Report on the key findings from the Theme Development Workshop "AI Mitigating Bias & Disinformation"

– November 2022 –

Executive Summary

The first cross-cutting Theme Development Workshop (TDW), co-organised by the ICT-48 projects AI4Media, Humane-AI-Net, TAILOR, and CLAIRE AISBL, under the lead of VISION on "AI: Mitigating Bias & Disinformation" took place on the 18th of May 2022 with the aim to discuss the importance and the use of AI to mitigate bias and disinformation. At this one-day workshop, experts from academia, industry and politics jointly developed initial input for the European Artificial Intelligence (AI) research and innovation roadmap. Inspired by introductory speeches and presentations from selected experts, the participants actively discussed a wide variety of topics during the breakout sessions and shared their main results in the subsequent plenary presentations. Furthermore, some initial ideas for follow-up activities and further collaborations have been identified.

This report contains a summary of the results from the Theme Development Workshop "AI: Mitigating Bias & Disinformation". To make the results available to a broader audience and the European AI community in particular, this report will be published via the organiser's websites.

Authors of this report are the members of the TDW Organising Committee:
(in alphabetical order)

- Kyra Bare, German Research Center for Artificial Intelligence (DFKI) & CLAIRE Office Saarbruecken, Germany
- Ricardo Chavarriaga, ZHAW Zurich, Switzerland
- Frank Dignum, Umeå University, Sweden
- Emanuela Girardi, Pop AI, Italy
- Christian Grimme, European Research Center for Information Systems (ERCIS), Germany
- Janina Hoppstädter, German Research Center for Artificial Intelligence (DFKI) & CLAIRE Office Saarbruecken, Germany
- Andreas Keilhacker, German Entrepreneurship, Germany
- Yiannis Kompatsiaris, CERTH-ITI, AI4Media Project Coordinator, Greece
- Philipp Slusallek, German Research Center for Artificial Intelligence (DFKI) & CLAIRE, Germany
- Georg Thallinger, Joanneum Research, Austria
- Marlies Thönnissen, German Research Center for Artificial Intelligence (DFKI) & CLAIRE Office Saarbruecken, Germany
- Heike Trautmann, European Research Center for Information Systems (ERCIS), Germany

Authors of this report outside the TDW Organising Committee:
(in alphabetical order)

- Dennis Assenmacher, GESIS Leibniz Institute for the Social Sciences, Germany
- Julia Metag, University of Münster, Germany
- Symeon Papadopoulos, Centre for Research and Technology Hellas (CERTH), Greece
- Eugenia Polizzi, National Research Council (CNR-ISTC), Italy

The authors would like to thank Nicolas Sponticcia (DFKI & CLAIRE Office Saarbruecken) for his support in generating this report.

Table of Contents

Introduction

In September 2020, four new AI networks were established by the European Commission via the call "Towards a vibrant European network of AI excellence centres" (ICT-48-2020). The aim of these networks is to foster the collaboration between the best research teams in Europe, and to address the major scientific and technological challenges in the field of AI. These four networks are coordinated and supported by the VISION project to foster activities that reach critical mass and enable the creation of a world-class AI ecosystem in Europe.

One of these activities are so-called Theme Development Workshops (TDWs), an innovative format bringing together key players from industry, academia and politics to jointly identify the key AI research topics and challenges in a certain area or for a specific industry sector. In December 2020, an agreement was made between the respective coordinators and leadership teams of TAILOR, VISION, HumanE-AI-Net and CLAIRE to plan and execute a series of Joint (co-organised) Theme Development Workshops, starting in 2021. This report is a result of the fifth Joint TDW organised and executed within the framework of this series of workshops.

Keynotes and introductory presentations

The TDW was opened by VISION Coordinator Holger Hoos and the Co-Chairs Emanuela Girardi (Pop Ai) and Yiannis Kompatsiaris (AI4Media) on behalf of the Organising Committee (OC), which included further representatives from AI4Media, CLAIRE, DFKI, ERCIS, German Entrepreneurship, Joanneum Research, Pop AI, University of Münster and ZHAW Zurich. The Co-Chairs outlined the objectives of the TDW as well as the agenda and programme, and introduced the invited keynote speakers to the participants.

The inspiring keynotes were provided by high-level experts from academia and industry. These introductory presentations served as the foundation for discussions on avoiding bias in data and misinformation in the field of AI, and provided some interesting examples of application areas. Accordingly, these presentations stimulated the expert discussions in the following breakout sessions.

Introductory presentations by Professor Dr. Sander van der Linden, Professor Dr. Virginia Dignum, Professor Miguel Poiares Maduro and Mijke van den Hurk.

