Funded by the European Union
Horizon Europe
(HORIZON-CL4-2021-HUMAN-01-27
AI to fight disinformation)

AI4TRUST

AI4TRUST

D4.1 SOCIAL DYNAMICS OF MIS/DISINFORMATION

PARTNERS

Document information

Title	D4.1 - First report on social dynamics of mis/disinformation
Editor	CNRS
Contributors	UCAM & UNITN
Dissemination level	☐ CO: Confidential, only for members of the consortium (including the Commission Services) ☐ RE: Restricted to a group specified by the consortium (including the Commission Services) ☐ PP: Restricted to other programme participants (including the Commission Services) ☒ PU: Public
Reviewers	☒ UCAM
Status	☐ Draft ☒ WP Manager accepted ☐ Coordinator accepted
Due date of deliverable:	30/11/2023
Actual submission date:	29/11/2023
Work Package:	4
Lead partner:	CNRS
Partner(s) contributing and/or reviewing:	UCAM, UNITN & FBK

• Summary of modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.1	22/09/2023	Paola Tubaro (CNRS)	Table of Contents
0.2	06/10/2023	Paola Tubaro, Emmanuel Lazega, Camille Roth and Yasmine Houri (CNRS)	First round of contributions
0.3	31/10/2023	Paola Tubaro, Emmanuel Lazega, Camille Roth and Yasmine Houri (CNRS), Hugo Leal and Stefanie Felsberger (UCAM), and Elena Pavan (UNITN)	Second round of contributions
0.4	10/11/2023	Paola Tubaro, Emmanuel Lazega, Camille Roth and Yasmine Houri (CNRS), Hugo Leal and Stefanie Felsberger (UCAM), and Elena Pavan (UNITN)	Final round of contributions
0.5	24/11/2023	Paola Tubaro, Emmanuel Lazega, Camille Roth, and Yasmine Houri (CNRS), Hugo Leal and Stefanie Felsberger (UCAM), and Elena Pavan (UNITN)	Internal review by CNRS, UCAM & UNITN
0.6	27/11/2023	Gina Neff (UCAM)	Review by designated Quality Assurance Manager, i.e. UCAM
0.7	28/11/203	Riccardo Gallotti, Serena Bressan, and Maria Vittoria Zucca (FBK)	First review by the project coordinator
1.0	29/11/2023	Riccardo Gallotti and Serena Bressan (FBK)	Final review by the project coordinator for report submission

Table of contents

List of abbreviations

ABBREVIATION	MEANING
AI	Artificial Intelligence
API	Application Programming Interface
FIMI	Foreign Information Manipulation and Interference
GDPR	General Data Protection Regulation
INSNA	International Network on Social Network Analysis
IRBs	Institutional Review Boards
LLM	Large Language Model
NFC	Need For Cognition
RECs	Research Ethics Committees
SNA	Social Network Analysis

List of figures

0. Executive summary

Deliverable D4.1 "Social Dynamics of Mis/Disinformation", led by the French National Centre for Scientific Research (CNRS), is a public deliverable of the European project **"AI4TRUST - AI-based-technologies for trustworthy solutions against disinformation"** (hereafter also referred to as "AI4TRUST") and part of **Work package 4 (WP4)** entitled "Human-Centred Explainability, Interpretation and Policy".

This report includes WP4's **first findings on the socio-contextual basis for mis/dis/mal-information.** It is based on a thorough review of the literature on how social actors, under uncertainty and embedded in their social networks, interactively elaborate information and participate in the production and dissemination of content that may be false, deceptive, or in some way misleading (Section 1). It highlights how any analysis of content alone would be too limited to explain the diffusion of mis/dis/mal-information. Instead, through analysis should consider the meanings that people collectively give to content and the interactive processes that lead to the emergence and spread of mis- and disinformation. This report presents social network analysis as part of the essential toolbox to tackle this complex phenomenon.

This report also examines the **tools and platforms already used to fight mis/dis/mal-information,** presenting some use cases and an analysis of the anticipated expectations of future users, especially fact-checkers (Section 2). The report concludes by highlighting **implications for the AI4TRUST project**, suggesting a cross-platform approach to data collection and analysis. D4.1 emphasises certain difficulties posed by adopting a cross-platform approach, especially concerning the diversity of data types and formats, as well as issues related to data access, ethics, and data protection. It proposes solutions to these challenges and for AI4TRUST in Section 3.

A **glossary of the technical terms** used in this report is available in Section 0.2.

0.1 Summary

This deliverable presents **the state of the art in the socio-contextual basis for mis/dis/mal-information,** relying on a broad review of extant scientific literature and on preliminary fieldwork observations, while also outlining **directions for further development within AI4TRUST**.

We begin by defining the scope and varieties of the phenomenon under study. In line with recent research, we see it as encompassing "**misinformation**" (i.e., inaccurate information unwittingly produced or reproduced), "**disinformation**" (i.e., erroneous, fabricated, or misleading information that is intentionally shared and may cause individual or social harm), and "**malinformation**" (i.e., accurate information deliberately misused with malicious or harmful intent). AI4TRUST should **extend beyond the fundamental task of solely identifying incorrect information and strive to encompass intent**. This involves considering the **social processes** that accompany the emergence and dissemination of problematic content.

We then review the literature, which often describes the characteristics of the process of diffusion of mis/dis/malinformation in terms of **"cascades"**, i.e., the iterative propagation of content from one actor to others in a tree-like fashion, sometimes with consideration of temporality and geographical reach. A key finding is that **network structures may facilitate or hinder propagation**, regardless of the characteristics of individuals in these networks. Instead, the actual **offline impact** of online disinformation is disputed. To move forwards, **AI4TRUST** could follow the most recent studies and rely on **hybrid approaches mixing network and content analysis** ("socio-semantic networks").

Mis/dis/malinformation campaigns are sometimes **driven by economic interests**, insofar as the business model of the internet confers value upon content that attract attention, regardless of their veracity or quality. A cross-country shadow market of paid engagements blurs the picture and invites caution when interpreting social media metrics such as users' ratings and reputation scores. Research to be done within AI4TRUST should also be mindful that **highly mediatised disinformation campaigns** are only the tip of the iceberg, **low-stake cases** being far more frequent and difficult to detect.

Spreaders of mis/dis/malinformation may be **bots or human users**, the former being increasingly controlled by social media companies. Not all humans are equally likely to play this role, though, and the literature highlights **"super-spreaders"**, particularly successful at sharing popular albeit implausible content, and clusters of spreaders – both detectable in data with social network analysis techniques.

Adoption of mis/dis/malinformation should not be taken for granted and depends on **cognitive and psychological factors** at individual and group levels, as well as on network structures. Actors use **"appropriateness judgments"** to give meaning to information and elaborate it interactively with their networks. Judgments depend on people's identification

to reference groups, recognition of authorities, and alignment with priority norms. Adoption can thus be hypothesised to increase when judgments are similar and signalled as such in communication networks. AI4TRUST could flag such signals to help users in their contextualisation and interpretation of the phenomena described.

Multiple examples of research in social network analysis can help develop a model of the emergence and development of appropriateness judgements. **Homophily and social influence theories** help conceptualise the role of inter-individual similarities, the dynamics of diffusion in networks sheds light on temporal patterns, and analyses of heterogeneous networks illuminate our understanding of interactions. Overall, **social network analysis combined with content analysis** can help AI4TRUST identify indicators of coordinated malicious behaviour, either structural or dynamic.

The report then uses results from **desk research and fieldwork** to trace the **state of the art in tools currently used** to fight mis/dis/malinformation, and highlights the gaps perceived by stakeholders. **Fact-checkers** rely on some companies like Meta and on dedicated websites which make specific tools available, while other commonly used solutions include reverse image and/or video search tools, software to recognise characters in images, automated translators, internet archives, and others. **Policymakers**, less interested in debunking a single piece of information and keener to understand the more general drivers of mis/dis/malinformation campaigns, rather rely on social listening companies like Graphika and the services they provide. **Researchers and academics** leverage a range of data collection tools, both quantitative and qualitative, commonly used in standard social-scientific data collection.

Use cases involve **detection**, undertaken by a variety of stakeholders — from journalists and fact-checkers to policymakers, academic researchers, and educators — who have different priorities and approaches, although all are under increasing strain due the growing magnitude of the phenomenon. Another use case is **verification/debunking**, which mainly concerns fact-checkers and for which a variety of tools are leveraged, despite technical and linguistic obstacles in some cases. **Analysis** of mis/dis/mal-information patterns is recognised as important by all stakeholders, though not all have the necessary resources to undertake it thoroughly. Finally, **communication of debunks** is perceived as an essential step to educate the public and to support journalists and other professionals.

The stakeholders who could be **potential users of the AI4TRUST toolbox and platform** include: (a) **policymakers**, who currently often outsource these services; (b) **fact-checkers**,

who work in fast-paced environments but are unevenly equipped to face an increasing number of challenges; (c) **journalists**, a more diverse group than fact-checkers; (d) **researchers** in both academia and in civil society and human rights organisations; and (e) **educators and organisations** aiming to build media literacy tools.

The **new toolkit and platform to be built within AI4TRUST** will have to address the gaps perceived by users, with a design that allows users to integrate it into their workflow. Among the **needs expressed by potential users**, the most salient are: multilingual capabilities; improved detection of AI-generated content; better solutions to deal with growing use of visual and image-based representations of mis/dis/mal-information ; enhanced understanding of the dynamics of spread phenomena; more effective communication with the public; greater coordination and data-sharing between stakeholders in different countries. While the wishes expressed are broad, varied, and in some cases unfeasible given the current state of technology, they provide a useful **general framework** from which AI4TRUST may select a subset of priority issues to address.

The last part of the report discusses **implications for AI4TRUST**. Against a literature that has most often limited itself to analysing diffusion of mis/dis/malinformation on a single platform, a desirable step forward is a **cross-platform study**, allowing for a comprehensive and comparative understanding of online flows of information. Such a study is more likely to capture people's simultaneous engagement with several platforms, and less dependent on the ontology of a specific platform, thereby allowing generalisation. However, this raises **technical, legal, and ethical challenges**.

To begin with, online social media come in various shapes and process different types of data, serving different purposes, and catering to diverse user needs. **Data are thus heterogeneous**: textual, multimedia, and metadata related to both content and users. While social network analysis *per se* does not need content and can be performed with metadata, efforts to **combine network and content analysis** are promising and should be extended to multiple data types. A first step can consist in establishing a **common ontology**, although it will not completely eliminate platform diversity and may require caution in interpreting findings. It is also important to **address linguistic heterogeneity**, avoiding the risk of losing meaning, cultural context, and nuances of content, through interdisciplinary collaboration. The AI4TRUST consortium may seek the **help of translation experts and (socio-)linguists**.

Another challenge is the **diversity of legal agreements imposed by social media platforms** on their users and on researchers requesting access to their data. Their specificities vary across platform, space, and time. Within AI4TRUST, a preliminary, detailed comparative analysis of the terms of use of target platforms may be needed to achieve the **methodological coherence** required by cross-platform analysis.

Finally, both cross- and single-platform analyses of social networks require **adaptation of standard ethical provisions** to ensure that research with human subjects does no harm and respects people's freedom and dignity. We transpose and adapt to the needs of AI4TRUST solutions that have emerged within the social network research community, like **anonymisation, pseudonymisation, and de-identification** (as names and identifiers are necessary to trace users across platforms and networks). These adaptations come with the recommendation to use reflexivity and examine the solutions and options most suitable in each case.

Overall, we propose to **address these challenges** and reach a deeper understanding of the social basis of mis/dis/malinformation through a **two-step approach** that starts from content, traces the networks that form around content in various online environments, and links content to the multilevel context of the user community that enables their circulation.

○ 0.2 Glossary of social network terms

This section includes a brief **glossary of social network-related terms** that are employed throughout the deliverable. This glossary is not intended to be an exhaustive resource. Rather, it aims to provide readers with a basic understanding of the network concepts and tools that are relevant for mapping information disorders and how they develop online. Social network-related terminology may vary across disciplines. Whenever possible, these overlaps are pointed out and the definition used in this report is presented.

Betweenness: One possible measure of node centrality. Betweenness measures the extent to which a node falls on the shortest path between two unconnected nodes. It provides an indication of nodes' potential to mediate relationships between others.

Centrality: A property of network nodes. It represents the extent to which one node/actor is involved in relationships with other nodes/actors in the network.

Cluster: Subset/area of a network where nodes are more densely connected amongst themselves than with others in the network.

Clustering coefficient: A measure indicating the extent to which nodes in a network tend to cluster together. The higher this coefficient, the more likely ties in a network concentrate in specific areas.

Degree: One possible measure of node centrality. It measures the number of edges a node has with others in the network. It provides an indication of nodes' prominence within a network.

Edge: The link between any pair of nodes or actors in a network. An edge can be also called a tie or link.

Edge direction: The orientation of an edge between any pair of nodes. The direction of an edge goes from the actor/node sending the ties to the one receiving it.

