FWO Project Proposal Outline

1. Rationale

The proposed PhD project seeks to understand the localisation of Russian disinformation narratives in the Netherlands, Belgium, and France through conversational Artificial Intelligence (AIs). Since the Russian interference in the 2016 United States elections, hybrid warfare has increasingly become an object of study. Russian disinformation is most notably a threat to Western democracies due to its effectiveness and volume. However, scholars point to a misconceptualisation of disinformation, as research attempts to propose solutions which utilise fact-checking mechanisms through algorithms or content censorship, leading to a simplification of the issue at hand (Kuznetsova et al., 2024; Makhortykh et al., 2024; Moschopoulos et al., 2023). At the core of online disinformation strategies lies narrative localisation, which goes a step beyond spreading misinformation. Recent technological progress and the current widespread use of conversational AIs have provoked new growth within online disinformation strategies. By looking specifically at five conversational AIs (ChatGPT, Deepseek, Gemini, Claude, and Perplexity), the manner in which Russian disinformation narratives are locally constructed will be examined through the use of research personas and data donations.

Over the past decade, Russian disinformation has come to be viewed as an existential threat to Western democracies due to its capacity to erode trust in institutions, distort political information environments, and foster polarisation through hybrid information warfare as an added layer to traditional military actions (Virtosu & Goian, 2023). This digital layer that overlaps with traditional military operations "makes hybrid warfare difficult to identify and counter by the involved adversaries" (Virtosu & Goian, 2023, p. 198). This has positioned Russian disinformation as a central component of hybrid warfare, one that deliberately targets the stability of democratic systems with effectiveness and volume.

Scholars warn that existing research on societal outreach often misses conceptual nuances and treats disinformation as mainly an issue of facticity (Kuntur et al., 2024; Michail et al., 2022). This way of thinking about disinformation risks mistaking it for a problem of spreading misinformation rather than a practice of narrative construction, identity manipulation, and cultural resonance. Such an approach may limit a fuller understanding of how disinformation actually operates. Too often, as argued by Kuznetsova et al. (2024), Makhortykh et al. (2024), and Moschopoulos et al. (2023), research attempts to propose

solutions which utilise fact-checking mechanisms through algorithms or content censorship. In this regard, the context and purpose of propaganda are often overlooked: Russian disinformation campaigns typically aim to exacerbate pre-existing societal divisions. Research needs to move beyond binary categorisations of true or false.

To effectively target diverse European audiences, Russian disinformation actors strategically adapt their narratives to local contexts, exploiting concerns about religion, migration, and resources by invoking historical narratives, cultural identities, and emotional appeals (Lemke & Habegger, 2022; Tyushka, 2021). A narrow focus on fact-checking risks overlooking the ways disinformation functions as a complex ideological tool. Rather than being limited to the spread of incorrect information, disinformation also involves intent, cultural framing, and the purposive use of narratives. In this context, the challenge lies in understanding how Russian disinformation adapts and localises narratives to resonate within specific environments. This calls for a need to further study the production of narrative localisation operations within the Russian disinformation context.

Recent technological progress has led to contemporary conversational AIs marking a new frontier in information warfare. Both social actors and academics sound the alarm, as the now widespread availability of LLMs (Large Language Models) and NMTs (Neural Machine Translation) allows for a significant growth in the production, amplification, and effectiveness of disinformation narratives (Bontcheva et al., 2024; Romanishyn et al., 2025). Such disinformation narratives can not only be persuasive but also be generated cheaply, in a coordinated manner, and through sophisticated AI-based disinformation campaigns (Bontcheva et al., 2024). In order to properly combat Russian disinformation, there is a new need to understand disinformation through conversational AIs. Consequently, given the relatively recent introduction of conversational AI for public use, scholars grappling with Russian disinformation and the localisation of narratives through LLMs and NMT are now calling for more insight into understanding conversational AI as a tool for disinformation (Makhortykh et al., 2024; Vykopal et al., 2023). Therefore, the proposed PhD project focuses on conversational AIs (ChatGPT, Deepseek, Gemini, Claude, and Perplexity), examining how they localise Russian disinformation in text- and image-based media.