Professor Dr. Sander van der Linden, University of Cambridge gave his introductory keynote on the issue of fake news. He defined the important difference between misinformation, which is understood to be false or incorrect information, and disinformation which describes false information that has been purposefully spread to deceive others. A popular solution that has been proposed to solve the issue of disinformation is fact-checking, however, there are several psychological limitations that limit the efficacy of this solution. One of these limitations has been labelled the continued influence effect, which describes the observation that people continue to retrieve false information from their memory, even when they acknowledge that they have seen a correction of that disinformation. To solve this limitation, Professor van der Linden has focused on pre-bunking which aims to protect people from disinformation before they even come into contact with it. The idea of psychological inoculation functions similar to vaccines, as it may be possible to protect people from misinformation by either warning them of the fact that they are about to be misled or by pre-emptively providing them with the correct information, if false information about an issue is currently being spread. However, just with fact-checking, there are issues of scaling this solution, as anticipating each new misinformation trend is incredibly difficult – there are, however, opportunities for automation.

Professor Dr. Virginia Dignum, AI Tech Center at ZF Group, **Umeå University** gave her keynote on the issue of responsible artificial intelligence. She started her keynote by defining what artificial intelligence is by distinguishing it from algorithms and data. Following this, Professor Dignum discussed the pitfalls of learning from data, with specific reference to the tendency of bias and discrimination included in data being passed onto AI systems. Thus, if

there are errors in the data that the AI system learns from, these will be inherited. Furthermore, Professor Dignum highlighted three major concerns related to the development and use of AI. The first being datification, which describes the fact that people are more than just data. The second concern being power, as it may sometimes be unclear who and for what purpose certain AI are developed. The last concern being sustainability, as AI consumes a lot of power and is very expensive to develop. Lastly, Professor Dignum presented various guidelines that may be used to develop responsible AI.

Professor Miguel Poiares Maduro, Chair of the European Digital Media Observatory (EDMO) gave his introductory keynote on the issue of disinformation. He highlighted recent changes to the speed with which information can be processed and shared. This difference in speed with which information can be shared is not necessarily a bad thing in itself, however, it has also increased the scale at which disinformation is distributed. Fake accounts are especially dangerous as they lend false credibility to disinformation and increase the chance that people that come into contact with it come to believe it and share it further. Furthermore, in the virtual public sphere, editing of information is no longer performed by trained professionals such as journalists, but instead by algorithms. This is important as algorithms tend to create information bubbles, as algorithms tend to only suggest new information that is linked with our previous interactions with other information. Additionally, information can now be targeted and edited for certain groups, for example politicians can target certain groups online with information that is specifically edited to suit their preferences to generate support from them. Combined, these issues bring new important challenges for democratic systems that need to urgently be addressed. Mr Maduro suggests that one solution could be to reintroduce editorial processes in the digital public sphere. An example could be to give users the choice of algorithm that suggests information to them – this could include the choice of an algorithm that gives the consumer a broader selection of information to avoid the creation of further information bubbles.

Mijke van den Hurk, Police of The Netherlands & Utrecht University started her keynote by highlighting the issue of defining who a terrorist might be and how broad categorisations may lead to many false positives. Such false positives are particularly important to avoid as they make identifying real terrorists far more difficult. Next, Ms. van den Hurk discussed the question of whether it is possible to predict an attack in the first place. She highlighted the issue of the Black swan effect which makes identifying possible attacks more difficult as many people might falsely threaten to carry out an attack. This translates into very large amounts of data being generated that the capacity to analyse for does not exist. However, at the same time there is also a lack of data, as many extremists may move in very shielded groups that are hard to track and quantify. Furthermore, many attacks may be carried out impulsively, which are particularly difficult to predict. Next, Ms. van den Hurk discussed the fact that implementing AI algorithms is very difficult for the police. One reason for this is the lack of explainability of machine learning and deep learning, which relates to the idea that

decisions made by an algorithm to flag a certain individual must be explainable in court – this is often not possible at the moment. Another reason is where to set acceptable boundaries for error, i.e., how many false positives or negatives is an AI system allowed to make. One solution may be to shift attention from the individual to the group, as radicalisation is a process and happens over time in groups it may be possible to instead focus on indicators of radicalisation within groups – such as social structure and group behaviour. Furthermore, a shift from prediction toward prioritisation would be particularly helpful – this is as prioritisation of resources is expected to be much more effective in comparison to prediction which is often inaccurate. Ms. van den Hurk ended her keynote by highlighting that AI can be useful in the context of the police if the right research questions are used, police rules are followed, experts are involved to develop indicators and the right sources of data are identified.

Key results from the Breakout Sessions

The "arms race" nature of DeepFake detection

The session focused on the issue of new generation DeepFake detection models that can evade their detection, leading to an "arms race" of AI methods and possible ways to stop or at least slow down this arms race by regulatory measures. The session was attended by participants of different backgrounds, including computer scientists, journalists/media professionals and industry professionals.

