable. Although AIs may still need to catch up to humans outside of such standardized benchmark tests, it is only a question of time until they spring forward again and surpass us in these as well[7].



[Source: Kiela et al. (2023) – with minor processing by Our World in Data[8]]

So, what does the future of AI have in store for us? Who better to answer than an AI itself?

"The future of AI promises transformative advancements across various sectors, from healthcare and transportation to education and finance. AI will enable personalized medicine, autonomous vehicles, adaptive learning, and smart city management, enhancing efficiency and quality of life. However, it also poses challenges like ethical considerations, data privacy, and job displacement. Balancing innovation with regulation will be crucial to fully harness AI›s benefits while mitigating its risks, ensuring a future where AI augments human capabilities and fosters sustainable growth."

ChatGPT 4; 22.05.2024

Artificial Intelligence and Hostile Influence

To begin with, we need to keep in mind that AIs are tools. As with all tools, the hand wielding the tools determines whether they are used to help or harm. In the right hands, this crafty AI tool can drastically reduce time spent and maximize efficiency in virtually any work process. They will also enable us to implement new methodologies that were previously too complex or cost-intensive. Unfortunately:

"AI's capabilities in data analysis and content generation can significantly enhance hostile influence campaigns, making disinformation more targeted and convincing. As AI tools become more sophisticated, the urgency for robust detection and countermeasures to protect information integrity and public trust intensifies."
ChaptGPT 4; 22.05.2024

Large Language Models

Large Language Models (LLMs) are advanced statistical models designed to understand and produce human language based on inputted language. They primarily learn context and meaning by examining the relationships within sequential data, such as consecutive words in a sentence.

With this, LLMs can analyze text, comprehend its meaning, generate appropriately coherent responses, and perform numerous language-related tasks. In recent years, they have become incredibly proficient at handling more complex texts, questions, and instructions to generate well-structured, grammatically appropriate responses. They have even mastered creating program code in multiple programming languages and proofing written code for programmers[9].

The ability for content creation is a game changer. Unfortunately, here we enter the realm of hostile influence. A small experiment we conducted in 2023 showcases what this means in practical terms. The objective was to find out what team size would now be necessary to run a news page updated with "unique" content daily. Fake news sites are frequently a key asset to hostile information campaigns. To specify, we stuck to the barebones – no blogs, forums, contact forms, SEO, or the like. We used an AI-generated template for a webpage, ChatGPT, to create content from scratch or repurposed existing content and applied the occasional stock photo or AI-generated photo. With this, we could comfortably run a daily updated news webpage with only two fully dedicated people. Given the technological advancements in AI, we are more than confident that a single person in 2024 will not only be able to run a fully functioning news website on their own but multiple websites in parallel and with ease.

Although they may not be perfect (yet), they would be highly cost-effective and could still generate a significant impact. In addition, we began translating our content into multiple languages through ChatGPT.

These translations were checked for accuracy by native speakers who only occasionally suggested minor changes to phrasing but were overall satisfied with the content translations.

There are five significant takeaways from this experiment:

1. The resources saved in content creation could be reinvested in design, marketing, etc., to professionalize the page.
2. The cost of running such a page is reduced drastically.
3. The resources saved in content creation could be reinvested in design, marketing, etc., to professionalize the page.
4. Using an LLM dramatically increases the potential to scale up a hostile influence campaign.
5. Producing content in a foreign language is easier than ever.

While this hands-on experiment already showcases the practical use and implementations of an LLM in hostile influence, it merely scratches the surface. A 2023 report by Europol investigated the potential usage of ChatGPT for criminal activities. The report primarily focused on criminal operations such as phishing campaigns, but it also recognized an important ability of ChatGPT - to not only recognize but also to reproduce speech patterns of individuals as well as groups. This has become a standard for the advanced LLMs and an invaluable tool for any hostile campaign.[10]
It doesn't end there, unfortunately.