Graph: A mathematical model to represent a network through a set of nodes and ties between them.

Heterogeneous network: A network composed of multiple categories of nodes (e.g., social actors, bots, images, videos).

Homogeneous network: A network consisting of nodes that are all of the same type (e.g., a network comprising social media users that employ a specific hashtag).

Homophily: A tendency of nodes in a network to form ties with other nodes sharing their same characteristics (e.g., fans of the same movie tend to be more connected amongst themselves than to fans of other movies).

Links: see *Edges*.

Network: A finite set or sets of actors (i.e., nodes) and the relation or relations defined on them.

Network member: see *Node*.

Node attributes: Any properties through which nodes can be classified and distinguished in groups/categories.

Node: Any entity belonging to a network. A node can be also called a network member or, especially in social network analysis, an actor.

Social relation: A specific declination of network edges/ties that indicates a link of social nature between any pair of nodes, e.g., friendship, mentioning on a social media platform, sharing the same interests.

1. State of art of socio-contextual basis for misinformation, disinformation, and malinformation

The **spread of problematic information** is a complex phenomenon. Below we define the term and clarify the scope and variety of forms of problematic information (subsection 1.1). Then, we trace the state of the art, focusing on four dimensions: (a) **the dynamics of (online) disinformation**, distinguishing the characterisation of spread itself as a collective process within a given social system or perimeter, also considering its (real or potential) impact onto the offline world (subsection 1.2); (b) **the economic incentives that favour the production and dissemination of disinformation**, and the market of non-genuine content production that has developed alongside social media and internet platforms (subsection 1.3); (c) **the characterisation of those who spread such content** (subsection 1.4); and (d) **the description of adoption dynamics** both at the individual and the group levels (subsection 1.5). Finally, we present **a typology of social networks** and how this can shed light on the phenomenon (subsection 1.6).

1.1. Misinformation, disinformation and malinformation in context

The spread of incorrect information, from falsehoods targeting individuals to elaborate narratives discriminating against entire human groupings, is not new. In fact, the

misinformation-disinformation spectrum (see fig.1) covers some **well-known historical practices**, from rumour dissemination to propaganda operations. The circulation of information distortions was traditionally limited by the reduced number of opportunities and outlets for their dissemination. For example, propaganda has been a quasi-monopoly of States, their agencies and respective communication vehicles. Even in plural democracies, propaganda was perceived as a type of information management by which the few (either elected or non-elected) could steer the many through persuasion rather than force. In the words of the political scientist Harold Lasswell "if the mass will be free of chains of iron, it must accept the chains of silver" (Lasswell, 1971, p. 222).

The dramatic **transformations to the technologies of information and communication** that took place in the 1990s democratised both the access to information as well as its production and reproduction. As a result, the information ecosystem itself underwent a radical transformation (Naughton, 2014). The number of opportunities and outlets to disseminate and/or distort information grew exponentially, particularly after the consolidation of the Web 2.0 (e.g., blogs and social media platforms). The scope, scale, and speed of narratives, true and false, were unprecedented. This "networked age" empowered new agents and created new spaces for both the circulation and distortion of online information. It is also in this period that we encounter a renewed interest in the practical distinctions between misinformation and disinformation. Previously addressed in domains such as information theory and philosophy of information (Floridi, 2003; Fox, 1983), the distinction was repurposed for information circulating on the internet where **"inaccurate information might result from either a deliberate attempt to deceive or mislead (disinformation), or an honest mistake (misinformation)"** (Hernon, 1995, p. 134). In the formulations adapted to the (inter)connected nature of the new information ecosystem, the focus is not only on the authenticity of the message but also on the authenticity of the messengers and their interconnections. Intent appears as the defining feature and differentiating factor between online misinformation and disinformation.
In the aftermath of some of the most consequential world-wide disinformation campaigns, from India to the US, Brazil, UK and Eastern Europe, Claire Wardle identified a new phenomenon referred to as **"information disorder"** (Wardle & Derakhshan, 2017, p. 4). She (re)defines some of the key concepts underpinning the information disorder, namely:

- **Mis-information:** "Information that is false, but not created with the intention of causing harm" (Wardle & Derakhshan, 2017, p. 20);

- **Dis-information:** "Information that is false and deliberately created to harm a person, social group, organisation or country" (*Ibidem*);
- **Mal-information:** "Information that is based on reality, used to inflict harm on a person, organisation or country" (*Ibidem*).

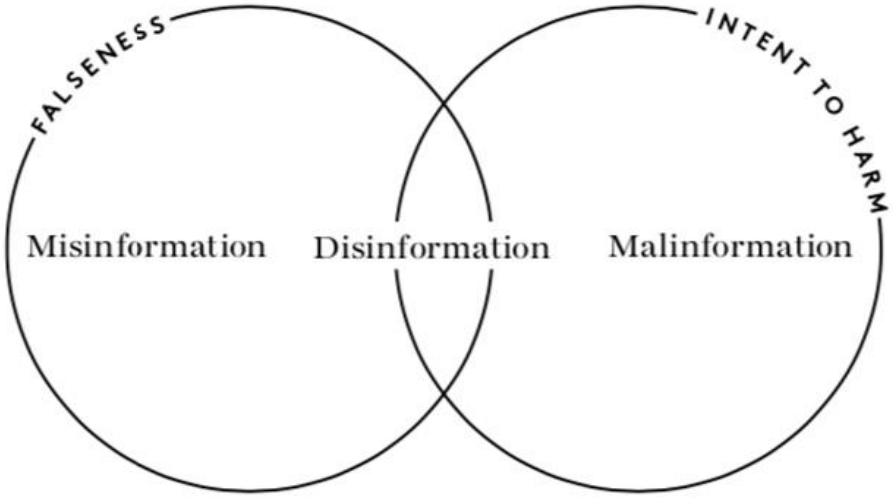

Figure 1: Information distortions – Source: (Wardle, 2020)

Duly adapted, this conceptualisation can be adopted as the AI4TRUST frame of reference. We suggest adaptations at three levels: **morphology, motivations of the actors, and scale of the actions.**

1. Morphologically, it is important to note that these terms have permeated the information landscape as prefixed – rather than just hyphenated – nouns. The suggested formulations should, nevertheless, absorb the grammatical evolution *and* the social uses across the working languages of the project. In this sense, while disinformation (and its linguistic variants) has been widely adopted, the same does not happen with the terms misinformation and malinformation. Both terms are well established in English but in most other languages we observe 1) **misuses of the word disinformation (and "fake news")** as an all-encompassing description of every type of information distortion; and/or 2) **adoption of borrowed words from English** (e.g., fake news, misinformation); and/or 3) **utilisation of an open compound word** (closed compound in German and French) that describes the phenomenon accurately (e.g., *cattiva informazione*, *Fehlinformationen*). It should be noted that we deliberately choose not to use the expression 'fake news'. Due to its scientific and

semantic inadequacy, the term dis/mis/malinformation encompasses not only false, but also genuine information that is used out of context and weaponised against some people or groups; likewise, it is not limited to 'news' and includes a variety of online content.

2. At the level of the motivations of the actors, here understood as those who produce and/or reproduce the mis/dis/malinformation, it should be stressed that intent goes beyond harmful intentions. This semantic clarification is important insofar as, except for malinformation whose main driver is indeed harmful intent, **the presence or absence of an intention to cause harm is not the only defining characteristic of the other information distortions**. Intent can also be an attempt to influence, persuade, dissuade, etc. It is the presence or absence of intent *latu sensu* rather than just harmful intentions that constitutes the boundary separating misinformation from disinformation.

3. Regarding the scale of the actions, we need to address two dimensions. At the micro level, we find **discrete informational content** (e.g., individual texts, videos, images) and, at the meso and macro levels, we must take stock of **networked informational contexts**, in particular, the social dynamics of circulation by actors and clusters of actors across different social media platforms and cultures. While Wardle's formulation focuses on the production of messages at the micro level, their reproduction and dissemination at the meso and macro levels is much more relevant from the perspective of a project like AI4TRUST, which aims to detect and counter mis/dis/mal-information by pre-emptively warning stakeholders about risks associated with the spread of all types of information distortions.

The following **conceptual framework** builds on the mentioned scholarly contributions, best fact-checking practices, institutional definitions laid out in official EU documents (e.g., European Democracy Action Plan[1]) and our own adaptations:

- **Misinformation**: Incorrect information produced or reproduced without neither knowledge about its accuracy nor harmful intent;

[1] The European Democracy Action Plan was produced by the European Commission in 2020. The document is part of the EU wider strategy to promote free and fair elections, strengthening media freedom and counter disinformation. The EDAP was an important steppingstone to the elaboration of the "Strengthened Code of Practice on Disinformation" in 2022 and it contains what we consider to be a very fitting definitions of misinformation, "false or misleading content shared without harmful intent though the effects can be still harmful", and disinformation, "false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm" (Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Democracy Action Plan, 2020, p. 18).

- **Disinformation:** Incorrect, fabricated, or misleading information that is intentionally shared by actors with the aim of disseminating it through social networks. The goals could go from deception to political or economic gains and may result in both individual and social harm.
- **Malinformation**: Typically, factual information deliberately used with malicious and harmful intent (e.g., doxing, revenge porn).

Although the socio-psychological elements of misinformation sharing are relevant, **AI4TRUST aims at identifying the dynamics of diffusion and the extent to which we can detect disinformation operations at scale.** The task of disinformation research is to unveil the central actors, processes, and overarching goals of *intentional* dis/mis/malinformation campaigns rather than just locating and debunking incidental misinformation stories. For that purpose, AI4TRUST focuses not only on discrete units of produced content but also on contexts of reproduction. Not texts but their broader contexts, not isolated posts but semantic networks of analogous false narratives, not individual users but interconnected actors with different positions and roles within identifiable clusters, not one specific communication channel but the connections across different platforms. The units of analysis to identify disinformation are plural and networked. Against this background, **it is important to look at the diffusion of mis/dis/malinformation with a methodological toolkit that addresses the dynamics of complex contagion** (Centola, 2020) within equally complex networked structures.

1.2. Diffusion of mis/dis/malinformation

A whole strand of (rather quantitative) research focuses on the **spread of mis/dis/malinformation at the level of an entire social media platform**. It principally aims at characterising the structural, temporal and, sometimes, geographical, or categorical properties of the spread.

Structural appraisals have typically built upon the widespread literature on so-called **cascades**, i.e., the study of the iterative propagation of content from an actor to further actors, in a tree-like fashion. Several early studies emphasised the role of the immediate network structure surrounding actors to explain diffusion phenomena, introducing a **meso-level perspective** that diverges from the idea that the features of individual nodes alone may be responsible for a successful propagation (Watts and Dodds, 2007; Weng et al., 2013). This quickly led to suggest **applications of social network analysis**, and even of

social and semantic network analysis, to the automatic detection of mis/dis/mal-information which is also very much at the root of AI4TRUST. Various endeavours accordingly aimed at characterising the arrangement of nodes and links in the vicinity of specific mis/dis/mal-information publication events (Ratkiewicz et al., 2011) and at searching for a relationship between the topology of the contextual network at a distance of a few hops (generally one or two) from a publication or a source and its trustworthiness. **A less complex structure of such networks has quite recurrently been observed**, be it between reposting actors or between co-cited information sources, whereby mis/dis/mal-information appears to be related to more compact, shallower, and perhaps more regular topologies (Jin et al., 2013; Singh et al., 2020). While a sizable part of the early literature focuses on either structural or semantic features exclusively, many recent structural approaches appear to rely on **hybrid approaches mixing networks and content analysis** – a combination that is likely critical to grasp the socio-cognitive dynamics of mis/dis/mal-information propagation (Conti et al., 2017) (see also section 1.5).

A third dimension relates to the **temporality of diffusion**, which is often framed in comparative terms by contrasting the diffusion of trustworthy vs. non-trustworthy content, as well as measuring the (generally weak) effect of debunking actions following mis/dis/mal-information publication. Starbird et al. (2014), for one and among others, shows that mis/dis/mal publication typically exhibits a peak followed by power-law temporal decay, while corrective content usually lags slightly behind and has a much smaller audience, one or two orders of magnitude below, without preventing the further spread of the initial publications at the system level. Several studies have made use of the rather strict definition dictated by the dichotomy of so-called **"hoaxes vs. non-hoaxes"** stemming from story labelling website Snopes, in order to gain insights on the temporal properties of mis/dis/malinformation spread. They have notably established that stories confirmed to be "hoaxes" spread faster, be it on Facebook (Friggeri et al., 2014) or on Twitter (Vosoughi et al., 2018). This research strand also quantitatively confirms the difficulty of correcting rumours once they have spread (Lewandowsky et al., 2012).