A first important finding of the session pertains to the gap in terminology and concepts between AI experts and media professionals such as journalists and fact checkers. It appears that the issue of deepfakes is only now starting to be discussed among media professionals, and is still considered as a future challenge (given the relatively limited number of real-world cases). More importantly, it appears that journalists have different levels of understanding regarding what is classified as deep fake content and what types of deep fake content exist. Their main interest seems to be on establishing the veracity and authenticity of media content independent of the underlying technical and conceptual details. Important steps towards increasing the level of understanding and clarity about the deepfake technologies could include the creation of simple and easy-to-understand training material for media professionals, the organisation of multi-disciplinary events bringing together media professionals and deep fake AI experts, and the development of tools that are trustworthy and transparent and support the analysis and verification of deep fake content.

A second important finding mainly coming from the AI experts pertains to the trends in deepfake generation and detection methods. On the front of deepfake generation models, it appears that diffusion-based models are now surpassing GAN-based methods in terms of realism and quality. In terms of detection approaches, a variety of approaches seem to be necessary, including for instance fingerprinting approaches, data augmentation (for more robust training), and person-specific biometric/semantic approaches. Fusing among different approaches and multiple modalities could be a helpful step to address the issue of adversarial attacks, decrease the number of false positives (i.e. mistakenly labelling authentic videos as deepfakes), and the big challenge of generalising to new types of deepfakes. Last, it appears that more research is needed on the front of explainability in order to make the results of deepfake detection more trustworthy and useful for journalists and fact-checkers.

The session concluded with the realisation that the challenge of deepfakes calls for a hybrid machine-AI solution, increased interaction among different disciplines (AI experts, media professionals), and more resources, including both training material and an extensive repository of labelled/documented cases from the real-world.

Explainability aspects in AI for disinformation

The session was attended by participants whose expertise covered artificial intelligence, computer vision, recommender systems, and journalism. It started with the invited experts' presentations of relevant aspects of explainability. They underlined the complexity of the topic and made clear that actionable explanations depend on the stakeholders, type of conveyed explanation, types of AI models used, and nature of the analysed data. The discussion then moved to the needs for automated explanations in the journalism domain, with particular focus on tackling disinformation.

A first important finding is that explanations should be tailored both to the level of technical expertise of journalists (and, more generally, of final users) and to the specific context in which explainability tools are used. Journalists take an evidence based approach when dealing with disinformation and a counterfactuals-based approach might be appropriate to assist them. However, a component which handles novelty should be included, and this is not straightforward since existing methods tend to assume that the data space is frozen. The need for an interactive explanation process was also underlined by final users. For instance, several counterfactuals could be aggregated into a sequence in order to assist journalists. The use of a neurosymbolic approach, which mixes logical and statistical methods, was evoked as an promising path toward improved explanations. On the other hand, it is also important to match explanations to the expertise on the topic, as confirmation bias can also occur among experts.

A second finding is that progress is needed towards providing more reliable and stable explanations. Reliability could be improved by a combination of better algorithms and more adequate training data. The quantity, quality and unbiased character of training data need to be ensured in order to obtain usable explanations. There is also a stability-plasticity dilemma which needs to be solved since provided explanations should follow the changing understanding of the world. This was illustrated with how knowledge about the COVID pandemics or recent wars has evolved through time.

Third, the need for better qualitative and quantitative evaluation of algorithms was discussed. Journalists pointed out that purely visual explanations are difficult to use and should be complemented with textual ones, which are easier to grasp by non-technical users. Shared evaluation exercises should be encouraged in order to have a fair quantitative evaluation of explainability algorithms.

The takeaway of this session is that, in contrast with the mainstream technical orientation of explainability-related work, the practical implementation should be done in a user-centric way. Aspects such as interactivity, cost/benefits of explanations, novelty, and bias reduction should be given more weight in the future.

Science Communication with and on AI

The session started with discussing the challenges when communicating scientific findings on AI to the broader public. One of the key aspects was that there is the perception that media coverage of AI predominantly focuses on fear and that certain narratives are very prominent in the media discourse, with robots and anthropomorphism in the context of AI as dominant ones. Books and fiction serve as some of the most important sources for these understandings of AI but nowadays also YouTube as a source is not neglectable. Particularly, anthropomorphism was considered as problematic from a science communication perspective. Also, economic interests behind AI results in the fact that strategic communication on AI as a topic is very dominant. This includes advertising AI, playing down problems associated with the implementation of AI and little interest in a politicised discourse on this topic.
In the following, it was concluded that a collective understanding of artificial intelligence is important. To reach this collective understanding - which does not exist at the moment as data from the Me:Mo:Ki-project has shown - should be one of the main aims of communicating AI. Research so far shows that the audience of AI as a topic is only a very small part of the public and that the general public has very little knowledge about AI. Many parts of the public are not reached by communication about AI. It was also discussed that the term “AI” is associated with some kind of pre-understanding and pre-existing beliefs which need to be considered (e.g. with regards to motivated reasoning processes) when communicating about AI. This can also be a problem in the context of dis- and misinformation about AI. With regards to dis- and misinformation, the role of AI for creating and disseminating false information was discussed and opposed to the potential of artificial intelligence to detect false information.
Based on this discussion and existing research findings, several recommendations were developed: (1) Raise awareness of omnipresence of AI in the current society to make its everyday implementation clear. This can also include more direct communication by scientists working in these fields and at the same time bringing the discourse out of the scientific bubble. (2) Target different audiences with science communication on AI, e.g. pupils and students who can serve as multipliers. (3) Provide new and different images and narratives, both textual and visual ones, when communicating AI. (4) Use AI to support science communication and to detect false information. Automated journalistic procedures can also help science journalists in their reporting. (5) Understand communication about AI as not solely technical and factual information but contextualised in political, ethical, moral views etc. This helps tailoring AI communication to people’s specific pre-existing attitudes and worldviews.