2. Positioning within the scientific state-of-the-art

Contemporary scholarship converges on two linked observations about the role of large language models within disinformation scholarship: (1) large language models (LLMs)

substantially change both the production and the detection of disinformation, (Nathanson et al., 2024; Shah et al., 2024), and (2) they do so in ways that are deeply conditioned by language, culture and platform affordances (Makhortykh et al., 2024; Vinay et al., 2024). Firstly, scholarship on LLMs within the context of disinformation suggests that researchers are optimistic about LLMs' potential to aid in detecting disinformation. In a wide-ranging literature review, Shah et al. (2024, p. 16) note that LLMs hold promise "to revolutionise the domain of misinformation and disinformation detection." Similarly, Shah et al. (2024) and Jiang (2024) show that LLMs have potential in aiding fact-checking, flagging manipulated content, claim extractions, and summarisation, capabilities which could meaningfully scale detection workflows. Given the relatively recent introduction of LLMs for public use, studies seeking to understand LLMs as a dynamic tool for adaptive localisation are still sparse. This gap is particularly significant given emerging evidence that conversational AIs are already capable of producing highly context-sensitive outputs. Barman (2024) and Vykopal et al. (2024) show that such AIs can rapidly generate persuasive, customised false narratives on demand, signalling a serious amplification risk if disinformation actors weaponise generative models.

Secondly, studies reveal systematic limits when LLMs operate across multilingual and culturally diverse contexts. Research on Russian disinformation shows inconsistent performance and political bias across languages, with lower accuracy and more hallucinatory outputs in lower-resource languages (Makhortykh et al., 2024). Most interestingly, Vinay et al. (2024) show that the manner in which an LLM is prompted 'emotionally' also impacts the success rate of disinformation generation: polite prompts result in high successful generation frequency rates, whereas impolite prompts cause the generation frequency rates to fall, as well as models refusing to generate disinformation altogether. This showcases the extent to which the production of disinformation is tied to the platform's affordances: the spread of disinformation through LLMs is impacted by language, culture, and interactional style. Further research is needed on how conversational AIs actively shape Russian narratives and their localisation.

The literature argues for reframing disinformation research away from a binary true/false paradigm toward narrative-centred analysis. LLMs often conflate rhetorical strategies (ethos, pathos, logos) with factuality. So, emotional framing may be misclassified as false or innocuous content as truthful depending on the LLM model (Sosnowski et al., 2024). Underlining concrete definitions behind the terms of misinformation, disinformation, and malinformation is crucial to resisting the problem of *information disorder* (Fallis, 2009),

which is, in its turn, an overarching, encompassing label for its modern-day form, which can serve to better understand nuances within the processes of misinformation, disinformation, and malinformation. For this proposed PhD project, let us look at Santos-d'Amorim and De Oliveira Miranda (2021), who offer the following definition: "disinformation is information deliberately deceptive, intending to deceive" (p. 16). What such definitions highlight is the underlying strategic intent within disinformation, which requires robust methodological tools to study. Thus, interdisciplinary, mixed-methods approaches, namely, computational detection and qualitative narrative analysis, are gaining traction to capture elements such as narrative localisation through conversational AI.

3. Research Methodology

Russian disinformation adapts its tactics regionally, leveraging a common history in Eastern Europe while seeking to fuel internal division in the West (Arribas et al., 2023). Focusing on localised divisive narratives, the proposed PhD project will be conducted using different case studies across France, Belgium, and the Netherlands. These countries are useful case studies for the study of localisation via LLMs as they have distinct national borders, but also intra-linguistic variations. Flemish Dutch, spoken in Belgium, differs in vocabulary, tone, and cultural reference points from the Dutch spoken in the Netherlands, just as in rather the same manner, Walloon French differs from the French spoken in France.

To exemplify one such localised narrative, the "Golden Billion" conspiracy will provide a case study. This particular conspiracy claims that a group of elites from the Western world exploit citizens of the South as a means of keeping everyone else poor and to hoard resources and wealth for their own means (Willaert & Tuters, 2025). Russian disinformation reframes this narrative in Europe to tap into unresolved debates about colonial history. In France and Belgium, where colonial legacies remain politically sensitive, these narratives get picked up and reframed to undermine trust in European institutions. Russia is hence able to present itself as a defender of the oppressed, while the EU and NATO are cast as the latest forms of colonial domination (Audinet, 2025).

