FUTURE OF EUROPE CAPSTONE PROJECT
DEMOCRACY GROUP 14

USING GENERATIVE AI TO DEBUNK SOCIAL MEDIA DISINFORMATION IN THE EU

Katherine Larson, London School of Economics
Fabio Schettino, Bocconi University
Arunima Singh, Stockholm School of Economics

I. **THE PROBLEM**: SOCIAL MEDIA FUELING DISINFORMATION, DISINFORMATION FUELLING EUROSCEPTICISM

If the EU was able to better counter disinformation on social media through fact-checking and media literacy, would the UK still have voted to leave the EU in 2016?

While such rhetorical questions are impossible to answer, **social media has certainly emerged as a powerful force shaping public discourse and political narratives** and has become a regular tool for political communication. As populists across the West have been early adopters of social media, many forms of social media – but especially Twitter and Facebook – have become linked with populist movements and contributing to polarization. Thus, social media has played no small part in fostering and amplifying Eurosceptic ideas, and social media-fuelled disinformation has contributed to a widespread emergence of Eurosceptic narratives (Gerburado et al. 2018, Galston et al. 2017, Henkel 2021).

This trend has had real and damaging effects. In the lead-up to the 2016 Brexit referendum, the "Leave" campaign utilized false and misleading statements, which had a destructive impact on the democratic function of news media, eventually changing citizen narratives (**Figure 1**) (Henkel 2021). This disinformation-fuelled rise of Euroscepticism generated affective polarization and partisanship, with measured increases in stereotyping, prejudice, and multiple evaluative biases (Hobolt et al. 2021).

Polarization online. Platforms like Facebook, Twitter, and Instagram have become primary channels for information dissemination, enabling rapid and widespread communication (Bavel et al. 2021, Fortunato et al. 2022). However, evidence suggests that increased exposure to populist social media channels strengthens populist views among those who already hold strong views, polarizing those at the margins, which has been observed in Europe (Jenkins & Fletcher, 2019). Further, extreme

viewpoints are amplified in social networks, which may lead to confirmation bias and echo chambers, as people become more entrenched in their opinions (Bavel et al. 2021).

Education and disinformation. There is also a strong relationship between exposure to online political activity and Euroscepticism among individuals with lower formal education. Thus, a key solution is to educate individuals about how to obtain information about politics and improve media literacy, as discussed further in the proposal portion of this paper (Fortunato et al. 2022).

The challenge of Euroscepticism. Affective polarization coming from social media platforms favors the emergence of Eurosceptic ideas, amounting to a foundational challenge and threat to the principles of unity and collaboration among EU member states (Condruz-Bacescu, 2014). If Euroscepticism constitutes an existential threat to the European project, then it is crucial to understand the ways in which social media influences and facilitates the spread of Eurosceptic ideas. As freedom of expression is a key tenet of the European human rights framework, it's critical to ensure protection of speech while still problematizing disinformation that can polarize and destabilize.

The power of social media and governance. The internet and social media have made Eurosceptic claims and positions visible and provided platforms for populist movements contesting EU policies. However, the relationship between social media and European democracy is not a fundamentally negative one – digital tools have the potential to hold governing power structures accountable, acting as the intermediary between citizens and their representatives (Caiani et al, 2017). Because social media platforms can pose fundamental challenges to the legitimacy of institutions and policies within the EU, social media regulation to address disinformation is critical to safeguard democratic processes (Barisione et al, 2017).

II. **EU SOLUTIONS THUS FAR** (AND THEIR SHORTCOMINGS)

The European Union has adopted comprehensive policies and initiatives to address misinformation and disinformation across media platforms, but especially regarding social media, given its mass potential to influence public opinion and rapidly spread potentially false information.

- **Regulatory Framework**. The EU's regulatory framework, based on the European Commission's Digital Services Act (DSA) and the Code of Practice on Disinformation, seek to establish clear rules for digital services, including social media platforms. They also aim to address the spread of illegal content, including misinformation and disinformation. However, the effectiveness of such voluntary measures relies on the willingness of platforms to comply, and doesn't address underlying media literacy challenges.

- **Administrative and coordination**. The EU has further implemented administrative and coordination-oriented initiatives, including cross-border alert systems to improve info-sharing on disinformation campaigns between member states. The EU has also allocated research and innovation funding through programs like Horizon Europe to support technologies that can help identify and counter disinformation, including generative AI.

- **Fact-checking initiatives**. The EU has set forth multiple fact-checking initiatives in order to problematize the spread of false information via social media. The European Digital Media Observatory (EDMO), EUvsDisinfo, and the European Fact-Checking Network (ECFN) are three such initiatives, all of which continually fact check and debunk harmful disinformation. While these fact-checking and transparency measures have been introduced at a number of levels, holding social media platforms accountable for the spread of disinformation remains a challenge. Such fact-checking platforms may have a limited ability to reach populations swayed or radicalized by disinformation, and their efficacy is constrained by how widely they are referenced and adopted.