A social cognitive perspective to AI and misinformation

Very much in line with the keynote talk given by Sander van der Linden, this session started with an introduction about some of the social biases that make people more susceptible to misinformation.
Just as in real world scenarios, also in human-AI hybrid systems people are likely to follow social norms, informal rules that prescribe or proscribe behaviours and guide our decisions according to what we perceive our peers do or think is appropriate to do.
In such systems however, the combination of structural (e.g., algorithmic, network features) and cognitive factors can easily distort such perception (e.g., “illusion of majority), possibly leading to overestimating support for unpopular opinions and reinforcing the persistence of biases in the system. This condition can influence people's likelihood to express opinions not in line with the (misperceived) majority, not taking actions against others engaging in undesirable behaviours (e.g. sharing fake news) and make biases more resistant to correction due to users (mis)perceiving a weaker norm against it. Recognizing these elements, e.g. by integrating research on social norms into the specific context of social interactions in human-AI hybrid systems, can help to better understand and limit the likelihood that biases spread.
From an AI perspective, one big challenge is how to do this in practice, e.g., building tools that help AI systems to “understand” human social rules, that recognize potential social biases, and possibly correct their effect on the system. Several important points were raised during the discussion. First, the research community is struggling with access to data: there is a general tendency to lean towards specific online platforms that provide less restrictions about data access but also with limited and selected user groups (e.g., Twitter). There is also a difficulty in combining data coming from different sources, such as controlled field work (e.g., experimental data gathered in labs, which often give information on micro/meso-scale aspects of human behaviour) with data collected in “the wild” (Big data, macro-scale) . Importantly, sharing of data is limited by privacy and legal issues (TOS of the different platforms). All these elements strongly constrain machine learning models and the generalizability of the findings.
Secondly, participants uniformly agreed that long-lasting changes can be achieved only by looking at the problem from a “systemic” point of view, which should also include education and training programs to improve people ability to recognize biases (and the mechanisms leading to their emergence) and avoid getting trapped in misinformation bubbles. In this regard, gamification tools as the one presented by Sander van der Linden seem to be a very promising and concrete example of currently available tools. In addition to recognizing biases, it is also important to know the magnitude of bias, because otherwise an overcorrection may occur. Lastly, a stronger collaboration between social sciences and AI communities overall is very much needed. By building theoretically-sound and empirically-grounded tools we can aim at keeping a healthy and unbiased online environment for public debates, and ultimately, increase the resilience of our societies.

Abusive Language Detection and Comment Moderation

Online media platforms currently invest substantial resources to moderate harmful online content and keep their comment sections clean. This is even amplified by the fact that unlawful content must be removed immediately, as mandated by law (e.g., Network Enforcement Act in Germany). While in the past, content moderation was conducted manually or via simple means, like relying on vocabularies of blacklisted words, the advent of deep-learning architectures promises a shift towards automation. In this breakout session, the participants discussed this direction's potential and associated risks. We got exciting insights from academia on the work of moderators on Reddit. Also, a former community manager of a large German newspaper reported how they incorporated ML models in their semi-automated moderation process.
While semi-automation for content moderation is a promising direction, the breakout session participants uniformly agreed that substantial unresolved challenges must be addressed interdisciplinary. These include:

- Dataset quality: Datasets often consist of unintended biases (e.g., spurious correlations) and disagreement in the annotation task. Datasets degrade over time. They become inaccessible due to data sharing restrictions leading to model incomparability because of distributional changes. This especially holds for "harmful" content such as abusive language. Additionally, there is a clear bias towards specific platforms (e.g., Twitter), which are not as restrictive when it comes to sharing content with the research community, leading to a bias towards specific user groups. Furthermore, we observe that existing datasets only focus on a uni-modular setting: the text of a comment. Contextual information (e.g., meta-data such as images) is often not considered.
- Dataset accessibility: It is hard to publish and share datasets because of privacy issues and the ToS of different platforms. This not only impedes training machine learning models, but also testing their applicability to realistic and up-to-date scenarios.
- Construct definition: There is no generally-accepted definition of Abusive Language/Hate-Speech. It is a subjective task that depends on individual and cultural background. Furthermore, what is moderated on social media only partially overlaps with definitions of abusive language/hate speech provided in legal frameworks. While research slowly moves away from the traditional binary setting (hate vs. no-hate), there is still room for improvement in conceptualising the construct and in translating it to real-world occurrences in an actionable way.
- Transparency in moderation decision: Usually, it is not clear to the users what content was subject to human evaluation and the exact reasons why comments were removed or kept on the platform. This intransparency has negative consequences for democratising moderation decisions as well as grounding the automation of moderation in explanations and rationales.