Vitally, these stories latch onto pre-existing tensions. In Belgium, the memory of Leopold II and the colonisation of Congo is still a societally divisive topic (Goddeeris, 2025). Russian disinformation thus attempts to exploit such tensions and twist them into broader claims that Europe has never stopped exploiting others and therefore cannot be trusted (Byford et al., 2024). Russian narratives of anti-Western sentiments are further used in Africa

as an attempt to gain influence there as well, seen in, for example, the use of Lumumba's assassination (Titeca, 2023). At the centre of the narratives, there is always a kernel of truth, as colonial exploitation did take place, but it gets expanded into a narrative of permanent Western domination, where NATO, EU policies, and even the global financial system, such as the International Monetary Fund, are framed as tools of oppression (Filip & Ablazov, 2024).

Furthermore, these narratives have a hybrid facet. They create a link between historical elements and current events, such as the ongoing Ukrainian war, to show that there is a continuity in Western hypocrisy. Because they are embedded in local cultural and historical reference points, they don't feel like an external imposition but rather an "organic" part of local debates (Dougherty, 2014).

The project will study conversational AIs directly, with particular attention to how they perform localisation when generating or translating content across linguistic and cultural contexts. LLM outputs leave detectable traces within the language they produce, visible in lexical preferences, syntactic regularities, collocation patterns, and idiomatic missteps (Dankers et al., 2022; Riley et al., 2020). Data will be collected through research personas and data donations. By analysing conversational AIs' features systematically, the project will classify and compare outputs across different national and intra-linguistic varieties (e.g., Flemish Dutch versus Dutch from the Netherlands, or Walloon French versus metropolitan French). This classification provides a foundation for understanding how LLMs adapt to local contexts, and whether their strategies of localisation reproduce, transform, or flatten cultural differences.

Bibliography

Arribas, C. M., Arcos, R., Gértrudix, M., Mikulski, K., Hernández-Escayola, P., Teodor, M., Novăcescu, E., Surdu, I., Stoian, V., & García-Jiménez, A. (2023). Information manipulation and historical revisionism: Russian disinformation and foreign interference through manipulated history-based narratives. *Open Research Europe*, *3*, 121. https://doi.org/10.12688/openreseurope.16087.1

Audinet, M. (2025). 'Down with neocolonialism!' Strategic narrative resurgence and foreign policy preferences in wartime Russia. *European Journal of International Security*, 1–22. https://doi.org/10.1017/eis.2025.10011

Barman, D., Guo, Z., & Conlan, O. (2024). The Dark Side of Language Models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning With Applications*, *16*, 100545. https://doi.org/10.1016/j.mlwa.2024.100545

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., Teyssou, D., Severio Nucci, F., Spangenberg, J., Srba, I., Aichroth, P., Cuccovillo, L., & Verdoliva, L. (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. In *European Digital Media Observatory*. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

Byford, A., Doak, C., & Hutchings, S. (2024). Decolonizing the transnational, Transnationalizing the decolonial: Russian studies at the Crossroads. *Forum for Modern Language Studies*, *60*(3), 339–357. https://doi.org/10.1093/fmls/cqae038

Dankers, V., Lucas, C., & Titov, I. (2022). Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.252

Dougherty, J. (2014, July 1). *Everyone lies: the Ukraine conflict and Russia's media transformation*. https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37376528

Fallis, D. (2009, February 28). *A Conceptual analysis of disinformation*. http://hdl.handle.net/2142/15205

Filip, D., & Ablazov, I. (2024). Russian activities to discredit international institutions in the light of the Russian-Ukrainian war. *Political Science and Security Studies Journal, 5*(3), 22-43.

Goddeeris, I. (2025). Leopold II, Kimpa Vita and the local decolonisation of the Belgian public space. *Journal of Modern European History*, *23*(3), 309–325. https://doi.org/10.1177/16118944251348781

Jiang, B., Tan, Z., Nirmal, A., & Liu, H. (2024). Disinformation Detection: an Evolving challenge in the age of LLMs. In *Society for Industrial and Applied Mathematics eBooks* (pp. 427–435). https://doi.org/10.1137/1.9781611978032.50

Kuntur, S., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Under the Influence: A survey of large language models in fake news Detection. *IEEE Transactions on Artificial Intelligence*, 1–21. https://doi.org/10.1109/tai.2024.3471735

Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2024). In generative AI we trust: can chatbots effectively verify political information? *Journal of Computational Social Science*, *8*(1). https://doi.org/10.1007/s42001-024-00338-8

