Journal Pre-proof

Trustworthy journalism through AI

Andreas L Opdahl, Bjørnar Tessem, Duc-Tien Dang-Nguyen, Enrico Motta, Vinay Setty, Eivind Throndsen, Are Tverberg, Christoph Trattner

PII: S0169-023X(23)00042-3
DOI: https://doi.org/10.1016/j.datak.2023.102182
Reference: DATAK 102182

To appear in: *Data & Knowledge Engineering*

Received date : 16 February 2022
Revised date : 23 April 2023
Accepted date : 24 April 2023

Please cite this article as: A.L. Opdahl, B. Tessem, D.-T. Dang-Nguyen et al., Trustworthy journalism through AI, *Data & Knowledge Engineering* (2023), doi: https://doi.org/10.1016/j.datak.2023.102182.

Trustworthy Journalism Through AI

Andreas L Opdahl (1), Bjørnar Tessem (1), Duc-Tien Dang-Nguyen (1),
Enrico Motta (1,2), Vinay Setty (3), Eivind Throndsen (4), Are Tverberg (5),
Christoph Trattner (1)
(1) University of Bergen, Norway
(2) The Open University, UK
(3) University of Stavanger, Norway
(4) Schibsted, Oslo, Norway
(5) TV 2, Bergen, Norway

Abstract: Quality journalism has become more important than ever due to the need for quality and trustworthy media outlets that can provide accurate information to the public and help to address and counterbalance the wide and rapid spread of disinformation. At the same time, quality journalism is under pressure due to loss of revenue and competition from alternative information providers. This vision paper discusses how recent advances in Artificial Intelligence (AI), and in Machine Learning (ML) in particular, can be harnessed to support efficient production of high-quality journalism. From a news consumer perspective, the key parameter here concerns the degree of trust that is engendered by quality news production. For this reason, the paper will discuss how AI techniques can be applied to all aspects of news, at all stages of its production cycle, to increase trust.

Introduction

The last two decades have put pressure on journalists, editors, and newsrooms (Siles & Boczkowski 2012, Caswell 2019). On the content side, young *digital natives* have different news habits from the older media consumers (Siles & Boczkowski 2012). They rely more on alternative and free information sources and are less likely to pay for news subscriptions (Siles & Boczkowski 2012, Chyi & Ng 2020). Other segments of the population shun mainstream media due to perceived political bias and distrust in authorities (Siles & Boczkowski 2012). These attitudes may become exacerbated by co-ordinated disinformation campaigns (Wintterlin et al. 2020) and amplified in sealed information enclaves, such as online echo chambers (Del Vicario et al. 2016) and alleged search and recommendation bubbles (Spohr 2017, Beckett 2019). Some population segments also lack the language, reading and other skills that are required to appreciate quality news. They may experience news fatigue or exhibit news avoidance (Skovsgaard & Andersen 2020, Gaillard et al. 2021). On the business side, media income from advertising, subscriptions and sales has dropped due to the availability of free online news sources, social media, search engines, and other intermediaries (Siles & Boczkowski 2012, Caswell 2019). In particular, social networking sites (Siles & Boczkowski 2012) exploit their platform power by leveraging information about consumer behaviours to offer individually targeted promotion at lower prices than traditional mass media (Lee et al. 2018). As broken business models lead to newsroom layoffs (Siles & Boczkowski 2012), these challenges become even harder to tackle, creating a vicious cycle. For example, in the US, more than a quarter of all newsroom jobs were lost between 2008 and 2020 (Grieco 2020). Similar developments have been experienced elsewhere (Siles & Boczkowski 2012).

At the same time, many of the same mechanisms that put journalists, editors, and newsrooms under pressure make quality journalism more important than ever. Rigorous evidence-based journalism that is based on robust democratic values and open about its positions and values can help to defuse growing populism and distrust in authorities. Indeed, many analyses[1] show that countries, such as Norway, which have a very strong quality media sector, have been able to address these issues successfully, showing high percentages of access to quality paid content and high levels of trust in the news. Analogously, international quality outlets like The Guardian have been able to manage the transition to the digital era successfully[2], emphasising a direct relationship with their readers based on trust and high quality content. In a nutshell, it is our view not only that quality journalism is extremely important in this new age of polarisation and fake news, but also that the same mechanisms that fuel these negative trends at the same time create an opportunity for quality news organisations to thrive, by providing both readers and advertisers with a platform characterised by trusted and high quality news content.

Figure 1. All four stages of news production can be augmented with AI-support for journalists, editors, and newsrooms. Blue arrows show the forward flow of information content from sources towards the audience, whereas red arrows indicate feedback on trustworthiness and other qualities from the audience back to the earlier production stages.

This paper presents a vision for how recent advances in Artificial Intelligence (AI) can support trustworthy high-quality journalism at "every stage of the journalistic value chain" (Moran & Shaikh 2022). According to the American Press Institute[3], this value chain involves "the activity of gathering, assessing, creating, and presenting news and information. It is also the product of these activities" (Figure 1).

[1] E.g., Index 2022, RSF - Reporters Without Borders <https://rsf.org/en/index>, accessed 2022-05.
[2] Guardian turns a profit for the first time in 20 years, aided by record online traffic, M. Bhattacharjee <https://whatsnewinpublishing.com/guardian-turns-a-profit-for-the-first-time-in-20-years-aided-by-record-online-traffic/>, accessed 2022-05.
[3] Journalism Essentials. American Press Institute <https://www.americanpressinstitute.org/journalism-essentials/>, accessed 2021-10.

Our vision centres on *trustworthiness*, which we see as a central challenge for newsrooms today, with the broader goal to contribute to the *AI for Good* endeavour (Taddeo & Floridi 2018), through innovative media technologies that help journalists and editors promote informed and engaged citizenship. We emphasise that trustworthy AI-supported journalism requires a fine balance between *AI-augmented* human tasks and *AI-automated* routine tasks, acknowledging that "technology is not simply an external tool journalists are forced to assimilate into newswork (though the drive for technological innovation is often fuelled by funders and external actors) but is instead a tool shaped by journalistic practices, needs and norms that similarly alters, sometimes dramatically, everyday newswork" (Moran & Shaikh 2022).