- **Media Literacy**. Finally, the EU promotes media literacy as a key tool in combating misinformation. Initiatives such as "Media Literacy for All" aim to empower citizens to critically assess information and navigate online content. However, gaps exist in reaching diverse demographics and ensuring that citizens are adequately equipped to navigate the digital information landscape. Of particular concern are more marginalized and less educated populations who are more prone to polarization via social media.

Gaps in the current policy status quo exist at all levels, especially considering how rapidly technology evolves and the challenges of law and policy to keep up. Key issues include improving access to fact-checking and media literacy programs for grassroots populations, as well as adapting law to current and emerging technology. Improvements in the policy landscape are critical not only to improve media literacy and consumption, but also to preserve democratic ideals and the European Union itself.

III. **OUR SOLUTION**: FACTUALLYEU

Our Proposal and its Objectives

Situating our proposal. The gaps in the current policy landscape that concern us most are the inability of existing policies to keep up with current technology, the challenge of media literacy programs and fact-checking programs to reach grassroots and marginalized populations who need them most, and the lack of any initiatives with a tailored focus on countering disinformation specifically related to EU matters, which is a critical nuance in the context of Eurosceptic narratives. Additionally, the relatively limited integration of AI, despite its disruptive potential, represents a gap in the current strategies. The dichotomy between initiatives primarily concentrating on fact-checking or enhancing media literacy and education, without an integrated approach encompassing both aspects, further underscores the shortcomings in the present policy framework.

AI as a solution. Given these premises, our proposal rests on the idea that Artificial Intelligence can be used to strengthen both institutional and societal defenses against disinformation. This can include incorporating generative AI-driven detection mechanisms within toolkits and applications designed to identify and signal deceptive or inaccurate content on social media platforms. Simultaneously, AI can serve an educational role by informing users about diverse methods of information manipulation, thereby improving their digital literacy. This dual functionality not only involves identifying and addressing misleading content but also empowers users with the knowledge and skills needed to navigate and critically evaluate information online.

We therefore suggest employing an AI toolkit, named "FactuallyEU," to identify and highlight deceptive or inaccurate content pertaining to EU matters and policies on social media platforms. Subsequently, users encountering such flagged content would be directed to an interactive platform where they can engage in a conversation with the AI. Within this dialogue, users would receive comprehensive and insightful explanations detailing the factual inaccuracies or misleading aspects of the statements found in the online content.

The FactuallyEU initiative consists of three pillars:

01 Innovative AI technology

At its core, the initiative would leverage the advanced GPT (Generative Pre-trained Transformer) technology, designed to understand and generate human-like text based on the input it receives. This GPT-powered system would serve as a robust fact-checking engine, specifically instructed to be capable of processing vast amounts of information related to EU matters and policies. Its machine learning capabilities would also ensure that it evolves with the dynamic nature of disinformation, adapting and improving over time to effectively counter emerging false narratives. This AI Toolkit would be designed to be multilingual, addressing the diverse linguistic landscape within the EU, ensuring accessibility and impact across all member states.

02 User-friendly Interface

The FactuallyEU initiative would prioritize the development of a user-friendly chat interface. Recognizing the importance of simplicity and accessibility, the chat interface will enable seamless interactions between users and the chatbot. Users will be able to easily navigate and engage with the system, receiving instant responses and explanations about EU-related information. Additionally, the initiative would incorporate engaging multimedia content to enhance user understanding. By integrating visuals, videos, and infographics, it will ensure that the educational experience is not only informative but also captivating, catering to a broad audience with varied learning preferences.

03 Structured debunking approach

FactuallyEU shall be instructed to reason in a structured way aligned with cognitive processes, aimed at debunking myths and disinformation circulating online. Following a well-defined framework, the AI shall employ a four-step method:
1) Stating the factual information;
2)   Issuing a warning about an incoming myth;
3)   Explaining the fallacy within the myth;
4)   Reinforcing the correct information.

This structured approach is rooted in effective communication strategies outlined in the "Debunking Handbook 2020", ensuring that corrections to myths and misleading content are clear, concise, and memorable. The tool will place emphasis not only on providing factual alternatives but also on highlighting logical and argumentative fallacies, contributing to enhanced cognitive alignment and critical thinking among users.

Implementation and Administration of FactuallyEU

Administrative Structure. FactuallyEU will be situated within the <u>European Commission's Digital Strategy</u>, aligning with the overarching goals of ensuring a secure and informed digital environment for EU citizens. The administration of FactuallyEU will be overseen by a dedicated cross-cutting team comprising:

- Software Engineers: Responsible for the development, maintenance, and continuous improvement of the AI model and chat interface.
- Policy Advisors: Tasked with staying abreast of misinformation trends, collaborating with relevant EU bodies, and ensuring that FactuallyEU aligns with EU policies.
- Social Media Specialists: Focused on understanding the dynamics of social media platforms, fostering collaboration with major platforms, and optimizing the user experience.

