Michał Kubiak, Kamil Kulesza

Trustworthy or knave? – scoring politicians with AI in real-time

"If you can bear to hear the truth you've spoken, twisted by knaves to make a trap for fools" (**If-** by Rudyard Kipling)

Introduction – AI as cognitive assistance

- The social media revolution overloaded people with news and degraded the democratic debate with polarization and disinformation. One solution could be to give humans cognitive assistance from AI to help cut across populism and unrest.
- All publicly available information on politicians and their actions throughout their careers (e.g., the content of speeches) can be evaluated. Let's take a politician in a presidential debate. With every word they say, we can incrementally score, e.g., if their statements are consistent with views presented in the past or evidence-based.
- Such a score can be presented live in a palatable, inviting way. A real-time, third-party visualization can be easily and independently added to the primary information stream (e.g., a live presidential debate video) between the source media and the viewer. Various "skins" can be used for attractive presentation, e.g., giving politicians features like a halo or devil horns, depending on cultural context and user choice.

Threat scenario

- The novelty of this AI cognitive assistance system lies in combining real-time, independent fact-checking and inviting live visualization.
- Scoring politicians in real-time is faster than the current pundits, and visualizations make the score more convincing. This could empower and engage the public, making democracy more hands-on.
- However, such perception-altering tools can a) malfunction, distorting the political process, b) be hijacked or originate from malicious third parties, disinforming under a veneer of objectivity, c) be shunned by politicians themselves, as such tools would profoundly alter how politics are run.

Demonstration of the proposed solution

- By running a series of experiments, we confirmed that such tools can be assembled with current building blocks, including popular AI tools.
- It seems only a matter of time before such AI tools will be deployed in politics, which calls for research on their safe design and usage.
- With this proof-of-concept, we consider how to make it robust and tamper-proof. Cryptographers have already addressed such challenges, e.g., with trusted third parties and majority voting protocols in fielded digital services and products.

Extrapolation into the future

- Such tools could empower the public and increase democratic oversight - or deteriorate it if used unsafely. We propose several next steps to address the critical future challenges.

Description of mitigation strategies

- We can mitigate the risks, e.g., by staged releases of said tools, making them secure by design, etc.

Figure 1: Donald Tusk, President of the European Council (2014-2019), who also happened to be a Prime Minister of Poland twice (over seven years and continuing) with many elections and campaigns, often dirty (on both sides). One of his videos was famously doctored to show him as untrustworthy via devilish horns and red skin. This symbolism became very popular with his opponents. The attractiveness of such visuals in politics could make them a tool of choice for displaying fact-checking results.

Appendices

Disclaimers

Under hackathon conditions, not providing sufficient—in our opinion—protection of our Intellectual Property, we can reveal only a general overview of our solution and a glimpse of industrial mathematics tools used. Hence, diving into the subject as deeply as in an academic seminar or commercial technical report is not conceivable. Instead, in the appendices, we show a helicopter view and general arguments supporting our solution.

Appendix A1 – Introduction – AI as cognitive assistance

- After reviewing the state-of-the-art in **fact-checking**, we can say without fear of contradiction that present AI technologies appear sufficient for real-time fact-checking of statements. This applies even to LLMs such as ChatGPT and the like. Setting up a good, real-time fact-checking and scoring system requires proper care with prompting, fine-tuning, and techniques such as Retrieval Augmented Generation. Care is also needed to properly debias the system. However, all of these tasks are currently doable, the overall performance of AI systems is improving, and it seems that decent progress is being made in reducing unwanted phenomena such as bias or so-called hallucinations in LLMs.

- In digital signal processing (DSP), the problem of **adding real-time overlays** to the signal (e.g., video) has been solved for a long time in TV broadcasts or filters in popular apps (Instagram, TikTok, etc.). Visualizations can also easily go beyond simple overlays. Synthetic video is possible to do very convincingly, as demonstrated, e.g., with Open AI's Sora or with lifelike AI-powered avatars created by companies such as Synthesia. It is also not excluded that visualizations could be run in parallel in virtual worlds, e.g., metaverse or online video games, attracting new users.

- By reviewing the state-of-the-art in fact-checking and video processing, we can conclude that the building blocks are already available to create an AI cognitive assistance system combining real-time, independent fact-checking and attractive live visualization.

Appendix A2 - Threat scenarios

- AI cognitive assistance tools can be **critical to the future of democracy** and could empower and engage the public. In turn, this will make such tools the target of many attacks and manipulation. While the catalog of such can be very wide, we present two threats below as examples.