Mis/dis/malinformation spread has also been **contrasted with geographical and categorical properties**. For instance, Cinelli et al. (2020) show that mis/dis/malinformation does not cross-national borders much in the context of the 2019 European elections. On a more global level, Gallotti et al. (2020), who measure likely misinformation via cumulative follower counts of users sharing messages pointing to web domains externally assessed as reliable or not, also show that likely exposure to such content related to Covid-19 varies

greatly according to countries. They further show that it is negatively correlated with the number of infection cases. Ceron et al. (2021) confirm that there is significant variation across countries, also because of distinct national political settings, in the propagation dynamics of misinformation by examining the spatiotemporal dynamics of hashtags related to Covid-19 in Latin America in 2020.

This last point hints at the **possible links between a given platform and its offline environment**. Of importance is the quite comprehensive study by Allcott and Gentzkow (2017) which shows that much of the content assessed as fake by fact-checkers and circulating during the 2016 US presidential election was favouring Trump, that the respondents of their own survey saw and remembered genuinely about 1.14 percent of such content. Compounding that with the observation (from the literature of the time) that people tend to believe quite often disinformation, that social media account for most of the traffic to such content, they suggest that disinformation may indeed have tilted the election in favour of Trump — nonetheless principally more a suggestion than a demonstration. At the time, such studies supported immediate calls for action. For one, the multi-author manifesto of Lazer et al. (2018) formulates two recommendations, building upon this notion that there is a good amount of bots and disinformation circulating without knowing really their actual impact: **empowering users** (featuring psychological experiments on disinformation reception, as well as critical education), and **empowering platforms** (using algorithms to counter disinformation, which is right in the line of several AI4TRUST objectives). More recent literature tends to admit that the extent to which exposure to online disinformation translates into actual behavioural change (for example, voting for a given candidate) is poorly understood, and may actually be much less strong than commonly believed. An established finding of public-policy research is that, first, if people largely consume disinformation they already agree with, or are predisposed to accept, then it is unlikely to radically change their beliefs or attitudes; and second, even if it does, attitudes are only weak predictors of behaviours (Altay et al. 2023). **Behaviour change** after contact with disinformation depends on people's own system of values and beliefs, and on the social environments that surround them, which validate or invalidate their reactions. Integrating a **social networks approach** into the AI4TRUST solutions that are being built is a way to take this form of complexity fully into account, in line with recent research.

1.3. The political economy of mis/dis/mal-information

The **business model of social media** like Facebook, Instagram and TikTok, and of most digital economy services, from search engines to online news websites, rests on advertisements. The public can typically access the service for free (or against a generally small subscription) while advertisers pay fees depending on reach and perceived likelihood of success. Companies thus compete on attention: they seek to attract "eyeballs" to ensure that ads get maximum visibility, and to charge correspondingly high fees to advertisers. Further, they use detailed data on users' online practices and behaviours to target content and ads specifically to those people who are most likely to read, like, and share them. To the extent that **elements of mis/dis/malinformation constitute potentially attractive content and may catch users' attention**, they thus acquire **economic value**, and may be shared to sell pricey advertisement spaces. Automated micro-targeting may expose users to false or misleading information, just as it exposes them to publicity of harmless products. In this sense, the very economic model that sustains most of the internet is also what may facilitate the spread of unverified or problematic content.

In recent years, a (more qualitatively-inclined) strand of research has unveiled how these **economic interests** contate social media with **problematic content**. Beyond the spontaneous, "organic" behaviour of users, misinformation and disinformation spread through actors that attempt to take advantage from the economy of attention and the value that accrues to attractive content. In particular, the so-called **"click-farms"** are online platforms or small (and sometimes informal) companies that pay workers to click, follow, and like accounts on social media like Instagram, TikTok, and YouTube. The clients of these services are influencers, brands, celebrities, and even candidates to political elections that seek to increase the popularity and visibility of their social media accounts —or to attract attention toward them. Hence, they buy followers, likes, shares, and sometimes written reviews of their profile, product, electoral promise, or whatever they offer online. These practices can be seen as examples of malinformation: the "like" action, for example, is genuinely taken by a user in reaction to some real online content, such as a post on Facebook or a video on YouTube, but it is misused to artificially boost the popularity of that content. In part, recourse to click farms is a reaction to social media companies having become tougher on the use of automated tools (bots), insofar as fraud remains more difficult to detect when it comes from real humans. On their side, the workers who provide these services typically receive less than a penny per task (Grohmann et al., 2022a).

Click-farming can be potentially very harmful if the purchase of likes and shares concerns medical (dis or mis)information, for example on vaccines or drugs, or sensitive issues that are likely to be taken into account in electoral choices and/or policy-making. An extreme example is the highly mediatized case of the disinformation industry that developed in Veles, a small town in central Macedonia, where many pro-Trump posts originated during the 2016 US presidential campaign. While its exact effects on election outcomes are difficult to assess, it is a clear example of **how the driving force was an economic, not a political motivation** (Hughes & Waismel-Manor, 2020). But the phenomenon is older: as early as 2012, it was noted that hackers leveraged the spam-as-a-service market to acquire thousands of fraudulent accounts which they used in conjunction with compromised hosts around the globe to flood out political messages in relation with Russian parliamentary elections (Thomas et al., 2012).

When these practices boost relatively innocuous content, such as a song or an entertainment video, consequences may seem milder; at large scale, however, they disrupt platforms' reputation, scoring, and ranking algorithms. Thus Lindquist (2021) speaks of **click farms as "impostor infrastructure" and "illicit digital economies"**. Grohmann et al. (2022b) see click-farms as **"parasite" platforms** because they depend on social media platforms infrastructures to survive, while at the same time threatening them. The very existence of click-farms is embarrassing for social media companies, because the former leverage, and simultaneously discredit, the same business model as the latter. In response, social media companies hire **armies of commercial content moderators** (Roberts 2019), whether internally or (more commonly) through subcontractors, to keep click farms (and their clients) under control. They tend to **ban suspicious behaviours** (for example, users who "like" or even review a product not sold in their country) **or accounts** (for example, accounts that only share third-party content without producing any on their own). In this way, click-farms push workers to scam social media platforms, all the more so as their low pay stimulates an underground, legally-borderline market of fake accounts and bots (Grohmann et al., 2022b).

Click-farms have been mainly observed and studied in Southeast Asia (Ong & Cabañes, 2019; Lindquist, 2021, 2022) and Latin America (Grohmann et al., 2022a, 2022b). Like other online labour platforms, and like content moderation contractors (many of them based in the Philippines according to Roberts 2019), they connect **low-tech workers from the periphery to core infrastructures of the digital economy** in North America and Europe.

So far, very few studies have engaged with the point of view of workers in click-farm platforms (Ong & Cabañes, 2019; Grohmann et al., 2022a). Even less explored are the structures and characteristics of the social networks of these workers, which may facilitate automated detection of relational structures in which mis- and disinformation are more likely to originate or propagate. The **proliferation of accounts for the same person on the same platform**, a commonly used strategy to escape platforms' sanctions, further complicates the task.

An implication of this, of which future actions within AI4TRUST should be mindful, is that **users' ratings and any reputation or ranking scores** calculated by the platform on this basis, **cannot be taken at face value as indicators of users' appreciation**. While experimental studies could set apart users' evaluations of some content or piece of information, for example to analyse the extent to which early ratings influence subsequent ones (Frey & van de Rijt, 2021), or to distinguish the competing effects of quality and popularity (van de Rijt, 2019), observational data from platforms that include potentially artificially-inflated scores may make such analyses impossible.

1.4. Diffusers of mis/dis/mal-information

Which nodes, then, are at the root of mis/dis/mal-information dissemination? One strand of research has explored **bots**, that is, computer algorithms that execute specific tasks in online social network sites, usually simulating the behaviour of, and interacting with, human members. Shao et al. (2018) show that most active users are more likely to be bots, and bots tend to target users with more followers. They are unequally distributed in the US (where the study is based — possibly because of the self-declaration of their location which may betray an effort at influencing specific geographical area), and most importantly, **humans appear to do most of the retweeting**. While bots can contribute to the diffusion of fabricated or manipulated content, social media companies have become increasingly efficient at recognising (and banning) them.

According to some studies, **human users** play an even more important role than bots in diffusing disinformation owing to inability to recognise unreliable information and to emotions, especially though not exclusively when content concerns political topics (Vosoughi et al., 2018). However, more recent research suggests that unreliable news represents a small portion of people's exposure to information, that most users do not share it, and that on average, people deem false content less plausible than verified ones

(Acerbi et al., 2022). Additionally, as discussed above (section 1.2), exposure to disinformation does not necessarily mean adoption of it.

It is therefore useful to distinguish, and focus on, specific types of human spreaders, rather than broadly assuming that any human user could become one. We have briefly presented above the universe of those who derive an economic advantage from the spread of dis/mis/malinformation, and the click-farmers they often employ to achieve their goals. Likewise, others may derive a political advantage from the spread of content aligned with their intentions, ideologies, or beliefs, and may equally rely on click-farming to increase their visibility and impact.

The literature has also identified **specific profiles** that, regardless of intent, adopt behaviours and practices that actively promote the diffusion of mis/dis/malinformation. For instance, the cross-platform study (Facebook, Twitter, YouTube) of Yang et al. (2021) reveals the existence of a minority of specifically low credibility-content sharing source accounts, or so-called **"superspreaders"**: in particular, while posts originate from a large number of accounts, successful posts i.e., widely reposted content is produced by a much more limited set of users (who furthermore do not often have a so-called "verified" status on the platform). This work also exhibits various clusters of coordinated sharers in terms of the similarity of the domains being shared, which resonates with the observation (section 1.2) that networks surrounding mis/dis/mal-information are rather compact and shallow.

1.5. Socio-contextual adoption dynamics

What factors drive the adoption of spreading disinformation by social media users? Broadly, there are **signs of a partisan bias in tolerating lies** (De keersmaecker and Roets, 2019) **or in being more or less susceptible to certain types of disinformation**. For instance, there seems to be a conservative (Havey, 2020) or respectively liberal (Borah, 2022) bias in adopting, or respectively not adopting, Covid-19 misinformation. The latter study also hints at the influence of cognitive variables too, such as what is often termed in the psychology literature as the **"need for cognition" (NFC)**. NFC may be summarised as a self-declared preference for complex over simple problems and tasks that involve deep thinking. This resonates with a further literature that examines the role of psychological variables in the adoption of mis/dis/mal-information. There seems to be much less partisan bias when so-called analytical thinking is used to discern malinformation from news (Pennycook and Rand, 2019). This suggests by contrast that when someone believes

mis/dis/mal-information, they are not engaging in analytical thinking, but could lack motivation, as Scheufele and Krause (2019) point out in their encompassing account of why disinformation is actually adopted. In their review, they differentiate **the individual, group, and "societal" levels of the phenomenon**. They build, in particular, on the work of Lewandowsky et al. (2012) whereby information most likely to be adopted appears to be cognitively efficient — it is logical, stems from an authoritative source, is aligned with prior beliefs, and enjoys social credibility too (it matters that others believe in it), while cognitive dissonance is avoided, possibly by using more motivated reasoning (i.e., interpreting information in agreeable ways) than selective exposure (i.e., looking for agreeable information).

As for the group level, groupthink is used in the research literature. Decreased consensus (diversity of beliefs and possibly, rumours) is generally linked to fragmented, clustered networks (limiting the possibility of interactions). Compact and shallow topologies are positively associated with mis/dis/mal-information. There is a vast modelling literature connecting rather large networks and increasingly sophisticated topologies with certain information dynamic — in terms of the emergence of various extremes, or of consensus; be it with continuous opinions (as in models of so-called "opinion dynamics", see Castellano et al., 2009) or categorical, often binary opinions (as in models of so-called "cultural dynamics", see Flache et al., 2017; Smaldino et al., 2017). But the experimental literature has rather focused on relatively small numbers of individuals and simple network shapes, aiming at characterising **the emergence of local echo chambers** and its underpinnings, including groupthink and the so-called "majority illusion" (Lerman et al., 2016). Building upon the idea that "people automatically infer how widespread the claim is from subjective familiarity", DiFonzo et al. (2013) validate, through a multi-network multi-participant experiment, the emergence and persistence of local rumour consensus in clustered networks, altogether with belief polarisation, "even when the rumours are unrelated to group identity."

This has direct implications in **the low-scale study of misinformation propagation dynamics**, as adequacy judgments have been shown to be performed under the influence of social connections, beyond the idea that crowds are "wiser" (i.e., the social aggregation of individual, almost independent judgments, leads to a better approximation of a ground truth). Stein et al. (2023) raise this question in a context where group interactions are not random and, more specifically, are rather homophilous. They construct a large-scale experiment where they control the composition of political leanings of users within given

groups, from partisan to cross-partisan networks, and (perhaps unsurprisingly) demonstrate that mixed groups are least susceptible to misinformation adoption; put differently, the majority illusion may not work. These analyses are important because they show that, in addition to psychological and cognitive factors, social structures as can be captured by network analyses play a significant role in accelerating (or conversely, slowing down) adoption. All other things equal (in particular, regardless of individual political orientations), **network segregation disproportionately aids the diffusion of messages** that are otherwise too implausible to circulate.