In Figure 1, the *gathering* and *assessing* activities have to do with the trustworthiness of the journalistic sources and of the information they provide, whereas *creating* and *presenting* relate to the users' actual and perceived trust in the news stories they receive. Hence, a vision for trustworthy journalism through AI must take into account both the producers (journalists, editors, newsrooms) and consumers (audience) of news. It must take into account both the regular audience and the adversaries who seek to exploit or diminish trust in the media (Hutson 2021).

A number of contributions have already discussed the relationship between AI and journalism, many of them addressing specific elements of this relationship. For example, (Miroshnichenko 2018) discusses the robotisation of journalism, while (Stray 2019) focuses on the use of AI in investigative journalism. Galily (2018) reflects on the possible development of automated journalism using sports journalism as the example, whereas Lewis et al. (2019) discuss responsibility for libellous content produced by automated journalism. Finally, Beckett (2019) presents a survey of journalists' views on how AI can impact on journalistic practices, carried out by JournalismAI[4], "a global initiative that aims to inform media organisations about the potential offered by AI-powered technologies". According to Lin & Lewis (2022), the "one essential thing" that journalistic AI might do for democracy is to "provide accurate, accessible, diverse, relevant, and timely news about public affairs". Compared to these efforts, our paper focuses more specifically on trustworthiness and on how AI can be harnessed to achieve it.

The rest of the paper is organised as follows: We first discuss the concepts of trust and trustworthiness in the news. We then discuss trustworthiness in each of the four activities separately. Finally, we offer conclusions and paths for further work.

Trustworthiness

Trust in news media is closely related to *credibility* (Flanagin & Metzger 2020). According to Strömbäck et al. (2020), it can be characterised in terms of *fairness*, *bias* (either lack of bias or, we add, being open about the position and values the facts are discussed from), *completeness* (telling the whole story), *accuracy*, and *factuality* (separating facts from

[4] JournalismAI, London School of Economics and Political Science <https://www.lse.ac.uk/media-and-communications/polis/ JournalismAI>, accessed 2021-10.

opinion). The Trust in News project[5] is an ongoing effort to investigate "what digital news sources people trust, why people invest their trust in them, and what publishers and platforms can do to help people make decisions about what news to trust online". Media trust can be investigated at different levels, from trust in *media content*, through trust in *journalists*, individual *media brands*, and *media types*, to trust in *news media* in general (Strömbäck et al. 2020). Our discussion in this paper will focus on the two most operational levels: *the media content* and *the journalistic production processes* behind it.

Trustworthiness is an impression formation process stemming from three qualities: *ability*, *benevolence*, and *integrity* (Plaisance 2014). Trustworthiness of an actor in a domain can be defined as *the actor acting responsibly* towards people that depend on that actor (Jones 2012) and on the actor being *identifiable* and *competent* in that domain. In the news media domain, this translates (from Strömbäck et al. 2020) into *both being and being perceived as fair, handling bias, and reporting in a way that is complete, accurate, and factual*. This is the definition of trustworthiness we will adopt in this paper, with focus on *the media content* and *the journalistic production processes* behind it.

A neighbouring concept from institutional theory is that of *legitimacy*. At the organisational level, it can be viewed as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions" (Suchman 1995). Generalised organisational legitimacy is composed of individual or collective legitimacy *judgements* (Bitekine & Haack 2015, Harmon et al. 2019) that follow different logics in times of institutional stability than in times of change (Bitekine & Haack 2015) - such as in the media today. Schiffrin (2019) also defines trust in journalism as a judgement "about how one assesses the reliability of information being provided". According to her review of credibility and trust in journalism, the main elements of trust are: source credibility, message content, and audience characteristics.

Gathering

Information gathering is the foundation of trustworthy news production. It can nourish trust by offering diverse sources of information and letting audiences trace news content back to those sources. This section reviews central problems of trustworthy information gathering. For each of them, it also discusses central AI techniques along with AI-related opportunities and risks. Table 1 presents an overview.

Routine harvesting: Information gathering involves both active collection of information on demand and passive routine harvesting, for example through subscriptions. In both cases, new uses of AI can seek to make content more trustworthy by relying on diverse and credible sources and by corroborating (or triangulating) overlapping information from independent sources (Bryman 2016). Specific types of *routine* information gathering are already automated with rule-based systems and natural-language processing (NLP) in many

[5] The Trust in News Project, Reuters Institute for the Study of Journalism, University of Oxford <https://reutersinstitute.politics.ox.ac.uk/trust-news-project>, accessed 2021-10.

news organisations. For example, the Wordsmith system gathers information from companies' published earnings reports in order to generate news stories automatically (Miroshnichenko 2018). Reuters has experimented with News Tracer (Liu et al. 2017), which gathers messages from social media streams and uses NLP and ML to detect pre-news events, giving "our journalists anywhere from an 8- to 60-minute head start" on "global news outlets in breaking over 50 major news stories"[6]. This idea can be extended to a wider range of sources, such as sensors (Atzori et al. 2010) to demonstrate agility and increase consumers' trust that the news they receive is up-to-date. An associated risk is over-reliance over the same old automated sources and, correspondingly, less incentive to introduce new ones. AI solutions for widespread routine harvesting must also be enabled to deal with changes in the source data formats, the APIs used, and the available information providers. Further hurdles are legal challenges with web scraping and data ownership.

Broader harvesting: AI techniques can also be used to monitor and identify new developments that engage and build trust with narrower audience segments. For example, U.S. newspaper The Atlanta Voice uses CrowdTangle to identify topics and monitor trends of special interest to the African American community, such as local elections and politics, local churches, homelessness, small business news, sports, and human-interest pieces.[7] NLP and other ML techniques can even be used to monitor and identify developments in the so-called *alt news* in order to investigate further what may contain grains of truth. In this way, mainstream media can reach out to news outsiders who might otherwise rely solely on alternative media sources. Whether the sources are mainstream or from the fringe, a risk is that seemingly trustworthy sources can be created or taken over by actors with bad intents. Hendrickx (2022) points to *dark participation* in the news, or to "different forms of deviant user engagement originating from malevolent actors", such as "hate speech, disinformation, and strategic attempts to influence public opinion" (Wintterlin et al. 2020). Hence, improved approaches are needed to debunk mis-information early (Thorne & Vlachos 2018), as we will discuss later.