- Interpretability: The output of an ML model is often a binary decision or a likelihood of a comment belonging to a specific class. Moderators should be able to understand the decisions made by the model without having profound knowledge of machine learning.

As a consequence of the identified key challenges, researchers should ground their modelling in a deeper understanding of the construct. This can be achieved by incorporating knowledge from disciplines such as psychology for discriminatory attitudes, or definitions, e.g., in legal frameworks. Due to the inhomogeneous definitions of the construct and the resulting diverse annotation strategies, an intermediate solution could be the creation of mapping for different construct definitions, enabling model comparison.
One of the key challenges is the creation of quality data from diverse sources, minimising different manifestations of bias and including as much relevant meta-data as possible. This can only be achieved if researchers from different disciplines and platforms work together.

Automation in Online Media

The breakout session started with three impulse talks by Stefano Cresci, Ralf Lüling, and Mike Preuss, who brought in their three perspectives on automation – the viewpoints of detection, content generation (language and image generation models) and content generation in the context of games. The experts and participants identified several challenges: (1) Automation in social media (sometimes also called bots) is evolving towards more credible content representation and distribution. This poses specific problems to current (often simple and unreliable) detection mechanisms and challenges the current research direction of detecting automation account-based. Since AI tools make automation more and more indistinguishable from genuine human accounts, research should focus on detecting coordination and identify individual actors in a top-down manner. (2) Generative models are very good but have a lack of self-consciousness in AI models. Nevertheless, the rise of multimodal deception attempts (e.g., combination of image and text) can be a next grand challenge. However, it is still unclear how good generated content is in general. If it has to be selected manually, (massive) human interaction remains necessary and will make content generation less attractive in terms of costs. (3) A major challenge is the evaluation of the performance of large language models. From both the production point of view and the risk assessment point of view it is important to know the quality and fit of generated text regarding the content and the message conveyed in it. (4) A challenge that addresses human perception beyond a pure technical focus is the question whether human users of social media are aware of being part of a virtual environment. They enter a technical infrastructure, which enables global communication with other accounts (not necessarily equivalent to human beings) but cannot be sure whether anything else is real. While this discrepancy of reality and simulation is usually clear in gaming environments, the environment of social media is often considered as “real”. (5) Finally, the challenge of

intervention addresses the methodological question of detection mentioned in (1) but also implies a danger of surveillance by AI-based methods that (again automatically) try to fight or at least control automation and coordination. During the discussion research in multiple disciplines was identified as a solution path for addressing the named challenges. It was considered essential to study more intensively how effectively automation can be used in the context of disinformation. The focus here is on the question of what developmental leaps automation in social media can make as a result of technology leaps in AI. At the same time, the influence that recommender systems in social media can have on the spread of disinformation and the extent to which these systems themselves can be influenced and misused must be investigated. It is equally important to examine how and in what environment disinformation is received and how it persuades (how do people consume and process disinformation and to what extent is the human psyche designed to construct narratives in order to perceive even automated communication as real?).

Measuring Polarisation, Radicalisation, and the emergence of Echo Chambers in online debates

In this session, the participants discussed how to identify, measure and characterise the interplay between users' social interactions in social media and the content they share and consume, also in terms of polarisation and radicalisation. It was talked about how polarisation can arise in several domains, i.e. mobility, public debate or economy, and that network science can help capture the aggregated behaviours that result from it. But biases and instabilities bear the risk of leading users in harmful directions like polarisation or echo chambers, which asks for an increased cooperation with digital platforms to avoid confounding variables, avoid sampling problems and test hypotheses. In this regard, algorithmic transparency for commercial systems would be ideal but changes in algorithms or issues of scaling make it extremely difficult to achieve this.
The key challenges identified in this session relate to the definition of effective measures to counter misinformation and to foster cooperation with digital platforms to promote data access for research and to increase transparency as well as the need for an external measure of the Rec Sys platform to elaborate on the possibilities of disentangling the contribution of users and of Rec Sys on polarisation phenomena. In order to solve these challenges, a better understanding of the emergent phenomena due to the interplay of users and machines like AI or Rec Sys is needed. Additionally, more data is needed to conduct experiments in different settings and domains, and the development of a standard baseline would enable comparing results, for example on the question of how much polarisation is taking place regarding a specific topic. The participants of this session agreed on the fact that a collaboration with social platforms is needed and to find ways to make results robust to allow for comparative analyses.