The above social and social-psychological statements are made at the highest level of generality. AI4TRUST should consider that **actors, especially policy- and decision-makers**, further **contextualise users' (and their own) adoption behaviour**, with consequences for adoption dynamics. Such actors are members of collectives, communities, organisations and institutions with goals, rules and strategies that impose constraints on live information management, in particular sharing and validation. Especially in times of crises (Lobera et al., 2023), these constraints can be challenged or reasserted based on individual and collective judgments with consequences for adoption dynamics. Behavioural assumptions about adoption must therefore include "appropriateness judgments" with which actors as members contextualise information, elaborate it interactively or collectively within their networks (personal and beyond). Understanding how actors make appropriateness judgments, and how these judgments coevolve with networks, requires specifying at least **three analytical dimensions** underlying these judgments when they are politicised: **identification to priority reference groups, recognition of authorities, and choices of beliefs or priority norms** (Lazega, 2014). Crises intensify the production of misinformation, as well as the search for the right appropriateness judgments among network members, here receivers in organisational contexts (Lazega & Quintane, forthcoming). Information diffusion is not mechanical. Adoption and diffusion of misinformation at meso and macro levels can thus be hypothesised to increase when similar politicised judgments are made and signalled as such by senders and receivers in communication networks. AI4TRUST could **track and flag such signals and indicators of collective mindsets** to help users in their contextualization and interpretation of the phenomena described.

1.6. Typologizing social network analyses of mis/dis/mal

The diffusion of mis/dis/malinformation on online social media presents unique challenges that social network analysis can significantly address. The social media landscape is **large-scale**, **multimodal**, and mostly fed by a **diverse** pool of individual and organisational users coming from different backgrounds and whose contextualization, preferences and intentions diverge and coevolve with the networks. As anyone can create an anonymous account without any cost, **malicious bots** can profuse and become powerful tools of misinformation. Most importantly, as defined above, disinformation and malinformation are characterised by their authors' deliberate **intent to deceive**. Yet, it is not trivial to detect this intent on the sole basis of content. Rather, researchers advocate for a hybrid approach that combines content traces such as linguistic cues and network-based contextual data (Conroy et al., 2015).

Social network analysis is the privileged perspective to address these challenges, as it can be articulated around **three dimensions**: content, social, and temporal (Shu et al., 2019). The **content dimension** is related to the news pieces, comments, social media posts and such, that are published online and which form the basic misinformation unit that is usually analysed by traditional, computational, or even linguistic approaches. The **social dimension** describes the relations between publishers, consumers, and spreaders of misinformation. The **temporal dimension** illustrates the dynamics of misinformative behaviours over time. All three of these dimensions can be represented in a tangible way in diverse types of networks (Borgatti 2009) which can, in turn, be used to detect and mitigate the effects of mis/dis/malinformation.

The most straightforward way of modelling the network of online diffusion of mis/dis/malinformation is with a "**friendship" network**. This is a homogeneous directed network where nodes represent social media users and edges represent whether a social relation exists between two of them (Shu et al., 2019). According to **homophily theory**, for example, people tend to form social links with like-minded partners with whom they share similar preferences and background (McPherson et al., 2001). **Social influence theory** further argues that individuals are more likely to share similar latent interest in new content with these similar and homophilous partners (Marsden and Friedkin, 1993). The analysis of the friendship network structure is the basic route to understand online news spreading. It involves examining the topology of the friendship network using network metrics such as degree centrality, betweenness and clustering coefficients to identify key actors and communities that play a role in the dissemination of mis/dis/malinformation.

This structural analysis of friendship networks is better **complemented by a temporal analysis of information flows**. This can be modelled in a **diffusion network** – a homogeneous, directed network where nodes represent entities which can publish, consume or spread news content, and edges represent the type and direction of propagation, which is usually associated with a propagation probability (Shu et al., 2019). Using this approach, researchers may examine how the spread of misinformation evolves over time, by observing features such as the speed of dissemination, peak periods, and changes in the network structure and content.

These basic network approaches to online news dissemination both rely on the latent phenomena of actors' **contextualisation of assertions, appropriateness judgments and subsequent credibility assessments** that are key to the spreading of mis/dis/malinformation. Internet and digitisation technologies have significantly lowered the cost and increased the access of information production and dissemination, which used to be limited to a number of sources endowed with enough authority and capital to justify and sell an information product (Metzger et al., 2007). This raises the issue of credibility assessment for users wanting to **differentiate between reliable and unreliable sources** in a context where information is presented on the same level of accessibility and not systematically subject to filtering through professional gatekeepers (Metzger et al., 2007). A **credibility network** can be represented as an undirected graph where nodes represent social media posts with corresponding credibility scores and edges represent the link type between these posts, such as "supporting" or "opposing". This credibility network can then be used to **evaluate the overall perceived truthfulness of news** by quantifying the credibility scores of each social media post (Shu et al., 2019).

During the news dissemination process, diverse types of entities are involved. Hence, to represent online relations from a multimodal and multilevel perspective, researchers have also used **heterogeneous networks**. Using this approach where diverse types of nodes and edges are represented in the same network, researchers can model **stance networks**. In these networks, nodes can be users (with data on as many attributes and (pre)dispositions as possible), news items and social media posts, and edges can be the link between them, such as "posting" between users and posts, or "stance" between two posts. This forms a first step towards **user behaviour analysis**, as it allows researchers to analyse how users interact with and respond to misinformation given assumptions about individual or collective judgments. In this network, mis/dis/malinformation will typically

generate controversial views among users in their social media posts about news items (Shu et al., 2019).

Still using a heterogeneous perspective, **interaction networks** go further into user behaviour description and analysis. These are networks where nodes represent publishers, groups of publishers, users, groups of users, news, and edges represent the interactions between them: a publisher publishes news, a user spreads it. This approach sheds light on the correlations of publisher bias, news stance, and relevant user engagements simultaneously (Shu et al., 2017).

Lastly, **knowledge networks** stem from a semantic approach to content, embedded in a social context. In a knowledge network, nodes represent knowledge entities (such as Wikipedia pages, dB data or Google Relation Extraction Corpus) and edges represent the relation between them. This is an inherently linguistic approach where fact checking can be approximated by finding the shortest path between knowledge nodes under properly defined semantic proximity metrics (Ciampaglia et al., 2015).

Such interactions and knowledge coevolve in complex ways. As mentioned in **Deliverable 2.1 of AI4TRUST (WP2)**, section 1.3.4, the primary purpose here is not to predict without a theory, but to **use social network analysis to identify indicators of coordinated malicious behaviour**, which can be **both structural** (e.g., identifying the sources disseminating the content) **and dynamic** (e.g., understanding how the content is being diffused across the social network). Doing so can then be enriched with flags providing embedded policy- and decision-makers as well as media professionals using AI4TRUST with indications about actors (individual or group) involved in the diffusion processes, their judgments, their strategies, and their goals.

2. State of the art in tools to counter mis/dis/malinformation

Tools used to counter mis/dis/malinformation not only consist of technological tools to **detect and verify information** but also include organisational infrastructure to **organise workflows and create archives**, tools to **analyse dynamics** of mis/dis/mal-information crucial to inform policy or organisational strategies, or tools to **facilitate the communication of material to counter mis/dis/mal-information,** such as educational

materials or articles debunking mis- or disinformation. This section presents the tools, platforms and services that are currently available, by type and use (subsection 2.1), then discusses how these tools are employed to counter mis/dis/malinformation and presents the gaps identified by, and the needs and wishes of, stakeholders (subsection 2.2).

2.1. Existing tools and platforms and state of the market

The following sections provide an overview of the **technical tools** and platforms that are currently **available to fact-checkers, journalists, policy makers, and researchers**.

2.1.1. Tools used for fact-checking & verification

Chat bots or public channels on popular messaging platforms such as WhatsApp or Telegram are used by fact-checking organisations to detect mis/dis/malinformation. Often members of the public share potential claims for verification with either specific organisations or on public channels, which are then monitored by fact-checking organisations.

Many fact checking organisations also work with and through **Meta's Third-Party Fact Checking programme.** Meta provides fact-checkers with a "queue" of content (either text posts, images, videos, or links) that Meta identifies as potentially inaccurate. How Meta determines the metrics that flag potential content for fact-checking is not known, but it is likely a combination of Meta's own algorithm and users flagging content. Fact-checkers then verify the post, and if it is false attach the debunking or fact checking article to the post on Facebook. The tool built by Meta thus detects potential misinformation for fact-checking organisations and also distributes the counter information. Nevertheless, due to the tool's proprietary status and control by Meta, it is not a tool that serves fact-checking organisations themselves.

There are a range of tools used to verify information. This ranges from journalistic methods of calling representatives for verification to technical tools. To provide context for images, **reverse image search tools** are used such as Google, Bing or TinEye, which has a browser extension. Websites like Foto Forensics[2] are used to determine whether a **picture has been**

[2] http://fotoforensics.com/.

edited or whether two images have been merged into one. Tools to **recognise characters in images**, such as NewOcr,[3] are used to automatically transcribe and recognise text in image-based content. This tool is used to save time. A set of **reverse video search tools** are also available such as InVid[4] or Amnesty International's tool to extract metadata from videos.[5] **Translation** of text into the language of fact-checkers or journalists is also a key activity. TGoogle Translate (as well as the Chrome automatic translation of websites) and DeepL are used. Tools like Namechk[6] are used to check whether a **username** found on one platform is also used on other platforms, which can help to locate people across networks. Similarly, services like Domain Big Data[7] help discover the **person or company behind a domain**. Tools used to **archive** false information found online (social media posts, articles, videos, etc.) are the Wayback machine or archive.is. Archiving is key to document the existence of false information as well as to provide evidence. Other tools serve to facilitate **collaboration** between fact-checking organisations, for example in case of a global news event. Google Spreadsheet is used to facilitate the work of fact-checking or Google's Fact Check Explorer.[8]

2.1.2. Tools used by policy makers

Policy makers do not often engage in active verification or debunking of mis/dis/malinformation, but rather rely on trusted news or fact-checking organisations. Policy makers are more interested in **understanding the dynamics of mis/dis/malinformation**, in order to understand underlying strategies, stem its flow, prevent harm, etc. To this end, policy makers—to different degrees—employ corporate **social listening platforms** such as Graphika, Logically, and Linkfluence (now part of Meltwater). Social listening describes sophisticated tools, including social network analyses, to monitor social media platforms and the web in general to analyse narratives and discourses on either a specific topic or keywords. It is often used by brands to understand the effectiveness of their campaigns, but also by policy makers and governments. This can include **public sentiment analysis** (gauging public opinion on a

[3] https://www.newocr.com/.

[4] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/.

[5] https://citizenevidence.amnestyusa.org/.

[6] https://namechk.com/.

[7] https://domainbigdata.com/.

[8] https://toolbox.google.com/factcheck/explorer.

specific topic or party), **issue identification** (identifying trends or concerns in the public sphere), **policy impact assessment** (how policy decisions are perceived in the public sphere), or **crisis monitoring and management** (tracking of narratives around crises or emergencies). Products and tools offered by social listening companies are not available for fact-checkers and journalists to the same extent as governments due to a lack of affordability.

2.1.3. Tools used by researchers & academics

Researchers and academics employ a range of **different tools and strategies to study the social dynamics of mis/dis/malinformation**. To provide a brief overview, these approaches can be divided in qualitative, quantitative, or mixed-methods. **Qualitative methods** include ethnographies, interviews, expert interviews, focus groups, participant action research that can be conducted online or offline. These methodologies begin with understanding people as the entry point into understanding the social dynamics of mis/dis/malinformation. Such research may ask for examples of why and how people engage with information online or offline, individually, interactively, or collectively, rather than map the phenomena. Online ethnographies might consist of the observation of specific online spaces and their dynamics, such as specific Facebook groups or forums. **Quantitative methods**, which have already been discussed at length in this report, can include social network analysis, content analysis, temporal analysis, sentiment analysis, or surveys.

2.2. Existing use cases and users

This section focuses on outlining a range of **different use cases for tools to counter mis/dis/malinformation** and on describing a set of existing users for such tools. It draws on a series of focus groups with experts working to counter mis/dis/mal-information, including civil servants, educators, AI developers, journalists, and policy makers and fieldwork conducted at a leading fact-checking organisation in Europe. Fieldwork consisted of a researcher spending a week at a fact-checking organisation to conduct participant observation of the organisation's way of working with information and against mis/dis/malinformation.[9] **Eight interviews** were conducted with people working in different

[9] The text addresses the range of concepts defined by Wardle, but the organisation did not distinguish between mis-, dis-, and malinformation internally. While all members understood these differences, the language in which they worked did not have translations for mis- or malinformation.

areas of the organisation about a set of topics: first, their work: *What does their work look like? What strategies against mis/dis/mal-information do they employ? What gaps in tools to counter mis/dis/mal-information can they identify?* Second, the potential contribution AI4TRUST can make to their work: *What kind of tool are people in need of? What can the AI4TRUST tool do for the organisation and their specific work? How should the tool fit into their workflow?* Third, general questions about the dynamics of mis/dis/mal-information.