Lemke, T., & Habegger, M. W. (2022). Foreign Interference and social Media networks: A relational approach to studying contemporary Russian disinformation. *Journal of Global Security Studies*, *7*(2). https://doi.org/10.1093/jogss/ogac004

Lu, Y., Wang, H., & Wei, W. (2023). Machine Learning for Synthetic Data Generation: A review. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.04062

Makhortykh, M., Sydorova, M., Baghumyan, A., Vziatysheva, V., & Kuznetsova, E. (2024). Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine. *Harvard Kennedy School (HKS) Misinformation Review*, *5*(4). https://doi.org/10.37016/mr-2020-154

Michail, D., Kanakaris, N., & Varlamis, I. (2022). Detection of fake news campaigns using graph convolutional networks. *International Journal of Information Management Data Insights*, *2*(2), 100104. https://doi.org/10.1016/j.jjimei.2022.100104

Moschopoulos, V., Tsourma, M., Drosou, A., & Tzovaras, D. (2023). Misinformation detection based on news dispersion. *2023 24th International Conference on Digital Signal Processing (DSP)*, 1–5. https://doi.org/10.1109/dsp58604.2023.10167997

Nathanson, S., Yoo, Y., Na, D., Cao, Y., & Watkins, L. (2024). A step towards Modern Disinformation Detection: Novel Methods for Detecting LLM-Generated Text. *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*, 615–620. https://doi.org/10.1109/milcom61039.2024.10773838

Riley, P., Caswell, I., Freitag, M., & Grangier, D. (2020). Translationese as a Language in "Multilingual" NMT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7737–7746. https://doi.org/10.18653/v1/2020.acl-main.691

Rettberg, J. W. (2024). How generative AI endangers cultural narratives. *Issues in Science and Technology*, *40*(2), 77–79. https://doi.org/10.58875/RQJD7538

Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1569115

Santos-d'Amorim, K., & De Oliveira Miranda, M. F. (2021). Informação incorreta, desinformação e má informação: Esclarecendo definições e exemplos em tempos de desinfodemia. *Encontros Bibli Revista Eletrônica De Biblioteconomia E Ciência Da Informação*, *26*, 01–23. https://doi.org/10.5007/1518-2924.2021.e76900

Shah, S. B., Thapa, S., Acharya, A., Rauniyar, K., Poudel, S., Jain, S., Masood, A., & Naseem, U. (2024). Navigating the web of disinformation and misinformation: large language models as Double-Edged Swords. *IEEE Access*, 1. https://doi.org/10.1109/access.2024.3406644

Sosnowski, W., Modzelewski, A., Skorupska, K., Otterbacher, J., & Wierzbicki, A. (2024). EU DisinfoTest: a Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives. *Findings of the Association for Computational*

Linguistics: EMNLP 2024, 14702–14723. https://doi.org/10.18653/v1/2024.findings-emnlp.862

Stahl, B. C. (2006). On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. *Informing Science the International Journal of an Emerging Transdiscipline*, *9*, 083–096. https://doi.org/10.28945/473

Titeca, K. (2023). Russian influence, anti-Western sentiments and African agency: The struggle for influence in the Democratic Republic of Congo. *Strategic Review for Southern Africa*, *45*(1). https://doi.org/10.35293/srsa.v45i1.4617

Tyushka, A. (2021). Weaponizing narrative: Russia contesting EUrope's liberal identity, power and hegemony. *Journal of Contemporary European Studies*, *30*(1), 115–135. https://doi.org/10.1080/14782804.2021.1883561

Vinay, R., Spitale, G., Biller-Andorno, N., & Germani, F. (2024). Emotional manipulation through prompt engineering amplifies disinformation generation in AI large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.03550

Virtosu, I., & Goian, M. (2023). *Disinformation using artificial intelligence technologies – akey component of Russian hybrid warfare*. https://scrd.eu/index.php/scic/article/view/493

Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D., & Bielikova, M. (2023). Disinformation capabilities of large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.08838

Willaert, T., & Tuters, M. (2025). From denazification to the Golden Billion: an inductive analysis of the Kremlin's weaponisation of digital diplomacy on Telegram. *Humanities and Social Sciences Communications*, *12*(1). https://doi.org/10.1057/s41599-025-05382-x