Dataset sharing and governance in AI for disinformation

This session focused on the balance between ethical, regulatory and technical aspects of dataset sharing, using disinformation as a guiding use case. Session participants brought expertise from both the regulatory and ethical field and the AI practitioner field. A first realisation among the session participants was the lack of clarity with respect to the limits of data access and sharing when carrying out research e.g. for studying the problem of online disinformation. Online platforms' terms and conditions and public application programming interfaces (APIs) are often confusing and don't address research community needs. An example of Facebook shutting down the accounts of researchers working on the AdObservatory at New York University (NYU) shows that APIs access can be restricted or eliminated at any time and for any reason.
It was noted that legal and regulation expertise would be extremely valuable for AI teams in order to better understand the limits and compliance aspects of their data access rights. The participants found it in particular relevant in light of the recent Disinfolab decision by the Belgian Data Protection Authority which imposed a fine on EU DisinfoLab – an NGO that fights disinformation – and on one of its researchers for violating the GDPR for publishing raw data from Twitter as part of a research study. It was concluded that there is a clear need for a legally binding data access framework at the EU level that provides researchers with access to a range of different types of platform data and guidance for researchers how to access and share datasets in a GDPR-compliant manner.
Legal and regulatory solutions have been discussed such as Art. 31 of the Digital Services Act (DSA) or EDMO's Code of Conduct on access to platform data. A promising proposal that was discussed was the establishment of a federated infrastructure at a European level that would be able to offer dataset access for research purposes under the proper safeguards and protection measures. This would be conceptually similar to the CERN model, i.e. a model for pooling resources under a single infrastructure that enables new types of research that are not feasible without it.
Another interesting paradigm that was discussed included the data donation model, i.e. to involve Internet users as an active stakeholder in the data collection and sharing process. This could be an additional avenue for data collection that could be combined with a federated data sharing infrastructure as the one described above.

What is bias and when is it bad?

This session focused on how bias can be beneficial. It is often assumed that bias is bad, but it is relative to some criteria that can change over time or context. These criteria were discussed in more detail in the session, e.g., regarding gender bias. The session started with a presentation of a use case from a media group in Switzerland, where an AI tool is used to identify personal entities in texts. The tool identified male and female entities, based on a curated list of names that are attributed to a specific gender, to show journalists if their texts

are biassed. The tool identified that most news articles are about male entities, so the feedback to the journalists was able to diminish that effect.
The discussion continued with the question of what bias is and how (bad) bias can be defined. There was common agreement that it is almost impossible to develop unbiased algorithms due to existing preconceptions and biases in the data that is used. An idea that was then discussed was to use recommender systems, i.e. in social media, to reflect desirable societal norms.
The key challenges identified in this session relate to the definition of acceptable and unacceptable bias, to promote transparency in algorithms and to promote transparency in recommendation systems. In order to overcome these challenges, the group agreed that there needs to be a debate about social norms in society and that tools need to be developed to obtain more transparency in algorithms. Also, to make the analysis of large amounts of data easier, the analysis needs to be focused, simplified and summarised.

Unformation vs. Disinformation?

The focus on this breakout session was on the concepts of "unformation" vs. disinformation. False information leads to disinformation, but information can be far more misleading if it does not contain the whole story. This "unformation" gives an incomplete view of reality and leads to misconceptions, for example in war propaganda.

SafetyTech

The field of SafetyTech (Technology to protect people from online harms) is nascent and despite a growing number of companies associated with it, developing viable SafetyTech solutions poses a series of challenges, the main one being: How can we ensure that SafetyTech solutions are designed in a way that their positives outweigh their negatives? Other challenges include accurate regulation that holds technology providers accountable without disincentivizing innovation, whether a taxonomy is possible, how misuse can be controlled and prevented and how technology regulation can become more expansionist instead of chronically lagging.
Many SafetyTech Providers (e.g. content moderation companies, parental control app developers, fake news detection providers) will rush towards answers that favour their products, and some are already collaborating (e.g. OSTIA). However, there is surprisingly little peer-reviewed research regarding the positive effects and the negative consequences of the many solutions that promise to make our online lives less miserable. As entrepreneurs set out to conquer new markets and fix our society, where the effects of now big techs mistakes, we must ensure that their good intentions do not ultimately result in more harm than good.

We propose to define and disseminate a taxonomy, to assess risks, source causes and how they can be eliminated, mitigated or transferred for every identified use case in the taxonomy. We aim to extrapolate and crystalise principles that need representation in the law based on the risks and solutions identified in the previous steps. (Safety-) Tech must be regulated with some assurance of compliance to prevent disincentivizing innovation (e.g. standards, Codes of conduct - clear guidance on how to comply with any requirement). We suggest that SafetyTech solutions are developed together with those that should be kept safe (e.g. develop parental control apps together with children), and the exploration of general obligations on companies to be responsible for harm they or their products/services cause (e.g. through product liability, tort or other civil liability law or through human rights law).

