3rd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies

Esma Aïmeur[1], Nicolás E. Díaz Ferreyra[2]

[1]Department of Computer Science and Operations Research, University of Montréal, Canada
[2]Institute of Software Security, Hamburg University of Technology, Germany

1. Preface

The rapid advancement and growth of Information and Communication Technologies in general and Artificial Intelligence (AI) in particular; has led to the seamless and yet indispensable integration of such technologies into our everyday activities.

Indeed, over the last decade, AI has infiltrated many aspects of our lives: people rely on it while driving or training; or when selecting which movie/song to play next, even when asking information about the weather or current traffic conditions. Moreover, individuals rely heavily on intelligent software applications across different domains including healthcare, logistics, agriculture, finance, education, defence, and governance. Particularly, AI systems facilitate decision-making processes across these domains through the automatic analysis and classification of large data sets and the subsequent identification of relevant patterns. To a large extent, such an approach has contributed to the sustainable development of modern societies and remains a powerful instrument for social and economic growth. However, recent events related to the discovery of biased AI, the massive spread of misinformation and deepfakes along with fears of AI powered autonomous weapons, have raised concerns among AI practitioners and researchers about the negative and detrimental impacts of these technologies. Indeed, like any other technology, AI can have some seriously negative consequences, whether intentionally or inadvertently.

Consequently, and due to the ubiquity of AI and the increasingly rapid rate of its development and adoption, there is an urgent call for guidelines, methods, and techniques to assess and mitigate the potentially adverse impacts and side effects of AI applications.

Workshop Objectives

This 1st Workshop on Adverse Impacts and Collateral Effects of AI Technologies (AiOfAi '23) is co-located with the 32nd International Conference on Artificial Intelligence (IJCAI-23). The objective of the workshop is to bring together experts and practitioners to explore how and up to

AiOfAi'23: 3rd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Macao, China

aimeur@iro.umontreal.ca (E. Aïmeur); nicolas.diaz-ferreyra@tuhh.de (N. E. Díaz Ferreyra)
http://www.iro.umontreal.ca/~aimeur/ (E. Aïmeur); https://www.ndiaz-ferreyra.com/ (N. E. Díaz Ferreyra)
0000-0001-6304-771X (N. E. Díaz Ferreyra)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3593/preface.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

which extent AI technologies can serve deceptive and malicious purposes whether intentionally or not. Furthermore, it seeks to elaborate on guidelines, countermeasures and mitigation actions to prevent potential negative effects and collateral damages of AI systems. We therefore invited AI researchers and practitioners across different disciplines and knowledge backgrounds to submit contributions dealing with the following (or related) topics:

- Hazardous AI applications:
  - Deepfakes.
  - Fake news and misinformation.
  - Online deception.
  - Malicious personalization.
  - Social engineering.
- Adverse impacts of AI:
  - Privacy and security breaches.
  - Backfire effects.
  - Guidelines and mitigation actions.
  - Ethical conflicts and challenges.
  - Risk assessment methods.
- Responsible AI:
  - Case studies.
  - Best practices for trustworthy AI.
- Generative AI tools:
  - ChatGPT, AIVA.
  - Stable Diffusion, DALL-E, AutoDraw

Special topics of interest:

1. **Crises and AI**: Crises like the COVID-19 pandemic and the war in Ukraine have plunged the world into a state of crisis. Although AI technologies can significantly contribute to mitigating the collateral effects of such catastrophic events (e.g., by helping administrate humanitarian aid), they can also act as lethal weapons and facilitate misinformation. Such is the case of autonomous drones capable of recognizing, selecting, and killing human targets; or biased news recommender systems polarizing public opinion. AiOfAi welcomes submissions elaborating on such controversial applications of AI technologies from an interdisciplinary and multistakeholder perspective.
2. **AI Regulations**: The new regulatory framework for AI systems drafted by the European Commission seeks, for instance, to promote profound changes in the way such systems are developed and deployed. Still, many challenges are upfront, particularly when it comes to the identification and assessment of risks potentially linked to AI solutions. We encourage the submission of papers elaborating on regulations, guidelines, methods and tools for assessing the risks of AI systems and their possible adverse impact on both individuals and societies at large.

2. Accepted Papers

Seven papers were submitted and peer-reviewed by 3 members of the program committee in a single-blind process. Out of these, six papers were accepted for this volume, two as long papers, two as short papers, and two as invited papers.