The use of artificial intelligence in psychology and psychiatry is a hotly debated topic and warrants an entire paper of its own. Many agree on the usefulness of AI in taking over menial administrative tasks or evaluating and sorting responses to standardized questionnaires. Ethical and regulatory questions arise regarding any role in diagnosing or advising treatment.[11] While these discussions are essential, it needs to be stated that neither ethical nor regulatory questions seem to be hampering or slowing down the advancements in any way in this direction of development. In a recent conversation with a researcher in the field of AI who asked to stay anonymous, he explained how he has been working on training ChatGPT to psychoanalyze specific subsets of the populations. In his case, he has been trying to find which tendencies might lead a subset of the population to believe campaign narratives, along with the answer to why they are susceptible to specific narratives. While the study is still in its early stages, he has already analyzed a grassroots campaign, of which several 100 users (on X) were identified using one LLM. The researcher used a second LLM to create psychological profiles, and a third LLM is now in use to suggest counter-narratives. In the context of hostile influence campaigns, AI utilizing psychological profiling is a frightening development as it now allows threat actors to not only fine-tune their messaging to a specific language with the language patterns specific to particular groups but it can also take psychological factors into account for precision targeting of groups.

LLMs (Large Language Models) and their application on social media

LLMs may have started as mere chatbots, but their application in creating automated social media accounts and content was a logical destination. Bots on social media are not a new phenomenon. A 2019 study by researchers at the University of California researched the progression of sophistication in Bots on X (then Twitter) in the aftermath of the 2016 election. According to their research, during the 2016 election cycle, bots were primarily used to pump out as much content as possible: quantity, quantity, quantity. A change in bots' behavior could already be observed in the run-up to the 2018 midterm election cycle, with bots shifting from solely focusing on a high quantity of output to better mimicking human behavior.[12] Fast forward to 2024 and the capabilities of the current generation of LLMs (June 2024), and it appears that information warfare has sprung from the bronze age straight into the era of precision-guided missile systems. AI-powered accounts can be generated at unprecedented speed and can now convincingly mimic human speech and behavior to a degree that makes distinguishing them an increasingly difficult task. For hostile influence campaigns, this means that dissemination networks can now be built with a minimum of effort, by a minimum of actors, at great speed, and arguably with much greater effect. We are now facing hyper-realistic accounts with natural user behavior and language patterns and armed with psychological profiles tailored to target and reach specific groups while seeming to be from the particular group itself.

These dynamic appearances have made the identification of both fake accounts and coordinated campaigns increasingly tricky. Prior to these AI capabilities, the most prevalent methodologies for hunting hostile influence campaigns had been coordinated behavior patterns, such as using the same image and/or the same taglines over a specific time period. For example, if 50 accounts post the same image with the same tagline and hashtags within an hour, it is relatively straightforward to find a potential case of inauthentic, coordinated behavior. Another factor in identifying fake accounts has been unusual activity hours, such as finding accounts that work 24/7. With the newest generation of LLMs, many of these identifiers no longer apply to LLM-driven campaigns. Given these rapid advances in AI, the pressing question is how to identify coordinated inauthentic behavior in the future. While some work is being done in using LLMs for defensive purposes, the fact remains that attackers are light years ahead of the defenders and will remain so in the foreseeable future. Add to this the ease and scale with which content can be produced with AI, and one can see the first wave of a Tsunami heading straight towards our shores.

Image Generation and Deep Fakes

> ***"AI-generated images and videos are increasingly used in disinformation campaigns to create convincing fake content, making it difficult to distinguish between real and fabricated information. These sophisticated deepfakes can manipulate public opinion, spread false narratives, and undermine trust in legitimate media sources."***
>
> (ChatGPT 24.04.2024)

Arguably, one of the biggest milestones in AI-generated imagery was the development of Generative Adversarial Networks (GAN) by Ian Goodfellow. As he explained in an interview with the Batch in 2020: "We were discussing the problem one night at a bar, and they asked me how to write a program that efficiently manages gigabytes of data per image on a GPU that, back then, had about 1.5GB RAM. I said that's not a programming problem.

