Digital Systems for Humans Graduate School

2025-2026 PhD Subject Proposition

Proposition de Sujet de Thèse 2025-2026

TITLE

Taming the Digital Persuasion Engine:

Human Oversight in the Age of Generative AI

Doctoral School: École doctorale DESPEG/ED513

Thesis supervisor: Caroline Lequesne

Host laboratory: GREDEG CNRS UMR 7321

Subject description:

The online information ecosystem is undergoing a profound transformation as generative artificial intelligence (AI) converges with the high-frequency recommender architectures of social-media platforms. Whereas early scholarship documented bots and coordinated inauthentic behaviour (Newman et al., 2017), today's hybrid systems go much further. State-of-the-art text, image, video, and audio generators can mimic human rhetoric, appeal to emotion, and exploit cognitive biases at near-zero cost, allowing malign actors to fabricate persuasive falsehoods at unprecedented scale and speed (G'Sell, 2024). By feeding such synthetic content into ranking algorithms that privilege engagement, generative AI intensifies what Boullier (2024) calls the *"self-replicating milieus"* of social media—feedback loops in which viral items autonomously amplify through shares, reposts, and algorithmic boosts.

A vivid example is Platform X's integration of "Grok," an in-house large language model. Here, (i) Grok *creates or suggests* speech, (ii) the platform's recommender *amplifies* it, and (iii) user engagement is *re-cycled* as training data. Such a co-production loop means harmful narratives can reach millions before traditional *flag-review-action* moderation ever intervenes. Ex-post takedowns are simply too late to neutralise behavioural effects. This situation can produce serious societal harms, spanning cyber-crime to broader social unrest and destabilization. For instance, reports have highlighted that far-right groups are increasingly employing AI-generated content to disseminate anti-immigrant sentiments and conspiracy theories across European countries, with a notable surge during election periods (Quinn & Milmo, 2024). The situation becomes even more alarming considering that generative AI can create realistic human faces or voices that do not belong to actual individuals. Drawing an analogy to the strict regulations against counterfeiting currency to safeguard the financial system, philosopher Daniel Dennett advocates for a governmental ban on the creation of "counterfeit people," such as social media bots, to preserve democratic systems founded on genuine

conversations between real individuals (Dennett, 2023). Furthermore, UN Secretary-General António Guterres has advocated for the idea of creating a global AI regulatory body, akin to the International Atomic Energy Agency (IAEA), to oversee AI safety and governance, particularly in light of the risks posed by generative AI, such as deep fakes and misinformation (Nichols, 2023).

Against this backdrop, this PhD project aims to answer the following question: Does the EU's trilateral regulatory regime, comprising the GDPR, the Digital Services Act (Regulation (EU) 2022/2065), and the Artificial Intelligence Act (Regulation (EU) 2024/1689), together provide a sufficiently precise, enforceable, and technologically attuned model of "meaningful human oversight" on artificial intelligence systems to mitigate the systemic disinformation risks posed by generative-AI and AI-driven recommender systems on Very Large Online Platforms, or is a supplementary harmonised standard under Article 40 AI Act and/or further legislative amendment required to fill the governance gaps?

References:

Boullier, D. (2024). *Social Media Reset:Redesigning the infrastructure of digital propagation to cut the chains of contagion*. Sciences Po. Retrieved March 1, 2025, from https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2024/06/Dominique-Boullier-Social-Media-Reset_compressed.pdf

Dennett, D. C. (2023, May 16). *The Problem With Counterfeit People*. The Atlantic. Retrieved December 15, 2024, from https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

G'sell, F. (2024). REGULATING UNDER UNCERTAINTY: Governance Options for Generative A. *Standor Cyber Policy Center*. https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2024-12/GenAI_Report_REV_Master_%20as%20of%20Dec%2012.pdf

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., & Nielsen, R. K. (2017). *Reuters Institute Digital News Report 2017*. Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf

Nichols, M. (2023, June 12). *UN chief backs idea of global AI watchdog like nuclear agency*. Reuters. Retrieved December 15, 2024, from https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/

Quinn, B., & Milmo, D. (2024, November 26). *How the far right is weaponising AI-generated content in Europe*. The Guardian. Retrieved December 10, 2024, from https://www.theguardian.com/technology/2024/nov/26/far-right-weaponising-ai-generated-content-europe

Candidater / Apply