On-demand gathering: AI can also be used to augment or fully automate information gathering on demand. For example, increasingly sophisticated search and recommender tools can use AI to take the journalists' and editors' backgrounds, competences, interests, and current work contexts into account (Bennett et al. 2012) and to encourage creativity (Maiden et al. 2018). The next section will even discuss information gathering in *real time* supported by AI-backed assessment techniques. Trustworthiness can be improved by tools that analyse social networks and that connect the right people inside a distributed and possibly global news organisation, e.g., to ensure that each news story is backed by a team with complementary competences and to avoid duplicate or even inconsistent reports about the same event. Trustworthiness is thereby increased through creative reporting by the most well-informed and interested journalists. The challenge of composing optimal teams of journalists is particularly evident in cross-organisational collaborative journalism, such as in the Panama papers investigation (Zhuhadar & Ciampa 2021), where journalists and editors that work under time pressure and in different locations worldwide must self-organise with little knowledge about one another in advance. Awareness of colleagues who work on

[6] Reg Chua, Reuters <>, accessed 2022-11.

[7] <https://help.crowdtangle.com/en/articles/4474020-the-atlanta-voice-s-strategy-for-using-crowdtangle-to-report-on-the-black-community>, accessed 2022-11.

related stories is already supported by current best-of-breed tools, but can be extended to identify similarities and differences in background and competency and in angles and perspectives on the unfolding story (Motta et al. 2020, Opdahl & Tessem 2021). Risks include journalists who *game* their profiles and networks to work on particular types of stories.

AI-backed tools can also collect, analyse and profile public information about available domain experts in order to suggest suitable informants for a story, maintaining up-to-date profiles of interest areas and levels of competency (Kazai et al. 2016). Tools can be used to analyse social media, possibly augmented with common-sense knowledge bases, to identify other relevant informants. As a result, trustworthiness can be increased through reporting that relies on a broader range of competent sources that reflect more diverse backgrounds and perspectives. Another benefit of diverse sources is that it reduces news organisations' dependency of global platform companies for gathering and analysing data (Simon 2022). A danger is that automated systems will repeatedly recommend the same trusted informants, reducing diversity and excluding new voices and less-connected informants from the news. Another danger is mischievous actors (Wintterlin et al. 2020) that establish trustworthy-looking digital facades, perhaps even forming fake networks of social-media accounts bolstered by botnets.

Process automation: The benefits of routine and on-demand information gathering for trustworthiness can be amplified by enhancing existing process automation platforms with AI. For example, journalists' and editors' information-seeking behaviours (Bennett et al. 2012) can be analysed using *process mining* to augment and automate information gathering and analysis routines: if a journalist repeatedly searches a particular online community forum to corroborate and enrich football match reviews, event logs can be used to *mine* a process description from this pattern (Van Der Aalst 2012). The mined process description can be used by rule- or ML-based *robotic process automation (RPA)* tools that mimic repetitive human tasks such as data entry, form filling, and data validation, e.g., so that the football community forum is searched automatically whenever the journalist starts working on a new review. Existing *workflow and business process management systems* can even be used to adapt the mined process description to similar contexts, improve the user experience and enable intelligent decision making, for example gathering background information for the journalist on-demand and in real time during interviews. AI-enhanced *low-code platforms* provide visual user interfaces that allow journalists themselves to orchestrate new (semi-)automatic processes by combining pre-packaged components for information gathering, NLP, prediction and other ML-augmented tasks. A risk is that the wrong processes are automated: for simpler processes, rule-based robotic process automation may be sufficient or even superior to more sophisticated AI-supported process improvement techniques. Indeed, learning which processes to automate and which to augment (Google 2021) is in itself a hard problem that can be investigated using process analytics.

Data integration: One key to achieving trustworthiness through corroboration is data integration. To prepare for corroboration (or triangulation, Bryman 2016), which we will discuss in the next section, ML-enabled agents and tools are needed to *integrate* and *interoperate* content and metadata formats at both the syntactic and semantic levels (Troncy 2008, Dong et al. 2018). Structured information is available in a wide range of formats such as tables, hierarchies, graphs, time series, and geo-tagged data. The W3C's Resource

Description Framework (RDF) offers a standard for representing both data and metadata as knowledge graphs, whereas the Web Ontology Language (OWL) and related techniques facilitate semantic integration (Hendler et al. 2020, Opdahl et al. 2022). An early example is Troncy's (2008) use of an OWL ontology to make the different metadata formats in a multimedia news production chain interoperable and enrichable using NLP techniques and knowledge from the semantic web. The BBC also adopted linked data early for integrating and adding value to news value chains in a non-disruptive way (Pellegrini 2012). Dong et al. (2018) describe Amazon's ML-fueled efforts to build "an authoritative knowledge graph for all products in the world" as it relates to everything available on their site. The ambition is to make people "[not] just come to Amazon to buy products [but] to see what's new or interesting".[8] A danger is that, while syntactic and increasingly also semantic integration become automated in a reliable way, new problems will become evident with handling the pragmatic and social aspects of meaning, which are harder to automate. For example, consumer and lifestyle stories in newspapers contain normative advice that must not be mixed with declarative facts related to the same entities.

Provenance: Trustworthiness concerns not only the potentially news-relevant *content* itself, but also its *metadata* (Pomerantz 2015, Gartner 2016). To facilitate AI solutions that can assess the trustworthiness of facts and opinions, their *sources* must be recorded along with the facts and opinions themselves. Managing meta information about sources thereby becomes as important as managing content. This includes managing the informants' contribution histories, connections, and other characteristics, such as psycholinguistic patterns (Giachanou et al. 2022). A side benefit is that trust can be increased by making the provenance information available to audiences on demand, as we will discuss later. Another benefit is that the independence of corroborating sources can be ensured: that they do not just relay the same primary information through different paths. To facilitate this, *secondary sources* must be traced as far as possible back to the primary source, such as an eyewitness account, an original document, or a verified live recording. Some crowdsourced public sources like online encyclopaedia keep traces of all edits made to their content, and some social-media sources like Twitter provide provenance metadata. Identity resolution techniques can be used to connect informants with their social-media accounts (Kitchin 2014). Otherwise, when content has unknown origin, it can sometimes be revealed by similarity searches in archives and across the net. As a last resort, natural-language (NL) inference techniques can be used to identify texts that have been derived from other texts (De Nies et al. 2012). A risk is that AI-based source management systems can be *gamed* by malicious actors that establish trustworthy-looking digital facades, possibly networked and boosted by botnets. Also, the provenance information may be false or tampered with, calling for non-repudiable provenance, e.g., supported by blockchains, and for "verifiable provenance".