It's an algorithm design problem. Then, I realized that a discriminator network could help a generator produce images if it were part of the learning process. I went home that night and started coding the first GAN."[13] The rest is history, as they say.

Like LLMs, AIs working on image and audio recognition and production are trained on large data sets. As Yilun Du, a PhD student in the Department of Electrical Engineering and Computer Science (2022), describes it, "Imagine all of the images you could get on Google Search and their associated patterns. This is the diet these models are fed on. They're trained on all of these images and their captions to generate images similar to the billions of images it has seen on the internet. Let's say a model has seen a lot of dog photos. It's trained so that when it gets a similar text input prompt like "dog," it's able to generate a photo that looks very similar to the many dog pictures already seen.[14]

There are no limits to the application of these capabilities in the realm of hostile influence campaigns. It begins with the simple creation of memes, for which a campaign team used to need a dedicated content creation team. Now, anyone can create a catchy meme with only a few prompts.

Due to this ease and speed, instead of mass sharing a single identical meme, more sophisticated actors wielding AI's capabilities can quickly reproduce dozens of visually distinct images yet propagate the same narrative at the tip of a button. This diversity of images significantly impedes the search for coordinated inauthentic behavior.

Deepfakes have gotten a lot of publicity in recent years, with technology having reached the point of being able to create videos that are difficult to distinguish from reality. This is especially true for the average user. Audio Deepfakes are already indistinguishable from original recordings and require some creativity when attempting to debunk them as the example of the Keir Starmer below highlights. This opens the door for malicious actors to misuse these tools to create the likeness of someone without their permission, with ease. In 2024, the story of Olga Loiek made the headlines. A Ukrainian college student, Olga, was running a YouTube channel when she suddenly started getting cryptic messages lauding her Mandarin capabilities. After receiving a link to the Chinese social media site Xiaohongshu, she discovered that a deepfake of her had been created and distributed to promote Sino-Russian relations, to criticize sanctions against Russia, and to promote Russian products in Mandarin, a language which she does not speak[16].

[Source: Three images of French President Macron yelling created within a few minutes using DeepAI (29.05.2024)[15]]

This is only one of many examples of what seems to be a trend of deepfakes in which Caucasian women are being used in Chinese social media networks. As the professor of international relations at Durham University, England, Chenchen Zhang describes it in an email to the New York Times: "This representation of young white women in sexually objectified ways is a typical trope of gendered nationalism or nationalistic sexism. Viewers can get both their nationalistic and masculine pride reaffirmed in consuming this content."[17]

It's not just deepfake videos that are on the rise. Fueled by startups such as ElevenLabs, Resemble AI, Respeecher, and Replica Studios, audio deepfakes are emerging as one of the preferred tools of hostile influence campaigns[18]. They are easy to make and, just as importantly, easy to distribute and very difficult to detect. Two days before the 2023 Slovakian election, an audio "recording" emerged on which Michal Šimečka, the leader of the liberal Progressive Slovakia party, appeared to discuss the rigging of the elections, in part by buying the votes from the Roma community[19]. This case is interesting because it touches upon cultural, technical, and regulatory issues that the attackers expertly exploited. It also highlights the difficulty of effectively countering such deep fakes when facing limited time.

The stigmatization surrounding the marginalized community of the Roma is a highly and emotionally charged topic. So, it is no coincidence that they were used in this instance to allege vote rigging. Detrimentally, it takes time and a lot of cumbersome to prove a fake, be it an audio file or any other. The wow effect, the sensationally charged deepfake proclaiming vote rigging, far overrides the technical explanations, which may sound challenging to understand. Furthermore, it took days to determine the inauthenticity of the recording. This time delay between the launch of a disinformation attack and the release of results of an investigation is an almost impossible problem for fact-checkers.

