Abstract for the general public

The project explores the intersection of language, racial bias and artificial intelligence (AI) in Poland, with a focus on the Polish language and its users. At its heart, it's about understanding how AI systems, like Large Language Models (LLMs) in Polish, can spread racist stereotypes and biases against people of color. The project aims to shed light on how racial bias is built into the way language and communication shape the understanding of race in Poland. The main goal of the project is to understand what happens when Polish language processing algorithms perpetuate racist stereotypes and racial bias against people of colour. At its core is an exploration of the complex m*ultimodal racialising discourses* which refer to multi-level semantic and pragmatic ideological processes within which race is constructed through linguistic, discursive and communicative practices.

By analysing AI-generated Polish content and gathering insights from diverse Polish-language users through field research, the project highlights the need for fairer and more transparent AI systems in Polish language processing. It looks at models such as the Polish version of ChatGPT and emerging Polish-specific models pre-trained on Polish data such as PLLuM to identify racial biases in AI-generated content and explore how these biases affect real people, particularly those from racialised communities. In relation to these key issues, this project aims to explore the relationship between AI-generated content and *multimodal racialising discourses* that may perpetuate racial bias in Polish large language models.

The project's primary aims are to:
(1) Understand, problematise and conceptualise how *multimodal racialising discourses* that perpetuate racial bias are organised and represented in AI-generated content, with a particular focus on Polish version of ChatGPT and PLLuM;
(2) Investigate how these paradigms are perceived and experienced by ethnically diverse Polish language users;
(3) Explore how Polish versions of AI language models reflect and shape social biases, and how this process contributes to a broader understanding of the social implications of AI technologies in Polish context. The findings will provide important insights into how these technologies can perpetuate or challenge existing stereotypes and inequalities;
(4) Examine how *multimodal racialising discourses* are presented and represented in Polish-language content generated by AI models and how they are experienced by users, particularly those from racialised communities in the country;
(5) Learn how local Polish experts in artificial intelligence and natural language processing perceive and manage these biases in the development of such systems.

Most research in this area has focused on English-speaking contexts, often overlooking how AI behaves in languages with unique grammar and cultural nuances. This project fills this gap by investigating racial bias in Polish AI-generated text and visual content, while also testing ways to make these systems more inclusive. The project uses a mix of methods, combining linguistic, sociological and computational tools to analyse not only text but also AI-generated images, giving a more complete picture of how bias operates in multimodal content.
By exploring these issues, the project will provide a better understanding of how artificial intelligence technologies interact with the social fabric of prejudice in Poland. The results of the research will contribute to a deeper understanding of the mechanisms by which biases operate in the content generated by Polish-language LMMs, as well as providing insights into user perspectives and experiences that can influence more equitable AI use practices in Polish communication. The project also aims to translate research findings into practical knowledge for policymakers, software developers and educators to promote a more inclusive digital landscape, and to deepen awareness and discussion of the ethical responsibilities involved in the development and implementation of modern technologies.

Ultimately, the project aims to set new standards for the study of racial bias in non-English AI systems, providing methods and insights that can guide the development of more inclusive AI technologies. As well as highlighting the risks of biased AI, it will also explore how these technologies can be used to combat bias in Polish language models, paving the way for a more equitable digital future in this geographical context.