FactuallyEU should be funded through the Digital Strategy budget, and initial innovation and startup funds can be allocated from research funds (for example, the Horizon fund and project).

Dataset Management. The dataset supporting FactuallyEU will exclusively consist of trustworthy official EU sources, including legislation, policies, meeting transcripts, and relevant data. To guarantee accuracy and comprehensiveness, the administration team will conduct thorough reviews and updates, ensuring that FactuallyEU is well-informed about current EU policies. Funding for dataset creation and maintenance can be sourced from the EU budget, and transparency in dataset management will be a priority – each new dataset update would be visible online.

Requirements for Social Media Platforms. FactuallyEU shall aim to address implementation gaps by becoming a mandatory add-on interface for all major social media platforms operating in the EU. This integration can be achieved through regulatory measures and collaborations with EU regulatory bodies. Compliance with this requirement will be a condition for social media platforms to operate within the EU, emphasizing the significance of countering misinformation as a shared responsibility.

Countering Misinformation, Not Propaganda. It is imperative to stress that FactuallyEU shall be designed to counter misinformation objectively and transparently, without veering into the realm of propaganda. To ensure this, the platform will undergo strict training in setting the dataset parameters and implementing the structured debunking approach. An independent oversight body, potentially constituted by experts in media ethics and information integrity, can provide periodic evaluations to safeguard against biases and maintain the tool's neutrality.

What if British citizens had access to FactuallyEU? In a hypothetical scenario where British citizens had access to FactuallyEU before the Brexit referendum, the AI could have played a pivotal role in countering the infamous myth that the EU took £350 million a week from British citizens.

01 Automated fact-checking

Before the user engagement, FactuallyEU's would have automatically fact-checked social media posts containing EU-related content, ensuring proactive identification and flagging of potentially inaccurate information.

02 Disclaimers on Social Media Posts

British citizens encountering misleading information would have seen a disclaimer under posts containing the myth, stating, "This post may contain inaccurate or misleading content. To know more, ask FactuallyEU," with a link to the chat interface. *(Figure 2)*

03 Debunking through FactuallyEU

Redirected to the chat interface, users would have engaged with FactuallyEU in a conversation structured to debunk the myth. *(Figure 3)*

04 Post-Debunking Interaction

After the debunking process, users would have had the option to explore related EU topics, ask additional questions, or access curated educational content, fostering a continuous and informed dialogue.

Figure 2. Example of FactuallyEU disclaimer under social media posts.
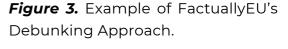
Why is this post's claim inaccurate?

FACT: The UK's decision to leave the EU carries uncertain economic consequences. The intricacies of the Brexit process would involve multifaceted economic considerations, and the outcomes are subject to various factors, including trade agreements, market dynamics, and policy changes.

MYTH: The claim that leaving the EU would result in the National Health Service (NHS) receiving £350 million a week is one of the central assertions made by the Leave campaign in favor of Brexit. This figure, however, is deemed a "clear misuse of official statistics" by the UK Statistics Authority, primarily because it does not consider the UK's financial inflows from the EU after contributing to the budget.

FALLACY: This assertion falls prey to the fallacy of oversimplification and cherry-picking of data. By focusing solely on the gross figure without accounting for the UK's financial interactions with the EU, the claim provides an incomplete picture of the financial implications of leaving the EU and creates a distorted view of the funds that would be available for the NHS.

FACT: In reality, the UK Statistics Authority's correction underscores that the actual amount, accounting for the rebate, is closer to £250m. Moreover, the intricacies of budgeting, economic impact, and post-Brexit deal financial obligations mean that the claimed £350m might not materialize as promised. The Office for Budget Responsibility has cautioned that potential savings from leaving the EU do not consider the broader economic consequences or financial commitments arising from the Brexit agreement.

Figure 3. Example of FactuallyEU's Debunking Approach.

IV. **CONCLUSION**

EU-related matters and policies often appear as a complex "black box" for European citizens due to the intricate nature of legislative processes and bureaucratic language. FactuallyEU's unique value lies in demystifying these topics, providing clear and accessible explanations. By integrating the AI toolkit into social media platforms, users would gain instant access to reliable information, fostering a more informed public discourse on EU-related matters.

In a world dominated by social media narratives, where misinformation has become a catalyst for Eurosceptic sentiments, FactuallyEU would provide a beacon of clarity. By countering misleading content with a structured debunking approach, this initiative would not only rectify false narratives but also equip users with the tools for critical thinking and media literacy. In short, **FactuallyEU is a policy proposal that empowers citizens, fosters informed decision-making, and, ultimately, fortifies the foundations of a united and well-informed European Union.**