[iTRUST] **I**nterventions against polarisation in society for **TRUST**worthy social media: from diagnosis to therapy

Katarzyna Budzynska, Warsaw University of Technology, **Poland** (lead); Marie-Francine Moens, KU Leuven, **Belgium**; Andrea Rocci, Università della Svizzera italiana (Lugano), **Switzerland**; Carles Sierra, Artificial Intelligence Research Institute (IIIA), **Spain**; Virginie Van Ingelgom, UCLouvain, **Belgium**

Digitalisation is rapidly transforming our societies, transforming the dynamics of our interactions, transforming the culture of our debates. Trust plays a critical role in establishing intellectual humility and interpersonal civility in argumentation and discourse: without it, credibility is doomed, reputation is endangered, cooperation is compromised. The major threats associated with digitalisation – **hate speech and fake news** – are violations of the basic condition for trusting and being trustworthy which are key for constructive, reasonable and responsible communication as well as for the collaborative and ethical organisation of societies. These behaviours eventually lead to **polarisation**, when users repeatedly attack each other in highly emotional terms, focusing on what divides people, not what unites them.

Focusing on two timely domains of interest – gender equality and public health – iTRUST will deliver:

- the largest ever dataset of online text, annotated with features relevant for **ethos, pathos and reframing**;
- a new methodology of large-scale comparative trust analytics to detect **implicit patterns and trends** in hate speech and fake news;
- a novel **empirical account** of how these patterns affect polarisation in online communication and in society at large; and
- AI-based applications that will transfer these insights into **interventions** against hate speech, fake news and polarisation.

Given the relevance for the **knowledge-based society**, the project puts great emphasis on outreach activities and users' awareness in collaboration with media, museums and other partners.

The **consortium** consists of five experienced PIs with expertise in rhetoric, comparative political science, corpus linguistics, natural language processing, multi-agent systems and computational argumentation. The group is complemented by senior experts (collaboration partners in academia and industry) in fields that provide valuable extensions, such as media studies and AI-based technologies. Our long-term ambition is to establish a pan-European network and foundations for **trustworthy AI** in response to the *European Commission* priority of "Europe fit for the Digital Age".