Table 1. Information *gathering* challenges and potential AI solutions.

Problem area	AI techniques	AI opportunities	AI risks
Routine	Rule-based systems, event	Select diverse and	Over-reliance of

[8] <https://www.aboutamazon.com/news/innovation-at-amazon/making-search-easier>, accessed 2023-04.

harvesting	detection, NLP, ML techniques	credible sources, corroborate information	established sources, stability of sources, legal challenges
Broader harvesting	Trend monitoring, NLP, ML techniques	Engage new audience segments, reach out to news outsiders	Source takeover, dark participation
On-demand gathering	Semantic search and recommendation, context mining and assessment, profiling, social network analysis	Adapt to journalists and editors needs, connect people within the news organisation	Gamed profiles and networks
Process automation	Process mining and automation; context mining and assessment; case- and rule-based systems; choosing tasks to automate/augment	Optimised information seeking, workflow management	Over-reliance on AI, automating the wrong tasks
Data integration	Knowledge graphs, ontologies, OWL, linked data, NLP, ML techniques	Syntactic and semantic integration of data and metadata	Pragmatic and social aspects
Provenance	NL inference, identity resolution, reasoning over provenance, semantic search	Identifying and managing primary and secondary sources, managing online identities	Fake sources and networks, gamed sources

Assessing

In the digital age, news can be altered - or used out of context - in order to attract attention, to influence behaviour or opinion, and to deceive and mislead. Information assessment involves detecting both intentional mis-information and unintentional errors in text, speech, audio, video, structured data streams from sensors, and structured or unstructured reference data (Seo et al. 2021). This section reviews central problems of trustworthy information assessment, along with related AI techniques, opportunities, and risks. Table 2 presents an overview.

Fact checking: Factually correct information is essential to trustworthy news. Fact checking - or fact verification - is the task of assessing the truthfulness of claims (Thorne & Vlachos 2018). It explores NLP and other AI techniques for: selection of checkworthy claims (Hassan et al. 2017); identification, retrieval and preparation of evidence; and using the gathered evidence to evaluate the claim (Popat et al. 2018, Augenstein et al. 2019, Mishra et al. 2019). The third of these tasks has been investigated most intensively. Research often uses real claims from Politifact, Snopes, FullFact, etc. but relies on pre-selected sources of evidence for verification, often Wikipedia. Relevant AI techniques include information retrieval, text classification, natural-language inference, and question-answering (QA) systems (Minaee et al. 2021). While many cutting-edge solutions have been proposed, tools are needed that make them available for journalists. The tools should support all stages of fact checking in practical settings, be trained on real claims, exploit a broad variety of

evidence types, and rely less on textual surface characteristics and more on fact-level inference from ground truths stored in reference bases. They should also be integrated with AI-supported fact checking sites. However, publicly debunking false claims requires careful editorial judgement: debunking mis-information that is not yet widely spread may just draw additional attention to it (Burel et al. 2021). Another risk is that, because fact-checking tools will become available to malicious actors too, they can be used to vet fake news. Automation might also increase pressure on fact checkers and journalists, giving them insufficient time for necessary manual assessment of claims and their sources. Finally, fact checking is highly context-dependent. In many cases, assessing a claim does not mean labelling it as true, false, or dubious, but establishing in which contexts (in terms of location, time, understanding of central concepts, etc.) the claim holds - and whether that context matches the context in which the claim has been made (more on context below).

Media verification: For journalists, verification of multimedia, in particular of images, is a tedious task that can require hours of using media monitoring tools, (reverse-)searching images, and searching general news. Existing verification tools tend to focus on text and must be extended to support verification of other media types. Multimedia forensics techniques assess whether digital multimedia contents are genuine and authentic through deep image analysis techniques that exploit traces imbedded in the digital content when it is created and processed (Farid 2016, Khan et al. 2023). Related deep image analysis efforts focus on *image provenance* (Caldelli et al. 2017). Verification of non-textual media again runs the danger that automation will increase time pressure on fact checkers and journalists. Despite many advances, it remains hard to assess user-generated content shared on social-media platforms, because images have typically been renamed, recompressed, resized, and even altered (e.g., cropped), which can significantly reduce their accuracy and blur the fine-grained distinctions that current verification tools rely on (Pasquini et al. 2021). As for textual fact-checking, there is a risk that media verification tools will become available to mis-informers too.

Deepfakes: One way to exploit media verification techniques are the recently proposed *generative adversarial networks (GANs* (Goodfellow et al. 2020, Guo et al. 2019) and related techniques which generate synthetic media content that is virtually indistinguishable to humans from genuine content (Viazovetskyi et al. 2020). GANs challenge the trustworthiness of media content through *deepfakes* that bear uncanny resemblance to our real world. For example, Descript[9] is a commercial tool that shifts modalities in multimedia editing: the user edits the transcript of a video or sound file, and the tool generates a convincing-sounding[10] and -looking version of the original content with textual edits integrated. In malign hands the tool can generate convincing fake multimedia news. Similar techniques can potentially be used to make the surface characteristics of fake news texts indiscernible from those of real texts (Hossam et al. 2021). This situation poses a serious threat to newsrooms, and puts researchers who develop content verification tools in a constant race against the counterfeiters. In order to detect GAN-based generated multimedia content, “fire must be fought with fire”, combining deep-image and GAN analyses to develop

[9] All-in-one video & audio editing, as easy as a doc. <https://www.descript.com/>, accessed 2022-12.
[10] Ultra Fast Audio Synthesis with MelGAN <https://www.descript.com/blog/article/ultra-fast-audio-synthesis-with-melgan>, accessed 2022-12.

robust and self-improving detection networks that attackers cannot easily fool (Verdoliva 2020).