2.2.1. Typology of use cases

The work to counter mis/dis/mal-information involves a wide range of activities that different stakeholders engage in. This ranges from **detecting** mis/dis/malinformation, **verifying or debunking** information, **analysing the context** and pattern of the spread of mis/dis/mal-information, to **counter strategies**, such as education and media literacy, and coordination between stakeholders countering mis/dis/malinformation, and policy. Many of these activities are overlapping and interdependent: for example, effective policy-making depends on detection and analysis of mis/dis/malinformation patterns. In the following paragraphs each of these **use cases** is explained further.

Detection

Different stakeholders deploy a range of different tools to detect mis/dis/malinformation depending on the structure of their organisation and the purpose of their work. Many fact-checking organisations either use public channels where members of the public can submit potential pieces of mis/dis/malinformation to be verified, in addition to monitoring social media and news sites. Many organisations also work with Meta and receive claims through a Meta provided platform. In general, fact-checking organisations often work in high-paced environments tackling fast-spreading pieces of mis/dis/malinformation making editorial decisions on what needs to be verified and what cannot be fact checked. Hence, they often face the challenge of assessing quickly how fast a piece of information is spreading (virality) and the potential harm it can cause. Both factors determine whether organisations fact check a piece of information.

Journalists and media organisations encounter mis/dis/malinformation from diverse sources that require verification. News agencies often have dedicated fact-checking departments that focus on monitoring and detecting the spread of mis/dis/malinformation, in addition to the verification processes taking place during regular reporting. News agencies are put under additional strain due to the rapid spread of mis/dis/malinformation

which shifts capacities away from other work. While fact-checkers and journalists focus much of their attention on identifying individual claims, they are also interested in detecting patterns in the spread of information. The latter is the focus of policymakers and civil servants: they largely rely on detection of patterns of misinformation by law enforcement, intelligence agencies, experts, and private service providers of social listening tools to detect both the spread of misinformation, targeted disinformation campaigns, as well as Foreign Information Manipulation and Interference (FIMI)—the threat of external actors or governments manipulating the information ecosystem. A major concern for policy makers is the impact of mis/dis/mal-information on elections and the health of democracy in its wider sense. Different governments have adopted AI social listening services at different speeds due to concerns for data protection.

For policymakers, multilingual capabilities are needed to monitor and understand misinformation across different linguistic contexts. While engaged in the process of countering mis/dis/malinformation through media literacy programs, educators are often not involved in detecting mis/dis/malinformation as such. But academic researchers employ a range of different methods to study this phenomenon: from developing algorithms to flag potentially suspicious content, employing social network analysis (see above), to qualitative methods such as online ethnographies, interviews, and focus groups to study discourses of mis/dis/malinformation and its effects on how people consume information.

Verification and debunking

The second use case described here is the verification of information or debunking of false claims. There are a range of different tools that fact-checkers and journalists use to check false claims in the forms of text, image, video, or audio. Fact-checkers toolboxes include technological tools as well as journalistic tools of verifying and contacting sources. Technical tools include Google Search, Reverse Image Search tools such as TinEye, Google's reverse image search, reverse video search, Maps, and translation tools, such as Google Translate, or DeepL. To verify information shared on social media, X's (formerly Twitter's) advanced search tool also allows fact-checkers to confirm whether a specific account has indeed shared a certain tweet. Similar tools are not available or as useful for other platforms such as Facebook, Instagram, YouTube or TikTok. For videos, tools such as InVid work by searching screenshots from a video and comparing those to existing videos, but are less precise. AI-generated content provides a new challenge for verification: existing tools detecting whether a video or picture is AI generated are still lacking and often only provide fact-checkers with accuracy confidence scores on whether something is

generated by AI. Fact-checkers rely on other strategies to verify AI-generated visual content, but audio clips present a challenge for which fact-checkers have found little work around. For example, they could verify if a person depicted in a specific place actually visited the place by checking their schedules or verifying the visit through a spokesperson. For AI-generated videos where the audio is altered, fact-checkers can detect fakes through observing mouth movements or by finding the original video. Technical solutions to these problems such as watermarking or fingerprinting content are limited due to the easy workarounds for those seeking to share disinformation. To analyse and archive videos, pictures, and audio clips, fact-checkers and journalists also rely on automatic transcription software which is not always available in all necessary languages. Arabic, Ukrainian, and Russian especially represent big challenges. As mis/dis/malinformation is also often related to global news (such as information on the war in Ukraine) or issues such as migration, fact-checkers regularly need to debunk content in different languages. Depending on the size of the organisation, institutions also often rely on their own database or public databases of debunked content, such as Google Fact Check Tools. In addition, online archiving tools such as the Wayback Machine are used to archive the original content that has been verified or debunked to provide evidence.

Analysis of mis/dis/mal-information

Analysing the patterns of mis/dis/malinformation is another important use case for tools to detect and counter mis/dis/malinformation. Analysing mis/dis/malinformation patterns can help different stakeholders to understand the ways in which information spreads across countries, platforms, and languages over time. Detecting the ways in which certain narratives or disinformation campaigns work similarly in different countries or travel from one context to the next can help stakeholders to prepare to develop prebunking content or debunking content. While not all previous research confirms that mis/dis/malinformation crosses boundaries (Cinelli et al. 2020), fact-checkers have the impression that specific claims or campaigns occur in different countries and contexts. This can also inform platforms and policy makers to develop better counter strategies. Many fact-checking organisations do not have technological tools that automatically analyse and visualise the spread of misinformation or specific disinformation claims. Policy makers have limited access through social listening tools provided by the private sector.

Better analysis can also help fact-checkers and journalists make better editorial decisions on where to focus their energies: currently different stakeholders have different methodologies by which they measure the virality (speed and spread of a piece of

information) and potential harm of mis/dis/mal-information, depending on their organisation. They often lack access to resources that help them to understand the spread of a claim across different platforms, whether this claim is associated with specific patterns or disinformation campaigns, and whether similar claims have been debunked or countered in different contexts.

Communication and education

Communicating debunks, for example when civil services leverage experts to offer counter-narratives, and educating both the public as well as other fact-checkers or journalists represents a large aspect of the work of countering mis/dis/mal-information. Tools to detect mis/dis/malinformation are an intrinsic aspect of communication with the public about the current information disorder, since they play a large role in how fact checking organisations and journalists demonstrate transparency in their fact-checking methodologies and debunking articles. When a fact-checking organisation publishes an article debunking a certain claim, this article usually explains in detail what tools were used so that readers can verify the process themselves and come to their own conclusions. These tools are also used in educational materials and workshops for the public or to develop capacities of other journalists and fact-checkers. See for example, AFP's toolbox and training material.[10]

2.2.2. Typology of users

This section outlines the set of stakeholders who would be **potential users of the AI4TRUST toolbox and platform**. The section briefly sketches out who these groups of users are to map the diversity of different groups but also to highlight the difference among each group of stakeholders/users. The following section synthesises the different needs for each group of users.

First, **policy makers and civil servants** make up the first set of potential users of the AI4TRUST platform and toolbox. As outlined earlier, policy makers often rely on external actors for the verification of information as well as analysis of the information ecosystem. These actors include trusted third-party organisations, journalists, or internal communication departments. Some institutions use private companies offering social listening tools, while others rely on open-source intelligence from human experts.

[10] https://digitalcourses.afp.com/.

Second, **fact-checking organisations** represent the second group of stakeholders that are potential users of the AI4TRUST platform. Fact-checking organisations engage in a range of different activities from detecting mis/dis/mal-information, verifying and fact-checking specific claims, communicating their debunks to the public, developing educational tools, community building, to research collaborations, and policy advice. Across Europe fact-checking organisations have different technical capacities, funding, workflows, target audiences, and institutional structures. Fact-checking organisations often work in fast-paced environments, as they often engage with rapidly spreading stories and news that require debunking and verification.

Third, **journalists** represent a similar yet distinct group of potential users. Much of the fact-checking work journalists do, overlaps with the work of fact-checking organisations, but goes beyond it. Journalists represent an even more diverse set of potential users, with news organisations ranging from large agencies and media organisations with international reach and dedicated internal fact-checking departments, to national and regional newspapers, working a large set of different languages.

Fourth, **researchers** are another important potential group of users. Academic research has provided crucial insights into the spread of disinformation (see section 1), as well as the diverse ways in which information can be misleading and decontextualised. Research on mis/dis/mal-information also takes place in civil society and human rights organisations, such as Amnesty International and Human Rights Watch or research organisations such as Forensic Architecture which use open-source intelligence tools to investigate human rights violations and international conflicts. The need for researchers interested in social network analysis is detailed in section 3.2.

Finally, a last set of possible users of this tool might be **educators and other organisations** aiming to build media literacy tools which rely on both tools to demonstrate the methodology of verifying information as well as access to use cases to build workshops and courses.

Academic research has been conducted on how journalists and fact-checkers work in the field of fact-checking, but **not much work has been done to study how policy makers or researchers conduct their work**.

2.2.3. User requirements

This section sketches out the set of differing and converging **needs that different users** have from a platform or a toolkit to detect and counter mis/dis/malinformation. The section first addresses the current **gaps between existing tools and strategies and user needs** for different use cases and, second, argues for the need to design any tool in a way that can **allow users to integrate a tool into their specific workflow**. As becomes clear in this section, the gaps between current solutions and needs are vast and too heterogeneous to be addressed by a single tool or platform. The section nevertheless attempts to scope the breadth of gaps and needs to assist a process of reflection on what specifically can be addressed by the AI4TRUST tool and platform.

Detection and verification

Mis/dis/mal-information is a very **dynamic and rapidly adapting phenomenon**. Fact-checking struggles to keep up with the pace at which false claims spread. Disinformation campaigns frequently occur in the aftermath of divisive news stories that attract global attention, such as the global pandemic, the war in Ukraine, or more recently, the conflict in Gaza. News and social media content from the ground are shared and misrepresented which means fact-checkers and journalists around the world and in Europe are faced with the challenge of having to verify audio and visual material in many languages and alphabets they are unfamiliar with. They are therefore in need of **reliable tools to translate content** (e.g., audio, video, and text) into their native language as well as tools that detect what language is spoken and where. Both types of tools can help to identify the origins of audio and video content that are often shared out of context. Especially in the wake of conflicts or natural disasters, out of context pictures are shared at large volume and require fact-checkers to confirm their origins. **Multilingual capabilities** are also crucial to monitor and understand misinformation across different linguistic contexts (more details below). The problem of language is also complicated by a lack of knowledge over local contexts or access to knowledge on who to contact locally to verify information. Furthermore, it also adds an additional layer in case fact-checkers want to call local institutions in another country to verify something because they need to be able to understand nuances and specifics of the information with which they are provided.

Intricately connected to translation tools are better detection tools of **AI-generated content.** In the wake of generative AI, fact-checkers say that more synthetically generated content is spread on social media and in news media. The verification of synthetical content brings new and rapidly developing challenges. Most urgently needed according to fact-

checkers is a tool to help detect AI-generated or manipulated audio content. As outlined earlier, AI-generated images or videos are easier to verify due to the additional context of the visual medium. In addition, AI generated or manipulated videos and images are often identifiable through visual clues, such as the display of additional fingers, mismatched mouth movement to the spoken text, or blurring where a face has been overlaid on an existing video. But for audio none of these options are available. In general **audio and video content** brings extra challenges to fact-checking since they require additional transcription for archiving and to search and verify the origins of an audio or video clip. Here automatic transcription tools could be helpful and save time. In addition, false information generated or hallucinated by large language models (LLMs) is currently difficult to detect and counter.

Another aspect observed by fact-checkers is the increased use of **visual or image-based content** to spread mis/dis/mal-information. Images travel further and spread faster since they do not need to be translated from one language to the other, allowing disinformation to also spread quickly from one country to the other. In addition, pictures are more visceral in how they communicate their message. One example is a picture of the Ukrainian President who was depicted wearing a t-shirt with a swastika which had been artificially added. While easily verified with finding the original image, the manipulated picture spread far and fast. Often images are reused to make similar or also unconnected false claims. These are referred to pictures out of context. Fact-checkers need tools that facilitate the process of accessing a database with previous debunks of pictures out of context, in order to speed up their work. Fact-checkers wished they had an automated tool to help facilitate the following: first, detecting the same image or content out of context shared on other platforms or in other places on the same platform, and, second, generating a debunking article for the fact-checking organisation (that mirrors the organisation's style of writing) which can then be posted next to the picture after being proof-read by a fact-checker. Additional suggestions for automation were made concerning the **detection of the same false claim in different formats**, so that debunking campaigns could target all versions of the same false claim. In addition, fact-checkers struggle when using reverse image search tools which they hope to use in analysing details in images. An example shared concerned the need to verify a badge on a soldier's uniform to verify where the soldier depicted in a video was from. When reverse-searching the badge, rather than show pictures with similar badges all reverse-image search tools only presented the fact-checker with images of similar shapes to the symbols depicted on the badge. Instead of scoping what should be achieved with the AI4TRUST tools, this example merely illustrates the vast gap between what fact-checkers *theoretically* need and the existing capabilities of image recognition

algorithms. It also illustrates that there is a gap between the expectations of AI-driven tools and what AI-tools can achieve.