Online manipulation

This session explored how human decision-making in highly mediated digital environments becomes the target of actors who tend to abuse their power at the expense of society. During the discussion among participants, this allowed for the identification of different challenges that need to be addressed in the coming years.
On the one hand, the existing interconnections between technical and social and political problems were mentioned. Since these are closely interlinked, they also require interdisciplinary approaches. At the same time, the challenge with online manipulation is not to overburden the affected users. While manipulation can be prevented by adding more information, at the same time, this still leaves a much larger amount of data available.
In addition, it was mentioned several times that measuring the impact of online manipulation is a major challenge. Isolating the effects of specific influence strategies and the effects of potential mitigation strategies are very difficult to measure. Also discussed was the fact that addressing these issues would need to be done in multiple languages. However, not only multilingualism, but also the way of communication (low vs. high context) have to be considered in this context. This increases the complexity even more.
Therefore, there is a challenge, or a need, to address the issues of certain types of online manipulation. This is necessary to better identify where the general population is affected, where marginalised groups or specialists are affected, and how to educate them to recognize, understand, and combat online manipulation.
Different approaches exist to address these challenges. It became clear that a general education and information strategy is needed. This must be done to generate social acceptance, or social awareness. However, it is difficult to communicate to the public and the general public the different types of online manipulation they can be affected by. This requires explanations of the entire digital ecosystem, which combines behavioural advertising, data economics, dark patterns, recommender systems, disinformation, etc.. Therefore, general media literacy, digital literacy and general education is needed.

At the same time, technological solutions must also be promoted. Accordingly, an AI strategy has been considered. However, it is important here that language is central (language-centric AI), because "whoever masters language also masters people's minds." Due to the central role of language, terminology and framing, this simultaneously leads to a technological consequence - "He who masters the terminology masters the mind".
One approach to this is to extract general mechanisms from social media platforms. These can also be approached individually, depending on their perceived impact and importance in the context of online manipulation (e.g., the function of sharing a document (without modifying it)). It should be noted here that people often use different platforms for different purposes - almost as if they present different sides of their personality there. This could present different risks of manipulation on different platforms.
Since AI-driven social media is one of the key arenas for shaping public opinion, political controls and regulations might be a necessary measure, as they play a central role in our societies. Companies cannot be solely relied upon to solve the challenges; government actors are not only a source of disinformation, but must also be part of the solution.
In conclusion, however, to address the problem of misinformation, one must be aware of the complexity of human communication, which is influenced by affect, personal interests, and social and cultural contexts. Only when there is sufficient understanding of this can online manipulation be effectively addressed.

Input for the roadmap

Based on the results summarised in the previous section, the Organising Committee identified several topics which could be a valuable input to a European AI research and innovation roadmap. These will be presented to and further discussed with experts from TAILOR, AI4Media, VISION and CLAIRE in order to enrich the respective roadmap activities.

The below topics are the ones that stood out most prominently and will thus provide the 'core' of the input. However, when the roadmaps will be constructed, all inputs from the Theme Development Workshop will be considered.

Sector specific

- Users should be at the core of the development of future XAI tools
- Communication surrounding AI should not only focus on solely technical or factual information but should be contextualised in the current political, ethical, and moral context
- Further measures to reach the public regarding AI and its capabilities and limits need to be developed and tested
- Abusive language detection and automated moderation may be misaligned in target constructs and broad goals
- AI should be able to make moderating processes more transparent for users on social media platforms
- The issue of data sharing at larger scale remains very challenging due to tension between technical, ethical, and regulatory aspects associated with it
- Clarification is needed regarding the data governance rules in order to reduce hesitation to share data on researchers' side

More general topics not limited to the sector

- It is incredibly important to make explanations of the behaviour of AI as flexible as possible to cater for the needs of users
- It should always be a priority to utilise models which minimise bias by design and properly integrate the uncertainty of predictions used
- Users should be at the core of the development of future XAI tools
- Tighter collaboration between computer scientists and legal researchers and lawyers is needed
- Creation of a large European-wide infrastructure enabling access to large datasets would be beneficial
- Set-up of sandbox environments to facilitate specific types of research is recommended

Summary and Conclusion

The high international interest that was expressed in response to the announcement of the Mitigating Bias & Disinformation Theme Development Workshop translated into excellent attendance of the event. Eighty-four participants joined the TDW, ranging from a diverse set of backgrounds. Fifteen (predominantly EU) countries were represented, with thirteen participants indicating that they are affiliated with industry, whilst fifty-six participants indicated that they are affiliated with academia (fifteen participants indicated “other”). The participation of major industry representatives, with companies like HENSOLDT Analytics, IBM, Deutsche Welle etc. is particularly noteworthy and testifies to great interest on the part of industry. The TDW, therefore, caught the attention of some of the most important actors in the field of Mitigating Bias & Disinformation and brought together representatives from key companies, supra-national institutions, and academia. The workshop thus successfully provided a platform for discussions between representatives from academia and industry Discussions that are key in unlocking the full potential of AI in Europe.