Cheapfakes: In contrast to deepfakes, so-called *cheapfakes* abuse media content through simpler techniques, such as distorting images mildly or re-using them unaltered but out of context (Paris & Donovan 2019). For example, a Facebook picture of Pope Francis kissing the hands of Holocaust survivors was purported to show him kissing billionaires David Rockefeller and John Rothschild[11]. Central techniques for detecting cheapfakes include deep image analysis; reverse-image searching; identifying the origin of the content; assessing its source and how it has been changed; and assessing the similarity of the original context to the use context (Bouquet et al. 2003, Aneja et al. 2021). In addition to the threat of adversarial techniques, cheapfake detection depends on pragmatic and social context, which are hard for AI solutions to assess reliably. For example, a stock photo of Russian tanks on parade in Moscow may be a relevant illustration for an article on the Russia-Ukraine conflict, but it is not strictly a picture of the conflict (different time and place). Does it therefore discredit the article it illustrates and, indirectly, its source and distributors?

Cross-modal content verification: When corroborating evidence has been gathered and integrated, consistency between the content units - both in isolation and in combination - must be ensured. Information in different *modalities* can be used for joint verification (Farid 2016, Boididou et al. 2018). Recent advances in multimodal representation learning (Guo et al. 2019) can detect errors and mis-information by enhancing the fact-checking process with visual and other information, e.g., checking the consistency between independent present and past still images, video, audio, text descriptions, data from traffic and weather sensors, map data, etc. which purport to describe the same event. For example, are the stated numbers of participants in a demonstration consistent with traffic data at the time, and are the supporting images consistent with weather data and light conditions? As for cheapfake detection, a challenge is that cross-modal analysis can depend on pragmatic and social context: for example interpreting a humorous illustration or caption as fact or refuting a factual article due to a humorous illustration or caption. New AI-based solutions are therefore needed that combine techniques from multimedia forensics, natural language processing, and knowledge representation and reasoning (Boididou et al. 2018). Assessment must consider not only the primary news content: the associated metadata must be trustworthy too (Pomerantz 2015, Gartner 2016). For example, social media can be used to prove or disprove that a photographer was indeed present at the time and place their picture purports to illustrate. Verified metadata can in turn be used to assess the trustworthiness of primary content. For example, metadata about the photographer and distributor of a picture can be an important indicator of credibility (Flanagin & Metzger 2020). Metadata verification relies on correct and untampered provenance information.

Proactive verification: Taken together, the challenges posed by cheap- and deepfakes suggest that current *reactive* approaches to fact checking and media verification alone may not be enough. *Proactive verification* is an alternative that issues *content certificates* with

[11] “Pope Francis kissing the hands of David Rockefeller & John Rothschild”, PolitiFact <https://www.politifact.com/factchecks/2021/may/19/viral-image/pope-francis-was-photographed-holocaust-survivors-/>, accessed 2021-10.

verifiable, tamper-evident signatures using techniques such as hashing[12] and blockchains (Zheng et al. 2018). Certificates can be issued both for metadata, for atomic content units (Caswell 2019, Jones & Jones 2019), and for composite content. Because the trustworthiness of a content unit is situational, this will require new AI solutions for mining and assessing context. For example, a picture taken of a politician in a white laboratory coat while participating in a TV-game show would be misleading as an illustration to the same politician making a statement about vaccination. Research is also needed to better understand the relationship between context and credibility: how different users will tend to trust different sources in different contexts. In any case, proactive verification risks relying too heavily on professionally provided mainstream information that is easily verified at the point of capture, at the expense of sources that are unusual, outside of the mainstream, and therefore harder to verify.

Contribution chains: Ideally, certification should be performed once and at the point of gathering or creation. This calls for representing and reasoning over provenance in the form of *contribution chains*, which provide *accountability* by recording how content has been derived from other content and by whom (or what). Such chains can help to avoid repeated integration, verification, or other assessment, and they can be used to detect mis-information from informants with vested interests. Contribution chains can be created automatically as part of the media production workflow, but can also be reconstructed using for example reverse content search, NL inference techniques, and deep image analysis. For example, an informant may edit a Wikipedia article in preparation of an interview to dupe the journalist. In such situations, natural-language inference techniques (Minaee et al. 2021) can be used to compare claims made by the informants with recent anonymous edits. PROV-O offers a standard vocabulary for representing and exchanging contribution chains (Lebo et al. 2013), and their integrity can be certified by blockchains (Zheng et al. 2018). Contribution chains can also be used to learn how different types of information and mis-information originate and spread through social media, especially when the messages themselves contain limited content and context to use for verification. Such and other AI uses depend critically on correctness of the provided provenance chain: that it is correctly recorded and not tampered with. The chains must also be properly managed to avoid exposing endangered informants and sensitive information gathering methods.

Context: When the origin of content has been established, the original context needs to be captured and represented as metadata too (Bouquet et al. 2003) as an extension of the contribution chain. For example, the context of online content includes the source web site, its domain, the person or organisation behind it, the known features and past activities of the author, and so on. Models for assessing context similarity for different purposes, types of content, and types of uses are also needed. This invites audience studies and feedback where the consumers' trust in news content is measured and used to learn their tolerance for edited, adapted, or otherwise changed content in different contexts and domains. Audiences nowadays are used to accepting, and even expecting, fake or manipulated content in entertainment and marketing contexts, but they may shun manipulation of news and documentaries. Tolerance for image editing in journalism is low and journalistic codes of ethics typically emphasise that the integrity of the journalistic photograph, along with other

[12] For example the Content Authenticity Initiative <https://contentauthenticity.org/>, accessed 2023-03.

journalistic content, must be protected[13] (Flanagin & Metzger 2020). Although many aspects of context can be accurately and automatically recorded - such as time and location, social and technical agents, etc. - the pragmatic and social context is harder to describe. It is not at all clear whether and how it can be done in a general and future-proof way independently of purpose.