Finally**, existing tools to reverse search videos were less than satisfactory for fact-checkers** because they could not verify the way the search worked, which meant they would spend time looking through entire videos to verify whether it was the one they were looking for. Many fact-checkers therefore resorted to taking a screenshot of a video they needed to verify which they looked up in a reverse image search. One fact-checker shared that in a dream world they would have access to a reverse-video search tool that worked smoothly — something that is far from today's state of the art. Finally, image or video-based platforms like Instagram or TikTok also presented additional challenges to the work of fact-checking: while Twitter has a tool for advanced search which allows fact-checkers to verify if a certain account shared a certain tweet or whether it was fake, neither Facebook, nor Instagram nor TikTok have any comparable feature. TikTok in addition was difficult to navigate for fact-checkers because the algorithm decides what users can view, making it exceedingly difficult to look for specific content.

Analysis of mis/dis/malinformation

While there is a need for policy makers and civil servants to verify content to ensure accurate information is represented in their communication to the public, they are predominantly interested in understanding both specific and general patterns of mis/dis/mal-information as they craft policy interventions. This also includes a need to **understand what strategies to counter mis/dis/mal-information work in which specific context**. The needs of different groups of civil servants and policy makers might differ across different countries. In addition, many civil servants lack technological expertise to use specific tools, which calls for the design of tools that can be easily accessed and utilised with little training. As outlined earlier, some governments and policy makers access this information through tools provided by companies such as Grafika, Logically or Storyzy. Adoption of these tools is less wide-spread in the EU as there are more concerns over data protection and AI ethics and safety. Policy makers stressed the need for a tool that would allow them to analyse the dynamics and patterns of mis/dis/malinformation which follows European data protection standards, ethics, and values. In addition, concern was raised over using tools which do not allow to keep data within the institution.

Fact-checkers also expressed a great **need for analysis tools to discover the dynamics of mis/dis/malinformation**, as they generally do not access the tools provided by private companies mentioned above. Something that researchers also urgently need. Fact-

checkers discussed how they observe patterns in disinformation campaigns that they urgently want to study: both for their own understanding and to be better prepared about future campaigns, but also in order to utilise this research and analysis as evidence for their policy advice and advocacy efforts. One aspect that was highlighted was a **need to better understand how disinformation campaigns and misinformation travel from one geographic context to the next**, as noted above. For example, fact-checkers observed how similar narratives about electoral fraud in postal votes popped up in several countries, even though each country had completely different electoral systems and different challenges. A mapping and better understanding of these dynamics could enable fact-checkers, journalists, but also policy makers to develop a better response. Especially for fact-checking organisations whose resources are often thinly stretched and who must make tough editorial decisions on what to focus their efforts on, nuanced analysis about the dynamics of mis/dis/malinformation is needed. This also includes more context about specific claims with regards to the spread and potential harm of content. This can assist organisations in their decision making on whether and how to take a specific case of mis/dis/mal-information. **Cross-platform information would be especially helpful**, since journalists and fact-checkers cannot independently verify how many times something has been viewed or shared on, for example, TikTok. The ability to detect coordinated campaigns and/or map misinformation networks was valued, rather than just flagging individual pieces of potentially false content. Fact-checkers also stressed that it was important that a tool could distinguish between flagging highly organised disinformation campaigns from cheap, simple hoaxes with apparently no purpose.

Communication and education

Fact-checkers were particularly concerned about understanding how to best communicate their debunking or prebunking articles to the public. **Understanding how to communicate** on which platform was crucial, as well as **understanding what format** was effective in countering what kind of mis/dis/malinformation. They argued that they also lacked tools to evaluate how effective a debunking article had been: how and where it had been shared. This would also assist smaller organisations in their reporting and in demonstrating impact to funders. Fact-checkers discussed wanting a tool that could help them turn a debunking article into “lyric video” style content which could be shared on more visually oriented platforms. For those developing educational materials for media literacy, access to a **platform that archives information** they can use for workshops and sessions was also mentioned.

Coordination

Given the highly interconnected and fast-spreading nature of mis/dis/malinformation, a large **need for collaboration and coordination** between different fact-checking organisations, journalists, researchers, and other stakeholders in different countries exists. Fact-checkers have in the past shared data on already debunked content across Europe, especially for global news events such as the war in Ukraine. A tool developed to help counter and detect mis/dis/malinformation can be helpful in facilitating this process of coordination by providing fact-checkers and journalists tools to help them **archive and share existing debunked content**. Similar platforms exist such as the EUDisinfoLab which conducts research and shares tools and knowledge around countering disinformation.

Design requirements

Different stakeholders — from policy makers to fact-checkers — stressed that **any tool needs to be flexible to adapt to changing disinformation tactics and social media platforms**. In addition, tools need to be able to fit into the workflow of different institutions and stakeholders: this means users ideally need the option to cherry pick which aspects of the tool should be able to be into their internal systems and databases through an API. Verification tools such as reverse image or video search are not necessarily internal to organisations but can be used in a browser. Different language capabilities were also stressed: the tool should be able to detect information in different languages, but also be used in different languages.

Conclusions

The scoping exercise in this section has revealed a significant gap not only in needs and technical solutions but also in the expectations regarding **what AI could contribute** to the efforts in countering mis/dis/malinformation. This section did not aim to specifically scope out the specific contributions the AI4TRUST tool should make to this work, but to underline the necessity to **understand this vast range of needs, hopes, and expectations different stakeholders have towards technical solutions**. Many of these are beyond the state of the art and the intervention the AI4TRUST tool will deliver. In this sense, this section calls for a nuanced understanding of these needs, hopes, and expectations which can inform both the specific intervention of the AI4TRUST tool and the way the usefulness of the tool is communicated to potential end users. It also **scoped the availability of existing tools and platforms**, thereby highlighting the necessity for the AI4TRUST platform to situate its own intervention distinct from existing tools and platforms.

3. Implications for the project

The creation of an automatic tool to counter mis/dis/malinformation on social media in real time is faced with several **practical, theoretical, ethical, and legal challenges**. This section begins with a **critical appraisal of digital cross-platform analysis** (subsection 3.1). Secondly, this section describes and puts into perspective the challenges raised by **platform data diversities**, beginning with data types (see 3.2.1) and their implications for compatibility across platform ontologies (see 3.2.2). From these observations stems specific attention to **language** (see 3.2.3) and terms of use diversity (3.2.4). Finally, we address the associated **ethical challenges** to bear in mind when addressing the challenges raised above (subsection 3.3), and suggest practical solutions that may be in place to face them (subsection 3.4).

○ 3.1. Cross-platform intervention

The sociological study of the Web has been fostered by the recent advances in digital methods (Gerrard, 2018). Simultaneously the **"platformisation"** of the **Web**, whereby online spaces such as social media are increasingly integrated with each other and the wider Web, has increased the flow of data and information circulating through and from the Web (Pearce et al., 2020). As a result, researchers may have access to massive flows of data through platform-specific application programming interfaces (APIs). This technicality of data access has ironically normalised single-platform research (Rogers, 2017). Indeed, despite the growing number and diversity of social media platforms, most existing research on online content sharing has focused on data obtained via X/Twitter's public API. Prominent research on major disinformation themes have claimed that these single platform analyses were generalisable beyond the specific Twitter space (Pearce et al., 2020). For instance, researchers have claimed that tweets provide 'a proxy for climate change discourse among the general public' (Kirilenko et al., 2015). However, seeing the "platformisation" of the Web and the popularisation of digital methods, **cross-platform research appears to be the most adequate approach** to assess integrated platforms, and presents numerous advantages.

Cross-platform intervention allows for a comprehensive and comparative understanding of online flows of information. Social media platforms are characterised by different affordances and features (Bucher and Helmond, 2018). These govern how users can contribute, share, follow and respond to pieces of content. Different platform ontologies define different platform-specific contexts of interaction and therefore influence the

structure of their social embeddedness. This diversity cannot be captured by a longitudinal, platform-centric approach. Furthermore, the communities that adopt each platform are characterised by different sociodemographic profiles and have different purposes. Thus, they develop community-specific norms and conventional practices (Yarchi et al., 2021). As a result, **platform-specific findings can hardly be generalisable** beyond the specific online space that they were drawn from. On the contrary, analysing multiple platforms enables researchers to identify patterns and trends in the spread and characteristics of online information. Going further, we might identify cross-platform communities whose dynamics of information sharing respond to their socio- demographics, norms, and values, which should be consistent throughout platforms. This type of analysis might shed light on within-platform heterogeneity. Cross-platform analysis could provide a holistic view of this issue across distinct online spaces and allow for more generalisable findings to emerge.

Despite their diversity and lack of common ontology, **social media platforms do not work in isolation**. First and foremost, individual users often hold accounts on multiple platforms, and integrate separate networks on each of them. Influential figures may also be present on various platforms, disseminate flows of content and attract groups of users across the online space. The organic flow of information across the Web is also highly encouraged by sharing features such as cross-platform messaging, chat integration, cross-platform logins, or cross-posting. In this regard, **cross-platform analysis is best suited to reflect the reality** of information and user flows via these interconnection channels.

Nonetheless, cross-platform analysis is challenged by the **limited accessibility of each online platform**. On the one hand, the task of collecting and standardising data from multiple platforms faces variability in data availability, access, and format across platforms. This will require researchers to employ different methodologies and tools for data collection and analysis. On the other hand, without resorting to a specific research design relying on sampled users' approval, research can only access public spaces of content sharing. But this is a very limited window to online information sharing as it fails to account for communications that are private (via direct messages) or semi-private (on closed personal pages or private channels). Cross-platform research also faces the **challenge of tracking units throughout platforms where they might occupy unreachable places**. The next subsection details the issues that arise in terms of heterogeneity of data types and formats across platforms on the one hand, and conditions imposed by platforms' terms of use on the other, while also highlighting potential ways forward. Furthermore, tracking users and communities across platforms poses ethical challenges that require adequate

solutions, discussed in sub-section 3.3 below. Sub-section 3.4 summarises and concludes with possible directions for future work within AI4TRUST.

○ 3.2. Data diversity

▪ 3.2.1 Data types

Online social media come in various shapes and process several types of data, serving different purposes and catering to diverse user needs. Platforms can be categorised into broad types, according to their features and affordances. The following list includes some of the **platform types which might be of interest** when studying online mis/dis/mal-information.

Social networking platforms invite users to create an individual profile (personal or professional) and to connect with other users. Platforms such as Facebook and MySpace follow this logic. They are the origins of online social media.

Microblogging platforms facilitate the sharing of short textual and/or media content with their followers in real-time. This is the privileged feature of X/Twitter, which is responsible for the interest it has famously attracted from social science research.

Photo and video sharing platforms focus on the sharing of images and/or videos which can be the original creation of the channel that publishes them. These platforms, which include TikTok, YouTube, and Instagram, increasingly capitalise on sharing features allowing for content to circulate across channels via reposting or remixing, thus connecting users through content dissemination and editing.

Discussion forums and channel-based platforms enable users to engage with specific interest-based communities. Among these platforms, Reddit and Telegram are some of the most famous ones.

Users on all of the aforementioned social media platforms share a diverse range of interactions and content, providing ample material for research.

Text-based posts, comments and replies, and media captions are present on most social media platforms. In textual content, users can both share any type of information ranging from news and opinions to activities and location. Text can also be descriptive of another piece of content contained within the same post, or interactive with content posted by

another user. URLs are a specific type of data generally shared in text-based content; they are one of the most forward ways for users to share information from a third party (user/platform).

Multimedia data containing images, video or audio content allow users to share information or experiences in a lively way. This content is potentially more impactful than text alone as it communicates emotions in a stronger and/or more direct way. As a result, multimedia content tends to go more viral (Pearce et al., 2020), and Internet culture is increasingly rooted in the creation and sharing of visuals that can convey information and emotion rapidly and memorably. “Meme culture” (i.e., the sharing of humorous or symbolic images and videos) is a major illustration of this phenomenon.

Content metadata provides information on the context of publication, categorised content, and interactions with a post. This may include timestamps, hashtags, or mentions, as well as view count, or search queries that brought users to this content.

User profile and engagement data provide information on the users’ profiles such as personal details or profile pictures, as well as friends/followers/connections count. These are of particular interest to social network analysis as they paint the picture of users’ online social identity and concerns.