- **Malicious third parties** (e.g., domestic and abroad authoritarian governments, radical political parties, non-state actors, etc.) could independently create such AI cognitive assistance tools which, while claiming to fact-check, would, in fact, manipulate, push propaganda and disinformation. This could be very effective if such tools become widespread and accepted. Moreover, even the tools coming from trusted parties could be hijacked and repurposed for malicious objectives.

- Several MEPs (Members of the European Parliament), a Polish MP, and other politicians consulted on the matter, speaking under the condition of anonymity, stated that for a significant fraction of politicians, real-time fact-checking with easy-to-digest visualizations would be an **existential threat to their way in politics.** Therefore, if such tools started gaining popularity, these politicians would do everything to stop or take over them.

Appendix A3 - Demonstration of the proposed solution

- With increasing disinformation and polarization, pressure will mount to deploy new fact-checking tools, including AI cognitive assistance systems. Such a perspective calls for research on their safe design and usage.

- With our previously outlined proof-of-concept, we primarily consider how to make AI cognitive assistance systems **robust and tamper-proof against malicious third parties.** Such parties can try to masquerade their services as legitimate ones or execute man-in-the-middle attacks to manipulate fact-checking and visualizations before they reach the end user.

- The challenge is non-trivial, but equally or similarly difficult things are already being done. For instance, cryptographers have already addressed such challenges in fielded digital services and products, e.g., with **trusted third parties, certification authorities, and majority voting protocols** that prevent identity spoofing and tampering with digital media streams.

Appendix A4 - Extrapolation into the future

- In the next 5-10 years, AI cognitive assistance systems in politics might take center stage and become the **primary tool that mediates the political experience** of vast groups of people, owing to conditions described in the appendices so far: the current availability of the technology (and scaling laws that will make it even more accessible), the importance of video as a medium, pressure to fight polarization and disinformation, etc.

- Will such systems deliver on their promise to empower the public and increase democratic oversight, or will they exacerbate existing issues depends on a) **democratic alignment and oversight over such tools**, b) continued technical research to improve fact-checking tools and diminish their biases and hallucinations, c) **ethics and safety research** examining how public relates to and utilizes such tools, and last but not least d) security R&D on subjects such as the ones described in Appendices A3 and A5.

- With more advanced AI, cognitive assistance could become part of the **ubiquitous digital realm** involving AI assistants, AI platforms, AI-first devices, and other digital advances. These combined could provide us with novel, seamless ways to engage with video and other data streams, e.g., in a metaverse-like fashion (but not limited to such). By researching social and technical safety and security of currently possible fact-checking and visualization tools, we can better anticipate future challenges and prepare to mitigate emerging threats to democracy.

Appendix A5 - Description of mitigation strategies

- In addition to the solutions described in Appendix A3 and A4, further mitigation strategies need to involve **staged releases** and making AI cognitive assistance tools **secure by design**.

- **Staged releases** would serve as a strategy in both social and technical senses — to familiarize the public with such tools and carefully observe their social impact while being able to respond to technical and security challenges as they arise.

  Security by design for AI cognitive assistance tools should involve integrating robust security measures into every stage of development, from conception to deployment. This approach would ensure that the tools are designed with security as a primary consideration rather than an afterthought. Example security measures in this paradigm include, but are not limited to: implementing protocols to protect data integrity, utilizing authentication mechanisms to control access, and incorporating mechanisms for detecting and mitigating potential vulnerabilities.

Appendix A6 – Authors

MICHAŁ KUBIAK
Researcher in CIAMSE in Warsaw, **AI Policy Officer** in the European DIGITAL SME Alliance in Brussels

Education: degree in Physics (Poland)

Experience: 10 years in IT, including commercial and quantitative problem solving, acting on the boundary between academia and business (e.g., now in Brussels)

Media: popularization of Science; writing on AI

KAMIL KULESZA
Head of CIAMSE in Warsaw

Education: degrees in Mathematics, Computer Science, Physics (all 3 - South Africa), PhD in Cryptography (Poland), Certificate in managing high-tech (Cambridge) - Business School while on mathematical postdoc

Experience: over 20 years in academia and IT, including commercial (project management) and quantitative problem-solving

Media: popularization of Science; about 100 media appearances and invited articles (mainly mainstream media, but also TEDx) with multiple recent articles on AI, as well as articles on the high-tech, role of Science, managing academic/high-tech projects, etc.

CIAMSE - Centre for Industrial Applications of Mathematics and Systems Engineering
(https://www.maths.com.pl/English_Info)