Source credibility: An important type of metadata describes the *sources* of content. Machine learning techniques can be used to train models that profile human informants through measures such as their historic reputation (Ceolin et al. 2012), their social-network connections, their knowledge background, position, sponsoring organisation, and whether they are referred to by other informants (Flanagin & Metzger 2020). For example, post-event historical analyses of Twitter feeds can be used to identify and positively weigh accounts that have consistently reported newsworthy events early and in a trustworthy manner. Such weighing could improve the signal-to-noise ratio in event detection. Background information from open or proprietary databases and archives can also be assessed in terms of the people and organisations that have provided the information and maintained the databases and archives - whether manually or with computer support. The editing histories of crowdsourced data can also be used for assessment. As for information gathering in general, a risk is that source metadata is manipulated by mischievous actors (Wintterlin et al. 2020) that establish trustworthy-looking digital facades, which can even appear in networks bolstered by fake social-media accounts. To counter this problem, AI-driven source assessment for information gathering may have to be designed to "err on the safe side", which in turn induces a risk of over-reliance over the same set of trusted sources. As for automated fact checking, automated source-credibility may weaken the focus on and time available for human assessment of sources.

Retraction: There are many examples of journalists who investigate stories based on information that later turns out to be false, such as misinformed or mischievous social media posts. In the aftermath of the tragic Boston Marathon Bombing, a missing student was mistakenly identified and accused on social media platforms as one of the perpetrators.[14] Some journalists passed the false accusation on to their followers, reporters started to call the student's family, and news vans began to stake out their home.[15] During Hurricane Sandy, CNN erroneously relayed a claim from social media that the trading floor of the New York Stock Exchange had been flooded.[16] To facilitate retraction of mis-informed stories, untrustworthy information must be traceable back to its sources, and an untrustworthy source must be forward-traceable to all the content it has contributed to (Lebo et al. 2013), calling for both backwards and forwards reasoning over provenance. But while erroneously published information can be retracted, its consequences cannot always. Such consequences include both how erroneous news has impacted the non-informational world

[13] For example, §4.10 and §4.11 in the code of ethics for the Norwegian press, <https://presse.no/pfu/etiske-regler/vaer-varsom-plakaten/vvpl-engelsk/>.
[14] New York Times, 2013-04-25. <https://www.nytimes.com/2013/04/26/us/sunil-tripathi-student-at-brown-is-found-dead.html>, accessed 2023-04.
[15] New York Times, 2013-06-25. <https://www.nytimes.com/2013/07/28/magazine/should-reddit-be-blamed-for-the-spreading-of-a-smear.html>, accessed 2023-04.
[16] Washington Post, 2012-10-30. <https://www.washingtonpost.com/blogs/erik-wemple/post/hurricane-sandy-nyse-not-flooded/2012/10/30/37532512-223d-11e2-ac85-e669876c6a24_blog.html>, accessed 2023-04.

and how erroneous information has fed downstream ML models. In any case, untrustworthy sources must be clearly marked for the future and used to learn social and behavioural patterns that can reveal similar untrustworthy sources and disinformation spreaders in the future (Del Vicario et al. 2016, Vosoughi et al. 2018). Whenever untrustworthy content is detected, the characteristics of both the content, its sources, and dissemination patterns must therefore be recorded and used to improve content assessments, including fake news detectors.

Transparency: For assessments to be trustworthy, the results, reasoning, and evidence must be accessible to the audience (Plaisance 2014). New solutions are therefore needed to ensure that audiences can understand the outputs of AI methods. For example, the classifications made by automated fact checks must be explained on demand (Kotonya & Toni 2020). Explainable AI (XAI) is a subfield of artificial intelligence that focuses on exposing complex AI models to humans in a systematic and interpretable manner, explaining AI decisions through, e.g., transparency, interpretation methods, and natural-language (NL) generation techniques (Samek 2019). This poses challenges when decisions and other assessments result from combining multiple AI-based methods and models. A danger is that explanations made by AI- and ML-solutions are hard to interpret by the layman, contributing to a widening gap between the more and the less *information competent*. Explanations of decisions behind news content should also adapt to how the audience perceives truthfulness and trustworthiness. Another risk is the accuracy of the explanations themselves, which too may be established using learning algorithms, while ensuring that explanations are optimised for correctness and not for believability.

Real-time assessment: In order to fuse gathering and assessment tasks efficiently, journalists and editors would benefit from tools that can suggest background information and verify content and sources instantly (Liu et al. 2017). During interviews, the veracity of the claims made and information provided could be assessed in real-time. Information retrieval and NL inference techniques could be used to suggest appropriate background information and follow-up questions (Minaee et al. 2021). Reuters' News Tracer (Liu et al. 2017) is an example of a tool that tracks social media in real time. Meta's CrowdTangle tool helps journalists and others follow, analyse, and report what's happening across Facebook, Instagram, Reddit, and Twitter.[17] A danger is that automated information streams in real time will leave less time for critical reflection and commentary, describing events as they unfold with little reflection over their meanings and consequences. Manipulative content and other types of mis-information may be spread rapidly and uncritically, perhaps destabilising the already critical situations they purport to report further. For example, mis-information about right-wing groups trying to disrupt a peaceful protest might encourage actual right-wing activists to seek out the event.

Table 2. Information *assessment* challenges and potential AI solutions.

Problem area	AI techniques	AI opportunities	AI risks

[17] <https://www.crowdtangle.com/>, accessed 2022-11.