Nonetheless, the very type of data (e.g., textual, visual) is not the focal point of social network analysis. Social network research studies the context of interaction around a piece of content, general theme, or event, and in this context, the type of content matters less than the sharing characteristics and the communities who disseminate it. In practice for our study, **social network analysis (SNA)** aims at identifying users who share content with many others (the “superspreaders” mentioned in subsection 1.4 above), or “bridge” users who connect two otherwise disjoint subnetworks, thereby allowing transfer of information from one to the other. A segregated network, for example, is one with no (or with very few) bridges, so that any view that spreads internally does not have a chance to get corrected by incoming counter-information. This is how social network analysis can shed light on diffusion phenomena, without necessarily adapting this framework to the specific content shared.

Nevertheless, there is a growing literature that endeavours to **combine the analysis of the social networks between users** (the superspreaders, bridges, etc.) **with an analysis of the content dynamics**. Indeed, the recent development of the wide range of digital platforms described above has simultaneously led to a diversity in interaction types between users

and a profusion of content sharing. As the dynamics of the latter seem to have much in common with the former, the creation of **novel modelling frameworks which jointly feature social structure and semantic characteristics ("socio-semantic networks")** might be the preferred way of analysis information sharing on social media where contexts of collective interaction between users also create ties between ideas (Roth, 2013).

▪ 3.2.2. Compatibility across distinct data ontologies (greatest common divisor)

For the sake of this research, it might be necessary to **establish a common ontology** to represent knowledge and information across distinct social media platforms. Ensuring that data from various sources are standardised to some extent will be a requirement for the development of a mapping and transforming tool that works with a fixed-type of input. To do so, the following **commonalities need to be identified across platforms**:

- units of analysis defining the context/environment of interactions (**where/when**);
- individuals that interact with each other and are connected with their multilevel environment (**who**);
- content that connects individuals and defines what they are interacting about (**what**);
- link type that defines how individuals or groups of individuals are connected to one another (**how**).

However, researchers should bear in mind that **objects found on different platforms might not be comparable** (Rogers, 2017). For example, hashtags are used more liberally on Instagram than on X/Twitter, and although this object has the same technical function on both platforms, its symbolic function greatly varies. Similarly, images play an important role on all of the main social media platforms, but they are not exhibited in the same way across these spaces: social networking platforms are built around the sharing of textual content, but ultimately, posts that include visuals are more likely to go viral; on the contrary, video-sharing platforms such as TikTok are built to connect users around visual content, but they increasingly include text in the videos, such as subtitles, headlines, etc.

These platform effects question the extent to which **cross-platform research is subject to digital bias** (Pearce et al., 2020). When comparing distinct social media data, researchers should acknowledge that platforms are a tripartite composition of users, algorithms, and data, which might result in several issues (Marres, 2017). Firstly, the data and content selected by the researcher is only a partial account for the entire platform, which might be

biased by the researcher's choices. Secondly, research instruments themselves might introduce bias through embedded algorithms for example. Thirdly, a few methodological issues might introduce bias in the data, due to platform accessibility (e.g. accessibility to private conversations) or social media research tools (e.g. social media research is biased towards textual analysis because its tools are more adapted to textual queries). These biases represent a **challenge for cross-platform analysis** because it might not be trivial nor valid to outline a common ontology across diverse platforms. Yet, adopting an "**affirmative approach**" **to digital bias** (Marres, 2017), i.e. accepting these biases as an inherent component of our object of study rather than attempting to neutralise them, could enrich our findings with insightful perspectives on platform effects (Pearce et al., 2020).

▪ 3.2.3. Language diversity

Social media display worldwide content in **multiple languages**. The analysis and interpretation of multilingual content requires researchers to acquire or outsource proficiency in the investigated languages. When resorting to translation, social scientists should make sure to preserve the social properties of language as a research material. Yet, **social media are simultaneously multilingual and multicultural**. Therefore, addressing linguistic and cultural differences raises **multifaceted sociological and technical challenges**.

Studying mis/dis/malinformation at the European level presents unique linguistic challenges, not only as **the EU encompasses 24 official languages**[11], but also because the global nature of massive disinformation campaigns connects Europeans with content and areas in foreign (non-European) languages (see paragraph 2.2.1 above).

Our research needs to capture the social features of content and exhibit network structures of information flows across languages (Eleta et al., 2014). When studying online information flows, a **geographical perspective** would help tracing content across borders and cultural groups, to evaluate how territorial and/or national differences interact with information production and dissemination. This will raise the **major challenge of conceptualising culture, and accounting for geographical distribution of online content**. Language itself is not a reliable proxy for geographical origin, as there is within-country variance in languages and as some languages are natively spoken or purposefully adopted on social media across the globe. For instance, on platforms where English is the majority

[11] https://european-union.europa.eu.

spoken language, new users are likely to adopt it independently from their native language (Eleta et al., 2014). However, nations cannot be treated as homogeneous cultural entities, because within-country cultural variance even exceeds the between-countries variance (Sheldon et al., 2020). Some languages such as Spanish, French, or Portuguese are natively spoken on different continents, by extremely culturally diverse populations, and are subject to dialectal variation. Language alone cannot reliably connect content to a geographical location.

Language use may also **vary substantially across demographics** such as age and gender (Schwartz et al., 2013), especially in informal contexts like social media. Research has found clear distinctions, such as use of **slang, emoticons, and Internet speak among younger internet users**, and progression of references to school, college, work, and family when looking at the predominant topics across all age groups (Schwartz et al., 2013). This represents a unique opportunity to shed light on **the relations between differences in language use across demographics** and the formation of homogeneous groups and information reception online.

To seize these opportunities, we may resort to interdisciplinary collaboration to ensure that the meaning of content and the cultural context and nuances are preserved. The AI4TRUST consortium may need to turn to **translation experts and (socio-)linguists** to treat raw semantic data before conducting sociological analyses.

From a technical perspective, **existing tools for information analysis might not be applicable nor perform equally on all languages**. Most existing tools for automatic assessment of online information are the result of Western-centric research, and therefore, they are best suited to work with English, but less so for other languages, especially Eastern ones. This might alter the generalisability of our findings, and makes way for further work to be done to improve algorithms' applicability outside of the West.

▪ 3.2.4. Terms of use diversity (access/life data in future)

Social media platforms impose a diversity of agreements on their users and on researchers requesting access to their data. These terms of use, also known as "**terms of service" or "user agreements"**, establish the rights, responsibilities, and expectations for both the users and the platform. Their specificities vary across platform, space, and time. Cross-platform analysis requires methodological coherence, but accommodating nuances in distinct terms of use makes it cumbersome to do so. The diversity and dynamics of terms

of use challenges data access and durability of a multiplatform lively tool that would be plugged into social media data.

First and foremost, data access and availability are contingent on platform-specific policies. **While some platforms provide open access to large volumes of data through public APIs, others are less research friendly and impose stringent restrictions or require explicit approval**. These conditions are likely to be altered over time, as we've seen with X/Twitter's API shutting down. The horizontal scope and vertical depth of social media research depend on the conditions of data access and availability, but they are entirely defined by platforms themselves.

Not only do the dynamics of terms of use vary throughout social media platforms, but they also show **heterogeneity across time and geographic locations**. The relevance of research findings may be affected by regional variances in use, given that the regulatory environments and patterns of disinformation might vary globally. Furthermore, because online language is dynamic and susceptible to periodic changes, researchers must be on the lookout for trends in the data and continuously adjust their approaches accordingly.

Social media platforms' terms of use frequently include content limits that **specify what kinds of content are allowed and what are not**. These are often a formalisation of platforms' compliance with their social responsibility, i.e. the legal and/or ethical obligations and commitments that these platforms have toward users, society at large, and the broader digital ecosystem. As influential places in the digital public space, social media are expected to ensure their users' online safety and well-being. These restrictions might impact users' spontaneity in expressing themselves online. Yet, in practice, internet users have collectively adopted **several workarounds and alternative practices that enable them to bypass these restrictions**. For example, many will replace letters with resembling numbers in forbidden terms, so as not to be signalled nor restrained. When studying online mis/dis/malinformation, we will need to identify these workarounds as they are likely to **signal controversial or potentially harmful content**.

○ 3.3. Associated ethical challenges and solutions

To ensure that research with human subjects does no harm and respects people's freedom and dignity, scientists are routinely expected to put in place procedures that protect privacy and data protection, most commonly through de-identification, pseudonymisation, or

anonymisation of data[12] (depending on the circumstances) and to obtain informed consent from data subjects. However, these commonly adopted solutions do not fit well with the requirements of social network analysis, and over time, significant efforts have been devoted to finding alternative approaches (Breiger, 2005). Below, we summarise these challenges and present **emerging solutions**, adaptable to the needs of research to be performed within AI4TRUST.

First, **full anonymisation** (and in some cases, even weaker forms of **pseudonymisation and de-identification**) may not be possible in social network analysis. As Charles Kadushin (2012, p. 188) puts it, unlike other fields of research with human subjects, names (or unique personal identifiers such as nicknames or initials) are not incidental but the very point of network data collection: they are necessary to match senders and receivers of ties, and to associate ties to relevant attributes. Common solutions to this problem consist in either requesting an acronym for each individual instead of the full name, or collecting real names that will be anonymized later. Institutional Review Boards (IRBs) and Research Ethics Committees (RECs) routinely approve the apparently straightforward acronym-based mitigation measure; however, extant evidence is that it entails severe disambiguation issues and results in significant loss of accuracy. For this reason, the guidelines formulated by the Social and Public Health Sciences Unit of the University of Glasgow (2022) recommend the "**delayed anonymization"** alternative, offering **"research exceptions" under the GDPR as justification**. It should be recognized in this respect that it might still be too early to anonymize or even pseudonymize network data at pre-processing stage (that is, *after* collection and *before* analysis). Depending on the research question and setting, it may be preferable to leave anonymisation to the post-processing phase (or to put it differently, *after* analysis and *before* presentation of results). **Delayed anonymization is a valid approach within AI4TRUST, especially with cross-platform analysis**. Of course, working with non-anonymous data requires extra security measures, such as storage in a protected institutional repository, encryption, and controlled access.

There is consensus that individuals should not be identifiable in any presentation of results to stakeholders, fellow scientists, or the general public (Tubaro, 2021). To achieve this,

[12] De-identification means that explicit identifiers such as names are hidden or removed. Pseudonymisation is the process of replacing identifying information with codes, which can be linked back to the original person with extra information (keys). Anonymisation is the strictest and most radical procedure, consisting in the irreversible process of completely removing all references to personal data, so that there is no way to re-identify data subjects.

removal of personal identifiers may be insufficient especially when data and/or results are presented in the form of visualisations. Even when node colours or shapes are de-identified and reflect broad categories such as gender or department, individuals' positions (such as those who are isolated or at the other extreme, those who are very highly connected) may reveal who is who, notably to audiences already familiar with the research setting.

Small networks are particularly vulnerable to reidentification. For example, it may be easy to recognize "the only high-ranking woman in the Boston office" (Borgatti et al., 2013, p. 48). In short, the power of visualisation is also a potential threat: 'Network analysis does its "magic" by making visible what was not visible before and reveals connections between individuals and groups who may not have wanted this information to be made public' (Kadushin, 2012, p. 188). However, It must be acknowledged that **such risks are unlikely to arise within AI4TRUST, where large datasets will be used**. In the event they do arise, researchers can harness the potential of network visualisations themselves to mitigate their effects: they can combine visual variables like size, colour, position, and shape in informative but privacy-protecting ways, and most importantly, they can **use network layouts that give less emphasis to rare or extreme network positions**, so that they avoid magnifying centrality, isolation, or segregation (Tubaro et al., 2016). Within AI4TRUST, a suitable approach consists in exploring, whenever possible, **alternative visualisations of the same data** along with provision of **contextual explanation** of their meaning and limitations.

Another standard requirement that can be problematic in social network analysis is **informed consent**, as it is difficult to obtain from the people who are part of the networks of data subjects as their contacts, friends, online followers, etc. but are not themselves participants to the study, and are frequently impossible to reach. Already known in more traditional social survey settings, this problem has become more prevalent in online communications, where data are almost never limited to describing individuals, and also include their social contacts and communication patterns. Current approaches to address this issue typically rest on the idea that the members of the network of an individual ("*ego*") are representations of the social environment of *ego* and therefore, they constitute ego's data, not the data of other people ("*alters*"). This implies that **only *ego*'s consent should be sought** (Robins, 2015; Perry et al., 2018). In practice, researchers often rely on this argument to request from their IRB (or REC) a waiver of consent from *alters*, which can be granted when strict confidentiality is guaranteed and the research involves minimal or no risks for these alters. In US institutions, this may involve qualifying *alters* as 'secondary

subjects.' In Europe, consent is only one of the legal bases for data processing under the GDPR, and **research with personal data may also be lawful if undertaken as part of academic or public interest research** providing there is no likelihood of substantial damage and distress to the data subject. More generally, it is usually expected that whenever possible and adequate, these subjects be provided with information about the study, even if this may occur after data collection and processing. According to the above, this information provision may be limited to "egos."