Lastly, the attackers understood that within the 48 hours leading up to the election in Slovakia, a moratorium was in place mandating silence from politicians and the media. This made it tragically more challenging to counter the false accusations.[20] Following this, Michal Šimečka lost the election by a slim margin. It is notoriously difficult to quantify the effect of a single hostile influence attack, but seeing how close the polls were in the run-up to the election, it is conceivable that it might have altered the outcome.

An example that showcases the difficulty of fact-checkers in today's information sphere is the audio of Keir Starmer, the leader of the Labour Party and soon-to-be Prime Minister of Great Britain, which allegedly had him swearing at a staffer. The audio started circulating on social media during the Labour Party convention in Liverpool in 2023, arguably a very opportune time to attempt to discredit the party's leading candidate[21]. Fullfact, a fact-checking organization, concluded the audio was inauthentic but not through one decisive piece of evidence. Instead, they relied on several technical indicators, such as there being no trace of background noise and repetitive intonations, which were used to confirm their findings. Furthermore, sources within the Labor Party and the conservative minister Tom Tugendhat stated that the audio was fake, which was offered as proof of the questionable nature of the X account. As Fullfact states on its website, "We've not been able to determine whether the clip was generated with artificial intelligence, edited in some other way, or is that of a vocal impersonator, but we've not seen any evidence to suggest it is real."[22] At first glance, these explanations may seem vague, but with today's advancements in deep fakes, fact-checkers need to take a holistic approach to their investigations, considering multiple indicators. Each of the indicators on their own may seem insignificant, but in their totality, they begin painting a more complete picture.

[Source: Three images of an "angry" Keir Starmer created within a few minutes using DeepAI (30.05.2024)[24]]

The very real downside is that this is a hard sell to an audience that all too often expects definitive and easily understood proof, which is increasingly difficult to come by.

To see what future deepfake involvement in elections may hold for us, one needs to look no further than present-day India. With nearly 1 billion eligible voters in the 2024 general election, Indian campaigners have resorted to making video and audio deepfakes of their candidates to create more highly personalized messaging for each potential voter group. There are very practical reasons for this. India has 22 national languages and thousands of regional dialects spread across a vast area.[23] A national campaign has always been daunting, and AI is helping bridge the linguistic, geographical, and micro-cultural gaps.

While it is easy to see the benefits of using deepfakes here, these tools are hitting a country already rife with misinformation and virtually no regulation on their usage. Both the ruling BJP and the Indian National Congress Party have accused each other of using deepfakes in the context of this election. The normalization of deepfakes usage by campaigners adds another layer of difficulty for voters to understand what is real and what is not.

Also, the question remains of what will happen with all the know-how, knowledge, and experience gained through deepfake development and deployment in this election. It would be naïve to believe these companies will pack their stuff and simply wait for the next general election.

So what to do?

This paper's primary focus was to look at the current methods of AI in the hands of hostile influencers and how they apply them. But, it doesn't seem fair to not at least mention a few ideas as to how to deal with this fast-evolving situation. This may come as a shock to policymakers, but regulation alone will not help. AIs are developing at such incredible speeds that even a supercharged group of EU bureaucrats could never keep up. Laws and regulations take time and are primarily based on compromises between different stakeholders. There are very good reasons for this, but in the context of what is occurring in the field of AI, the regulatory system is at its limits, if not beyond. This is not to say there should be no regulations, but policymakers must leave their comfort zone. Too much of the discussion focuses on regulatory solutions instead of exploring different paths that could be implemented in parallel with the passing of regulation, and this is also true of researchers.

This paper's primary focus was to look at the current methods of AI in the hands of hostile influencers and how they apply them. But, it doesn't seem fair to not at least mention a few ideas as to how to deal with this fast-evolving situation. This may come as a shock to policymakers, but regulation alone will not help. AIs are developing at such incredible speeds that even a supercharged group of EU bureaucrats could never keep up. Laws and regulations take time and are primarily based on compromises between different stakeholders. There are very good reasons for this, but in the context of what is occurring in the field of AI, the regulatory system is at its limits, if not beyond. This is not to say there should be no regulations, but policymakers must leave their comfort zone. Too much of the discussion focuses on regulatory solutions instead of exploring different paths that could be implemented in parallel with the passing of regulation, and this is also true of researchers.