Fact checking	NLP, text classification, information retrieval, NL inference, question answering, context mining and assessment	Assessing checkworthiness, finding evidence, claim evaluation	Giving false claims more attention, over-reliance on automation/less time for manual assessment, sensitivity to context (pragmatics), adversarial use
Media verification	Multimedia forensics, reverse image search, deep image analysis (CNN-based for image quality and camera noise analysis), representing and reasoning over provenance	Extend verification tools to multimedia, multimedia forensics, image provenance	Over-reliance on automation/less time for manual assessment, low-quality social media images, adversarial use
Deepfakes	Deep image analysis, generative adversarial networks (GANs)	Detect deepfakes	Fake multimedia news, adversarial use, racing the counterfeiters
Cheapfakes	Reverse image search, deep image analysis, context mining and comparison	Detect cheapfakes	Fake multimedia news, adversarial use, pragmatic and social context
Cross-modal content verification	Multimodal representation learning, multimedia forensics, context mining and assessment	Ensure consistency between text and other media, detect mis-information, verifying metadata, using metadata for verification	Pragmatic and social aspects, manipulated contribution chains
Proactive verification	Context mining and assessment	Content certificates, understanding trustworthiness in context	Pragmatic and social context, over-reliance on mainstream sources
Contribution chains	Representing and reasoning over provenance, reverse content search, NL inference, deep image analysis	Maintaining contribution chains, understanding how mis-information originate and spread	Manipulated contribution chains, exposing vulnerable informants
Context	Context mining and similarity assessment, learning from trustworthiness measures	Managing the original context of content, understanding trustworthiness in context	Pragmatic and social aspects
Source credibility	Social network analysis, profiling, ML	Managing informants, verifying crowdsourced information	Fake sources and networks, over-reliance on trusted sources, less time for manual assessment
Retraction	Backwards and forwards reasoning over provenance	Rectifying consequences of mis-information, understanding how mis-information originate	Unrectifiable consequences of information that has been used in further analysis and

		and spread	training
Transparency	Explainable AI, NL generation	Explaining assessments to the audience, understanding perceptions of trustworthiness	Explanations for the few, optimising for believability over correctness
Real-time assessment	Fact-checking techniques, information retrieval, NL inference, question answering	Suggesting background information, content verification, follow-up questions	Less time for reflection, accelerated spread of mis-information

Creating

Creating and presenting news consent are interconnected tasks. Borrowing from literary theory and computational creativity (Gervas et al. 2009), we separate the *fabula*, i.e., the selection and combination of content, from its *discourse*, which is how the content is presented (or narrated). This section reviews central problem areas in creating the fabula, while the next section discusses its discourse. For each problem area, related AI techniques and potential AI opportunities and risks are discussed. Table 3 presents an overview.

Robot journalism: Robot journalism typically employs rule-based approaches to NL generation based on structured data from highly trusted sources, such as public databases (Leppänen et al. 2017). For example, the Heliograf tool broadens news coverage by automatically generating short reports for The Washington Post's live blog. First used to report results during the Rio Olympics, it has since expanded its reach to report on subjects like congressional races and high-school football games.[18] Although the input data are not always structured, keywords and standard sentence structures are used to extract key data. This type of automatic content creation is trustworthy because it follows explicit rules and is directly based on facts, which are sometimes even publicly available and that can be considered transparent and unbiassed - at least to the extent the underlying sources are unbiassed. A danger is gradually dehumanising the journalistic profession through *automation creep*: even when the intention of introducing AI is to augment rather than to automate, business and market considerations may create increasing automation pressure. Another danger is that robots drive journalism towards stereotypical, automated news reports.

Augmented journalism: Beyond rule-based robot journalism, journalists and editors must continue to be essential both to shape the narrative of news content and to ensure its overall quality and trustworthiness. In *augmented journalism* (Marconi et al. 2017), the journalist takes responsibility for the final product, but uses partially automated AI-supported workflows to identify content and sources; to ensure they are trustworthy; to combine them in a trustworthy manner; and to find a news angle and background information that is relevant to the context in hand (Motta et al. 2020, Opdahl & Tessem 2021). A risk is again that augmented news production pipelines initiate an irreversible process, driven by business concerns, towards increasingly automated news, in which journalists are gradually turned into high-level overseers and maintainers of journalistic information flows, increasingly

[18] <https://www.digitaltrends.com/business/washington-post-robot-reporter-heliograf/>, accessed 2022-11.

detached from essential human capabilities such as handling sources, their credibility and diversity, and providing balanced and nuanced perspectives on unfolding events.

Content units: Augmented journalism can benefit from AI to connect the available and verified sources of information with one another and with pre-existing content units. We envision that new content can be created by recombining small, certified, and self-contained semantic units into composite narratives (Caswell 2019), for example using case-based reasoning (Aamodt & Plaza 1994). Examples of such units are short texts, links, images, video clips, animations and other visualisations. This idea is similar to *structured journalism*, in which content is composed from *units* or *atoms* expressed as data (Jones and Jones 2019). Journalistic content composed of reusable atomic units can only be as trustworthy as the descriptions of these units, emphasising again the challenge and risk of pragmatically and socially sensitive context descriptions. To support trustworthy content composition, the atomic units must be described by rich and precise metadata. For example, each unit must be labelled with its related topics to make it retrievable, with its sources to ensure it is trustworthy, and with its original context to ensure that it is not reused in an inappropriate way. Hence, automatic extraction of rich, precise and multi-faceted metadata becomes essential for trustworthy content composition.

Trustworthy composition: Composing narratives from self-contained units does not only require trustworthy content units and metadata, but also demands that the narrative organises the units in a reliable, credible, and explainable way. The granularity of content units, where they come from, how they have been extracted, how well they fit the context, and other metadata can be used to optimise internal cohesion and perceived trustworthiness. In a composition, content units can be *repurposed* for other uses than the original intention, for example by paraphrasing (Burrows, Potthast & Stein, 2013). Similarly to moving content across contexts, inappropriate repurposing can damage trustworthiness, as when pictures of a violent protest are used to illustrate another protest that is peaceful. On the other hand, repurposing can also increase overall trustworthiness, for example by using pictures collected for traffic reporting to show that the protest was indeed peaceful. Learning materials and training goals are another concern. If learning relies too strongly on audience preferences, it is a risk that automated news composition becomes biassed towards producing believable, compelling, and engaging news at the expense of factually precise and ethical news: atomic news systems must be trained to be ethical rather than clever.

