WP 8.4: The Social Implications of AI: Data and Models, Acceptance and Use of Generative AI and ADMs

WP8.4.1: ChatGPT as Communication Partner

Goal:

The project is focused on a sociological investigation of the use of Large Language Models (LLMs) in communication, especially on their use in search engines like Bing (and other prototypes). Our approach is based on the notion of "artificial communication" (Esposito 2022) and intends to assess the results of the interaction with algorithms not in terms of machine intelligence or "human parity" (Turing test) but in terms of their communicative competence. In this view, the discussed pitfalls of Microsof's chatbot Sydney and similar models are not errors but the unintended result (and demonstration) of the machine's competence in perceiving and evaluating the attitude of its interlocutor: after hours of intensive conversation with a person the algorithm deduces, for example, that there is personal involvement and makes a declaration of love. And indeed, the measures that have been proposed to counteract these unintended outcomes are communicative and not structural: they do not intervene in the way the machine works but, for example, in the allowed duration of communication.

Methods:

Data scraping and text mining, quantitative experimental studies, qualitative analysis of literature and mediatic debate, ethnomethodological analysis. Our research relies on ongoing collaboration with colleagues at Bielefeld University, with the goal of analyzing the communicative features of Large Language Models (LLMs).

Task 1: Desk research through two main strategies. 1. Literature review and cognitive analysis of the international debate and research on LLMs and their applications - in specialized journals but also in a sample of generalist media;I.2. Data scraping and text mining techniques to map usage and diffusion of Generative AI. Depending on data availability, mapping and forecasting of Generative AI use and diffusion by sectors (public, private, small-medium-large companies, etc.) and/or population subgroups (young, old, men, women, etc.).

Task 2: Quantitative analysis of the relationship between the contextual information that the models gather during the interaction with its interlocutor and the reference dataset, comparing and assessing different LLMs – to be implemented in collaboration with the WP8.8 *Experimental studies on pervasive AI system*;

Task 3: Quantitative analysis of the production and reproduction of bias and data blindness in the use of LLMs (as interaction partner and incorporated in search engines), comparing and assessing different LLMs and the corresponding de-biasing tools – to be implemented in collaboration with the WP8.8 *Experimental studies on pervasive AI system*.

Task 4: Ethnomethodological analysis of the variation of the meaning of linguistic expressions during the interaction with LLMs, investigating "third position repair" contextual adjustments (Schegloff et al. 1977).

Outcomes:

- An analysis, which is not yet available, of the structure and use of LLMs starting explicitly from their communicative features;
- Description and problem analysis of existing models from the perspective of their communicative use and its impact with their computational setting, as well as of the resulting forms of bias and blindness;
- Concrete guidelines for communicatively effective regulatory intervention in the use of LLMs;

- A systematic survey of the international debate and existing research on Generative AI;
- A study of standard socio-demographic variables of interest on the use and adoption of generative AI.

WP8.4.2: ADMs

Goal:

Reconstruction and evaluation of instances and interests of stakeholders (designer, adopter, user/receiver, and general public) who design, adopt, undergo, and accept an ADM. For each of the 4 stakeholders (designer, adopter, user/receiver and general public) the socio-technical reference imaginaries that influence and shape choices and perceptions will be outlined.

Specific to the design phase: different sources of bias in the data, construction and model application steps will be evaluated. The collaboration and synergies that will develop within FAIR will enable the selection of one or more case studies.

Methods:

Mix method approach: ADM's reverse-engineering, ADM audit, in-depth interviews, quantitative survey (representative of the respective population), predictive models and randomized controlled trials.

Task 1: Stakeholder 1 - Designers (designers that model and train the ADM).

- Reverse-engineering of one (or more) ADM (based on the case studies that can be identified in collaboration, particularly with WP8.8 (Experimental studies). This involves the deconstruction of both programming and training process of an ADM, focusing on the design team and the client/adopter (public or private). If there were an opportunity to follow the construction of a WMD from scratch, one could take a qualitative approach made of qualitative techniques (such as participant observation and in-depth interviews) to understand the strategic choices in programming based on the client desiderata and the database selected for training. Thus, potential sources of bias in the design and training steps can be identified.
- Reconstruction of socio-technical imaginaries through a survey (possibly representative of a specific population of designers).

Task 2: Stakeholder 2 - Adopter: public entities (ministries, regions, municipalities etc.) or private entities (SMEs or large enterprises) adopting (in-house, outsourcing or buying it off the shelve on the market) an ADM.

- In-depth interviews and surveys of different actors in this category to understand their motivations, evaluations, opportunities, awareness.
- National survey of sociotechnical imaginaries.
- Predictive models of the use of ADMs in the coming decades in the public and private sectors to predict the evolving role of ADMs in decision-making processes, also in light of the potential role that other novel types of AI, such as generative AI, may play.

Task 3: Stakeholder 3 - User/receiver (Recipienti della decisione). Awareness, acceptance, and trust toward ADMs output. Selection of individuals who are affected by the decision (e.g. job seekers or public services' recipients).

- Survey of sociotechnical imaginaries (if the available sample allows for it)

- Randomized experiments by comparing homogeneous groups subject (treated) and not subject to (control) ADMs' decisions. We want to test whether it is also true in the Italian case (there is similar research in France and the U.S.) that individuals judge humans by their intentions and machines by their outcomes (the experiment concerns the evaluations of those who have been the subject of an ADM in their job search and CV evaluation. Another option is to test people's acceptance of driverless cars in novel urban mobility scenarios, eg. in collaboration with Task 8.8.2).

Task 4: Stakeholder 4 - General public.

- Representative survey of the Italian population (not yet available in Italy) that reconstructs sociotechnical imaginaries, awareness, trust, hope and fears.
- Two types of experiments: 1. standard RCT comparing treated and control and 2. innovative approach using a survey instrument to elicit counterfactuals (Aucejo et al 2020).

Outcomes:

This combination of techniques and methods will allow a clear picture of:

- what is the relationship between ADMs and the reproduction of social inequalities (identification of sources of bias all along the pipeline, upstream and downstream of the algorithm);
- what social and institutional dimensions are to be considered for fair and equitable ADMs in different sector and case studies;
- how to build a solid and representative governance of all stakeholders;
- how to raise awareness of the social implications of ADM in different social groups and in the general population (based on their sociotechnical imaginaries)
- what factors explain the use and adoption of ADMs by different stakeholders in selected industries and sectors;
- the predicted impact of the ADMs on the labor market in the coming decades.

References

Aucejo et al 2020
Esposito (2022)
Schegloff et al. 1977.