Even when formal informed consent cannot be obtained and alternative solutions are leveraged as illustrated above, researchers are now globally more mindful of the expectations of privacy that users of online networking sites might have (Chu et al., 2021). For example, there is growing consensus around the idea of **"contextual" privacy** (Nissenbaum, 2009), whereby information shared on an online social networking site (such as Facebook, Twitter/X, Reddit, or other) is not intended for other uses and cannot be taken as 'public' in the same way as, say, the formal speech of a politician or a journalist's article in the press, deliberately intended to reach out to large audiences. Similarly, Marwick and Boyd (2014) claim that privacy is achieved in networked publics through **joint negotiation of boundaries and relationships**, even when perfect control over one's data is impossible to achieve. This suggests drawing a distinction between social media profiles that are manifestly public, such as those of politicians, institutions, press outlets, and public figures, and those that are unlikely to be intended as such and thus need greater efforts for protection.

While these considerations suggest that social network analysis raises some **ethical challenges**, this method is also an ideal lens to attest that even well-established rules and practices are not necessarily universal, and may need contextual adaptation. Even when general guidance (notably two issues of the journal *Social Networks*, in 2005 and 2021 respectively, and a state of the art in *Network Science* planned in 2024) and umbrella protocols are in place (in preparation jointly by the scholarly associations International Network on Social Network Analysis, INSNA, and the Network Science Society, NetSci), there may be no one-size-fits-all solution. Thus, Tubaro et al. (2021) recommend that researchers adopt a **reflexive approach** and carefully examine the applicability of different solutions to each particular problem. Likewise, ethical authorities such as institutional review boards (IRBs) and RECs should work on a case-by-case basis. Within AI4TRUST, this approach can help members to **proactively identify potential problems** and propose appropriate solutions.

○ 3.4. Potential outcomes

To address the aforementioned challenges of **cross-platform analysis**, we suggest a lateral and innovative entry point to **perform an analysis that starts from the content substantiating information disorders and moves on to their multilevel context**. Doing so could reveal the same communities of agents that exist around content. More specifically, we suggest beginning by building **different networks** that investigate different and specific digital spaces. This will generate **systematic knowledge and understanding of the contextual dynamics** of mis/dis/malinformation within each observed space. Then, we would seek to tr**ace the circulation of a common piece of content** that channels mis/dis/malinformation (e.g., identical images, videos, clusters of words) which we consider as a **mark of similarity** insofar as it points to a common sensitivity towards circulating/spreading the same instances of distorted information. This sort of **two-step approach** allows us to achieve important outcomes. This would result in **deeper knowledge** of how online mis/dis/malinformation is relationally built within different digital spaces. It would also circumvent the above-mentioned challenges of cross-platform analysis. This means that AI4TRUST could investigate, in an ethically and legally compliant way, the percolation of information disorders across digital spaces.

References

- Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School (HKS) Misinformation Review*.
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236.
- Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: conceptual and methodological challenges. *Social Media + Society*, *9*(1).
- Borah, P. (2022). The moderating role of political ideology: Need for cognition, media locus of control, misinformation efficacy, and misperceptions about Covid-19. *International Journal of Communication*, 16, 26.
- Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916), 892-895.
- Borgatti S.P., Everett M.G., & Johnson, J.C. (2013). *Analyzing Social Networks*. Sage.
- Breiger, R.L. (2005). Introduction to special issue: ethical dilemmas in social network research. *Social Networks*, 27(2), 89–93.
- Bucher, T., & Helmond, A. (2018). The affordances of social media platforms. The SAGE handbook of social media, Sage, 1, pp. 233-253.
- Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2), 591–646.
- Centola, D. (2020). *How Behavior Spreads: The Science of Complex Contagions*. Princeton University Press.
- Ceron, W., Gruszynski Sanseverino, G., de Lima-Santos, M.-F., & Quiles, M. G. (2021). Covid-19 fake news diffusion across Latin America. *Social Network Analysis and Mining*, 11(1), 47.
- Chu, K.H., Colditz, J., Sidani, J., Zimmer, M., & Primack, B. (2021). Re-evaluating standards of human subjects protection for sensitive health data in social media networks. *Social Networks*, 67, 41-46.
- Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS one*, 10(6), e0128193.
- Cinelli, M., Cresci, S., Galeazzi, A., Quattrociocchi, W., & Tesconi, M. (2020). The limited reach of fake news on Twitter during 2019 European elections. *PloS one*, 15(6), e0234689.

- Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Democracy Action Plan, COM/2020/790 final (2020).
- Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1), 1-4.
- Conti, M., Lain, D., Lazzeretti, R., Lovisotto, G., & Quattrociocchi, W. (2017). It's always April fools' day!: On the difficulty of social network misinformation classification via propagation features. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, IEEE.
- De keersmaecker, J. & Roets, A. (2019). Is there an ideological asymmetry in the moral approval of spreading misinformation by politicians? *Personality and Individual Differences*, 143, 165–169.
- DiFonzo, N., Bourgeois, M. J., Suls, J., Homan, C., Stupak, N., Brooks, B. P., Ross, D. S., & Bordia, P. (2013). Rumor clustering, consensus, and polarization: Dynamic social impact and self-organization of hearsay. *Journal of Experimental Social Psychology*, 49(3), 378–399.
- Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41, 424-432.
- Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation (JASSS)*, 20(4), 2.
- Floridi, L. (2003). From Data to Semantic Information. *Entropy*, *5*(2), Article 2.
- Fox, C. J. (1983). *Information and misinformation: An investigation of the notions of information, misinformation, informing, and misinforming*.
- Frey, V. & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, 67:7, 4273-4286
- Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Rumor cascades. In *Proc. 8th ICWSM Intl. Conf. Weblogs Social Media*, pp. 101–110. AAAI.
- Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & Di Domenico, M. (2020). Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature Human Behaviour*, 4, 1285–1293.
- Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12), 4492-4511.

- Grohmann, R., Aquino, M.C., Rodrigues, A., Matos, É., Govari, C., & Amaral, A. (2022a). Click farm platforms: An updating of informal work in Brazil and Colombia. *Work Organisation, Labour & Globalisation.* 16(2), 7-20.
- Grohmann, R., Pereira, G., Guerra, A., Abílio, L.C., Moreschi, B., & Jurno, A. (2022b). Platform scams: Brazilian workers' experiences of dishonest and uncertain algorithmic management. *New Media & Society*, 24(7), 1611-1631.
- Havey, N. F. (2020). Partisan public health: how does political ideology influence support for covid-19 related misinformation? *Journal of Computational Social Science*, 3(2), 319–342.
- Hernon, P. (1995). Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, *12*(2), 133–139.
- Hughes, H. & Waismel-Manor, I. (2021). The Macedonian fake news industry and the 2016 US election. *PS: Political Science & Politics,* 54(1), 19-23.
- Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*, pp. 1–9.
- Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press.
- Kirilenko, A.P., Molodtsova, T., & Stepchenkova, S.O. (2015). People as sensors: Mass media and local temperature influence climate change discussion on Twitter. *Global Environmental Change,* 30, 92-100.
- Lasswell, H. D. (1971). *Propaganda technique in world war I*. MIT press.
- Lazega, E. (2014). Appropriateness and structure in organizations. *Research in the Sociology of Organizations*, 40, 377-398.
- Lazega, E., & Quintane, E. (submitted, 2023). In and out of a politicized crisis in a controversial institution: Social rationality driving network dynamics.
- Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380), 1094–1096.
- Lerman, K., Yan, X., & Wu, X.-Z. (2016). The "majority illusion" in social networks. *PloS one*, 11(2), e0147617.

- Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3), 106–131.
- Lindquist, J. (2022). "Follower factories" in Indonesia and beyond: automation and labor in a transnational market. In M. Graham & F. Ferrari (eds) *Digital Work in the Planetary Market*. MIT Press, pp. 59-75.
- Lindquist, J. (2021). Good enough imposters: The market for Instagram followers in Indonesia and beyond. In S. Woolgar, E. Vogel, D. Moats & C. Helgesson (eds) *The Imposter as Social Theory: Thinking with Gatecrashers, Cheats and Charlatans*. Bristol University Press, pp. 269–292.
- Lobera, J., Santana, A., & Gross, C. (2023). Are we looking at crises through polarized lenses? Predicting public assessments of the official early responses to the COVID-19 pandemic in eight countries. *European Sociological Review*, jcad016.
- Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. John Wiley & Sons.
- Marsden, P.V., & Friedkin, N.E. (1993). Network studies of social influence. *Sociological Methods & Research*, 22(1), 127-151.
- Marwick, A.E., & boyd, d. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16(7), 1051-1067.
- Metzger, M.J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13), 2078-2091.
- McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415-444.
- MRC/CSO Social and Public Health Sciences Unit, University of Glasgow. 2022. *Guidance on processing data for social network research*, available at: <https://doi.org/10.17605/OSF.IO/KVDPT>.
- Naughton, J. (2014). *From Gutenberg to Zuckerberg: Disruptive Innovation in the Age of the Internet*. Quercus.
- Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books.
- Ong, J.C. & Cabañes, J.V.A. (2019). When disinformation studies meets production studies: Social identities and moral justifications in the political trolling industry. *International*

Journal of Communication, 13(20), available at: <https://ijoc.org/index.php/ijoc/article/view/11417/2879>.

- Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Rabello, E. T. (2020). Visual cross-platform analysis: Digital methods to research social media images. *Information, Communication & Society*, 23(2), 161-180.
- Pennycook, G., & Rand, D.G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.
- Perry, B.L., Pescosolido, B.A., & Borgatti, S.P. (2018). *Egocentric Network Analysis: Foundations, Methods, and Models*. Cambridge University Press
- Ratkiewicz, J., Conover, M.D., Meiss, M., Goncalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (ICWSM'11).
- Roberts, S.T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
- Robins, G. (2015). *Doing Social Network Research. Network-based Research Design for Social Scientists*. Sage.
- Rogers, R. (2017). Digital methods for cross-platform analysis. *The SAGE Handbook of Social Media*, Sage, pp. 91-110.
- Roth, C. (2013), "Socio-Semantic Frameworks", Advances in Complex Systems, 16(4-5): 1–26.
- Scheufele, D.A., & Krause, N.M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16), 7662–7669.
- Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9), e73791.
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1), 4787.
- Sheldon, P., Herzfeldt, E., & Rauschnabel, P.A. (2020). Culture and social media: the relationship between cultural values and hashtagging styles. *Behaviour & Information Technology*, 39(7), 758-770.
- Shu, K., Wang, S., & Liu, H. (2017). *Exploiting tri-relationship for fake news detection*. arXiv preprint arXiv:1712.07709, 8.

- Shu, K., Bernard, H.R., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. *Emerging research challenges and opportunities in computational social network analysis and mining*, 43-65.
- Singh, L., Bode, L., Budak, C., Kawintiranon, K., Padden, C., & Vraga, E. (2020). Understanding high-and low-quality URL sharing on Covid-19 Twitter streams. *Journal of Computational Social Science*, 3, 343–366.
- Smaldino, P.E., Janssen, M.A., Hillis, V., & Bednar, J. (2017). Adoption as a social marker: Innovation diffusion with outgroup aversion. *Journal of Mathematical Sociology*, 41(1), 26–45.
- Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R.M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *iConference 2014 proceedings*.
- Stein, J., Frey, V., & van de Rijt, A. (2023). Realtime user ratings as a strategy for combatting misinformation: an experimental study. *Scientific Reports*, 13(1), 1626.
- Thomas, K., Grier, C.,, & Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats* (LEET'12). USENIX Association, USA, 13.
- Tubaro, P. (2021). Whose results are these anyway? Reciprocity and the ethics of "giving back" after social network research. *Social Networks*, 67, 65-73.
- Tubaro, P., Ryan, L., Casilli, A.A., & D'Angelo, A. (2021). Social network analysis: New ethical approaches through collective reflexivity. Introduction to the special issue. *Social Networks*, 67, 1-8.
- Tubaro, P., Ryan, L., & D'Angelo, A. (2016). The visual sociogram in qualitative and mixed-methods research. *Sociological Research Online*, 21(2), 180-197.
- Van de Rijt, A. (2019). Self-correcting dynamics in social influence processes. *American Journal of Sociology*, 124(5), 1468-1495.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
- Wardle, C. (2020). Understanding information disorder: Essential guides. *First Draft*, available at: <https://firstdraftnews.org/long-form-article/understanding-information-disorder/>
- Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Report. Strasbourg: Council of Europe.

- Watts, D.J., & Dodds, P.S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4), 441–458.
- Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific Reports*, 3, 2522.
- Yang, K.-C., Pierri, F., Hui, P.-M., Axelrod, D., Torres-Lugo, C., Bryden, J., & Menczer, F. (2021). The Covid-19 infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1), 20539517211013861.
- Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1-2), 98-139.