1. *On the Importance to Study Fringe Social Networks and Their Impending Use of GenAI to Promote Mal-Info: Gab as a Case Study* (short)
   Florian Barbaro and Andy Skumanich
2. *From Hype to Reality: Transformer-Based Models for Fake News Detection Performance and Robustness Revealed* (long)
   Dorsaf Sallami, Ahmed Gueddiche and Esma Aïmeur
3. *AI in Healthcare: Impacts, Risks and Regulation to Mitigate Adverse Impacts* (short)
   Retno Larasati
4. *Disincentivizing Polarization in Social Networks* (long)
   Christian Borgs, Jennifer Chayes, Christian Ikeokwu and Ellen Vitercik
5. *A GPT-based Practical Architecture for Conversational Human Digital Twins* (invited)
   Bart Knijnenburg and Nina Hubig
6. *The role of AI in the misinformation ecosystem: virtuous or vicious?* (invited)
   Claudia Baptista Claros and Kimiz Dalkir

3. Invited Talks

One keynote talk was included as part of AiOfAi's technical programme.

Keynote - "Evaluating GPT-3 Generated Explanations for Hateful Content Moderation"

Roy Ka-Wei Lee | Singapore University of Technology and Design

Recent research has focused on using large language models (LLMs) to generate explanations for hate speech through fine-tuning or prompting. Despite the growing interest in this area, these generated explanations' effectiveness and potential limitations remain poorly understood. A key concern is that these explanations, generated by LLMs, may lead to erroneous judgements about the nature of flagged content by both users and content moderators. For instance, an LLM-generated explanation might inaccurately convince a content moderator that a benign piece of content is hateful. In light of this, we propose an analytical framework for examining hate speech explanations and conducted an extensive survey on evaluating such explanations. Specifically, we prompted GPT-3 to generate explanations for both hateful and non-hateful content, and a survey was conducted with 2,400 unique respondents to evaluate the generated explanations. Our findings reveal that (1) human evaluators rated the GPT-generated explanations as high quality in terms of linguistic fluency, informativeness, persuasiveness, and logical soundness, (2) the persuasive nature of these explanations, however, varied depending on the prompting strategy employed, and (3) this persuasiveness may result in incorrect judgements about the hatefulness of the content. Our study underscores the need for caution in applying LLM-generated explanations for content moderation.

4. Organization and Committees

Workshop Organizers

- **Esma Aïmeur**
  - University of Montréal, Canada
  - Website: http://www.iro.umontreal.ca/~aimeur/
  - Email: aimeur@IRO.UMontreal.CA
- **Nicolás E. Díaz Ferreyra**
  - Hamburg University of Technology, Germany
  - Website: http://www.ndiaz-ferreyra.com/
  - Email: nicolas.diaz-ferreyra@tuhh.de

Publicity Chair

- **Dorsaf Sallami**
  - University of Montréal, Canada
  - Email: dorsaf.sallami@umontreal.ca

Programme Committee

- Gabriel Pedroza (CEA-LIST, France)
- Juan Carlos Nieves (Umeå University, Sweden)
- Jean-Gabriel-Ganascia (Paris-Sorbonne University, France)
- Josep Domingo-Ferrer (Universitat Rovira i Virgili, Spain)
- Daniela Godoy (ISISTAN CONICET-UNICEN, Argentina)
- Timotheus Kampik (Umeå University | Singavio GmbH, Sweden)
- Alison R. Panisson (Federal University of Santa Catarina, Brazil)
- Julita Vassileva (University of Saskatchewan, Canada)
- Bart Knijnenburg (Clemson University, United States)
- Stefan Stieglitz (University of Potsdam, Germany)
- Sibel Adali (Rensselaer Polytechnic Institute, United Satates)
- Abdessamad Imine (University of Lorraine, France)
- Antonela Tommasel (ISISTAN CONICET-UNICEN, Argentina)
- Yuanpeng Zhang (Nantong University, China)
- Xiangmin Zhou (RMIT University, Australia)
- Pamela Briggs (Northumbria University, UK)
- Steven Furnell (University of Nottingham, UK)

- Reda Yaich (IRT SystemX, France)
- Ronald Arkin (Georgia Institute of Technology, United States)
- Paolo Rosso (Universitat Politècnica de València, Spain)
- Jeremy Clark (Concordia University, Canada)
- Michael Floyd (Knexus Research Corporation, United States)
- Fang Yu (Jinan University, China)
- Hella Kaffel (Faculté des sciences de Tunis, Tunisia)
- Maryline Laurent (Institut Mines-Telecom, France)

Acknowledgments

This work was partially supported by Canada's Natural Sciences and Engineering Research Council (NSERC).