The Organising Committee would like to express its deep gratitude to all experts for their valuable input and contributions to this Theme Development Workshop! Their active participation in the workshop and engagement in the breakout session discussions paved the way for the excellent results presented in this report.

List of participants

(in alphabetical order)

Name	Affiliation
Arya, Vijay	IBM, United States
Assenmacher, Dennis	Leibniz Institute for the Social Sciences (GESIS), Germany
Backfried, Georg	HENSOLDT Analytics, Austria
Balzert-Walter, Silke	German Research Center for Artificial Intelligence (DFKI), Germany
Bellomo, Lorenzo	Scuola Normale Superiore - Pisa, Italy
Bontcheva, Kalina	University of Sheffield, United Kingdom
Brinkmann, Wiebke	German Research Center for Artificial Intelligence (DFKI), Germany
Cresci, Stefano	CNR, Italy
Dahi, Zakaria Abdelmoiz	University of Malaga, Spain
Dutkiewicz, Lidia	KU Leuven Center for IT & IP Law (CiTiP), Belgium
Ebert, Nico	Zurich University of Applied Sciences, Switzerland
Elflein, Dennis	German Entrepreneurship, Germany
Fernández Peralta, Antonio	Central European University, Austria
Gallotti, Riccardo	Bruno Kessler Foundation, Italy
Gatica-Perez, Daniel	Idiap-EPFL, Switzerland
Gilotta, Elena	Bizsalt, Germany
Grimme, Christian	European Research Center for Information Systems (ERCIS), Germany
Hitrova, Christina	PwC, Czech Republic
Hrckova, Andrea	Kempelen Institute of Intelligent Technologies, Slovakia
Keilhacker, Andreas	German Entrepreneurship, Germany
Kieslich, Kimon	University of Düsseldorf, Germany
Kompatsiaris, Yiannis	CERTH-ITI, Greece
Kramer, Olaf	University of Tübingen, Germany

Kuczerawy, Aleksandra	KU Leuven Center for IT & IP Law (CiTiP), Belgium
Lensink, Saskia	TNO, The Netherlands
Lu, Meng	Peek Traffic B.V., The Netherlands
Lüling, Ralf	Aleph-Alpha, Germany
Maas, Jonge	Delft University of Technology, The Netherlands
Mair, Stefan	Ringier, Switzerland
Meseberg, Kay	ARTE G.E.I.E., France
Metag, Julia	Westfälische Wilhelms-University Münster, Germany
Mezaris, Vasileios	CERTH-ITI, Greece
Müller, Kilian	European Research Center for Information Systems (ERCIS), Germany
Monderkamp, Hannah	Rheinische Post, Germany
Mondon, Ariane	Sekretariats- und Sprachenservice Mondon, Germany
Morini, Virginia	University of Pisa, Italy
Onchis, Darian	West University of Timisoara, Romania
Paiva, Ana	IST University of Lisbon, Portugal
Papadopoulos, Symeon	CERTH-ITI, Greece
Pedreschi, Dino	University of Pisa, Italy
Petrak, Johann	University of Sheffield, United Kingdom
Pohl, Janina	University of Münster, Germany
Polizzi, Eugenia	CNR, Italy
Popescu, Adrian	CEA, France
Primiero, Giuseppe	University of Milan, Italy
Preuss, Mike	Leiden University, The Netherlands
Rossetti, Giulio	CNR, Italy
Samory, Mattia	GESIS Institute Cologne, Germany
Sarris, Nikos	CERTH-ITI, Greece
Schlicht, Ipek	Deutsche Welle, Germany
Schmidt, Christian	Hensoldt Analytics GmbH, Austria
Schulz, Konstantin	German Research Center for Artificial Intelligence (DFKI), Germany

Slijepcevic, Djordje	St. Pölten University of Applied Sciences, Austria
Smacchia, Marco	Università degli Studi G. d'Annunzio Pescara, Italy
Smith, Lucy	AIHub, Germany
Spangenberg, Jochen	Deutsche Welle, Germany
Srba, Ivan	Kempelen Institute of Intelligent Technologies, Slovakia
Stockinger, Elisabeth	ETH Zürich, Switzerland
Suesser, Daniel	Pennsylvania State University, United States of America
Teyssou, Denis	AFP, France
Thallinger, Georg	Joanneum Research, Austria
Tintarev, Nava	Maastricht University, The Netherlands
Trautmann, Heike	European Research Center for Information Systems (ERCIS), Germany
Treullier, Célina	Université de Lorraine, France
van Aulock, Raphael	Alfred Landecker Foundation, Germany
Verdoliva, Luisa	University Federico II of Naples, Italy
Weber, Marie	Freelance translator, Germany
Zollo, Fabiana	Università Ca' Foscari, Italy

In addition to this list, 7 participants of the TDW preferred not to be mentioned publicly by name and affiliation.

The organisers would like to thank all participants for their valuable input and contributions to the Theme Development Workshop!