Countries need to begin monitoring narratives and hostile actors deploying narrative-spreading campaigns. Monitoring online discourse and, more explicitly, hunting for hostile campaigns is a huge challenge for liberal democracies. While countries such as China and, to a slightly lesser degree, Russia control their information space with an iron fist, liberal democracies allow free discourse to a much greater degree, and robust data protection laws significantly impede the work of security services. This creates a window through which autocratic regimes and other threat actors can attack. Far from advocating that we give up on these freedoms, as societies, we need to discuss how we can preserve them while maintaining a stringent level of oversite of such searches for hostile influences. For example, a significant amount of our work is based on identifying, hunting, and researching narratives and their dissemination networks on a macro scale.

While not ideal, there are methods of investigation that do not infringe on individual users. In too many discussions with members of security services and policymakers, data protection is used as a scapegoat to not deal with the issue instead of trying to put effort into finding tangible solutions.

The single most effective and crucial method of countering hostile influence is education. If people understand how hostile influence works, how easily we can be manipulated, what methods are used, how AI is deployed to trick and manipulate us, and how social media algorithms work, we may have a fighting chance. Education is also the sector that is one of the least appreciated and least funded. To quote an OECD study: "Overall, on average, the PISA 2022 assessment saw an unprecedented drop in performance across the OECD. Compared to 2018, mean performance fell by ten score points in reading and by almost 15 score points in maths."[25] Democracies must understand that education is much more than a system to develop a new workforce. It is also the frontline of defense in the realm of AI-driven information warfare.

Last but not least, stay up to date with our research. We regularly publish reports on digital influence and are always looking for contributors!

Refrerences

1. Ancient myths reveal early fantasies about artificial life | Stanford Report
2. Ancient myths reveal early fantasies about artificial life | Stanford Report
3. Ancient myths reveal early fantasies about artificial life | Stanford Report
4. Artificial intelligence - Alan Turing, AI Beginnings | Britannica
5. Mighty mouse | MIT Technology Review
6. The History of Artificial Intelligence - Science in the News (harvard.edu)
7. The brief history of artificial intelligence: the world has changed fast — what might be next? - Our World in Data
8. The brief history of artificial intelligence: the world has changed fast — what might be next? - Our World in Data
9. Was sind Large Language Models? Und was ist bei der Nutzung von KI-Sprachmodellen zu beachten? - Blog des Fraunhofer IESE
10. ChatGPT - the impact of Large Language Models on Law Enforcement | Europol (europa.eu)
11. AI is changing every aspect of psychology. Here's what to watch for (apa.org)
12. View of Evolution of bot and human behavior during elections | First Monday
13. How Ian Goodfellow Invented GANs (deeplearning.ai)
14. 3 Questions: How AI image generators work | MIT CSAIL
15. DeepAI
16. A UPenn Student Started a YouTube Channel. Her Face Was Stolen in China. (businessinsider.com)
17. In China, Deepfakes of 'Russian' Women Point to 'Nationalistic Sexism' - The New York Times (nytimes.com)
18. Audio deepfakes emerge as weapon of choice in election disinformation (ft.com)
19. Slovakia's Election Deepfakes Show AI Is a Danger to Democracy | WIRED
20. Slovakia's Election Deepfakes Show AI Is a Danger to Democracy | WIRED
21. An Alleged Deepfake of UK Opposition Leader Keir Starmer Shows the Dangers of Fake Audio | WIRED
22. No evidence that audio clip of Keir Starmer supposedly swearing at his staff is genuine - Full Fact
23. DeepAI
24. Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve | WIRED
25. Decline in educational performance only partly attributable to the COVID-19 pandemic - OECD

CRC | CYFLUENCE RESEARCH CENTER

The opinions expressed in articles published by the CRC are the author's alone.

www.cyfluence-research.com