Suggesting perspectives: Going beyond existing content units, intelligent tools can increase trustworthiness by providing alternative viewpoints on the same event from sources with different political or other orientations (Resnick et al., 2013). For example, Oh et al. (2009) attempt to distinguish between liberal and conservative viewpoints, while the NewsCube browser (Park et al., 2010) goes further by partitioning the space of articles about an event according to different viewpoints, called *aspects*. Combining information gathered from independent sources with complementary profiles can increase trustworthiness by ensuring that alternative perspectives on an event or situation are covered by a story. NLP techniques can be used to identify sources and related content that describe the same situation, but that emphasise different facts, present different explanations, and express different sentiments about it (Trabelsi & Zaïane 2015). In this way, even weakly backed or known false positions can be identified and potentially mentioned in the news report, as long as they are clearly

presented as dubious or false, along with grounded counter evidence or arguments (Thorne & Vlachos 2018, Kotonya & Toni 2020). However, all perspectives are not equally valuable. Balanced journalism does not mean that belief-based or malignant perspectives must be granted as much attention as factually grounded and ethically defensible ones.

Live reporting: There can be many sources of information that need to be fact-checked and validated to ensure they have not been manipulated. In a breaking news situation, this must happen in real time (Liu et al. 2017). Trustworthy live reporting is closely connected to on-demand information gathering, fact checking, and real-time assessment. It supports trustworthy news content by enabling frequent and immediate updates of stories about unfolding events, showing everyone that the content is up to date (Marconi 2020). It emphasises the need for interactivity, speed, and online availability and for suggesting informants, interview questions, and open paths to pursue during an interview. Natural-language inference techniques (Minaee et al. 2021) and automatic angle detection (Motta et al. 2020) can potentially be used to direct interviews in real time. Again, it is a risk that, with less time for human reflection, real-time news reporting can spread mis- or dis-information rapidly. Mis-information - whether live or not - can even create self-fulfilling prophecies, for example when a system detects frequent mal-practice in a particular hospital - a possible statistical fluke - leading to increased attention by patients of that hospital.

Iterative journalism: Taking real-time journalism one step further, *iterative journalism* (Marconi 2020) emphasises frequent story updates, while taking into account not only how the news event itself unfolds, but also how it is reported and how the audience's reactions, opinions, interests, tips, and information needs evolve. The audience can be profiled directly through explicit ratings or indirectly through, e.g., counts of shares and reposts. Such quantitative measures can be supplemented with qualitative data, e.g., from NLP analyses of users' textual comments. A danger is journalism that too reactively follows the crowd, producing populist news that leaves its critical watchdog role behind. Another challenge is monitoring news consumers and their interactions through social media. German newspaper Der Spiegel uses *conversario* to moderate user comments both onsite and on its social media accounts, managing well over 2 million comments per month. For European broadcaster RTL, "the software makes our work in social media management easier [...] through the comment categorisation of the conversario AI [...] around the clock [...] reliably and with a low error rate."[19]

News discovery: Journalists and editors can even be assisted in *detecting* newsworthy events, as demonstrated by Reuters' Tracer tool (Liu et al. 2017). Trustworthiness in news can be expected to increase when unfolding news events are quickly detected and continuously updated. In addition to supporting the detection of new events, pattern-matching mechanisms can be used to identify new information that can shed new light or suggest unexpected angles (Motta et al. 2020, Opdahl & Tessem 2021) on unfolding events. The sheer amount of sources explored, interpreted and verified with intelligent algorithms can potentially contribute to making the content of a developing story more trustworthy. Similarly to automated information gathering, automatic news discovery runs the risk of relying on a too-limited range of set sources, ignoring events that originate in the fringe. Different time scales mean that rapidly evolving events may be easier to detect than slowly

[19] Frank Kohls, RTL <https://www.conversar.io/en/product>

developing ones that are perhaps more important. Finally, automatic news detection runs the risk of promoting the most unusual - or surprising - rather than the most impactful and in that sense important events.

Table 3. Content *creation* challenges and potential AI solutions.

Problem area	AI techniques	AI opportunities	AI risks
Robot journalism	Rule-based reasoning, NL generation	Utilise structured data from public sources, broader news coverage	Dehumanising journalism, automation creep, stereotypical news
Augmented journalism	AI-supported workflows for, coordinated use of other AI techniques	AI-backed journalistic workflow support	Automation creep, dehumanising journalism
Content units	Context representation and assessment, case-based and other reasoning	Composing content from verified content units, structured journalism	Inaccurate metadata, in particular about context
Trustworthy composition	Explainable AI, context representation and comparison	Trustworthy and explainable composition, repurposing content units	Inappropriate repurposing, understanding trustworthiness in context, optimising for believability over correctness
Suggesting perspectives	Information retrieval, profiling, NLP (e.g., stance and sentiment analysis)	Present multiple perspectives on an event, seek complementary sources,	Dealing with ungrounded and malignant positions
Live reporting	Fact-checking techniques, NL inference, question answering	Frequent story updates, suggest informants and follow-up questions	Less time for human reflection, accelerated spread of mis-information, self-fulfilling prophecies
Iterative journalism	Social-media monitoring, profiling, short-text NLP (of social-media messages)	Frequent story updates, understand and react to audience responses	Uncritical, populist journalism; managing social media
News discovery	Event detection, news-angle mining, pattern matching	Detecting new events, detecting new information about unfolding events, suggesting news angles	Over-reliance on trusted sources; ignoring the fringe; focus on the rapid over the slow, the unusual over the usual

Presenting

Selection and combination of content units constitute the *fabula*, or what is told, whereas how the fabula is *presented* (or narrated) to the intended audience is referred to as the *discourse* (Gervas et al. 2009). This section reviews problem areas related to news

presentation and discusses related AI techniques, opportunities, and risks. Table 4 presents an overview.

Narrative generation: An AI system can suggest a specific narrative to a journalist or directly create a textual presentation. Recent developments in deep learning (Goodfellow et al. 2016) include large generative, transformer-based *large language models* (LLMs) like OpenAI's[20] *general-purpose transformers*, to which we will return in the conclusion. GPT-3 (Radford et al. 2018, Brown et al. 2020) wrote a published newspaper essay already in 2020.[21] Its successors, such as ChatGPT (Ouyang et al. 2022), may radically transform how media is produced. Other recent breakthroughs include large generative models for voice-to-text[22], text-to-voice[23], text-to-image[24], and text-to-video[25] translation. The most recent GPT-4 (OpenAI 2023) model from OpenAI is even *multi-modal* in its ability to input images as well as text and to output both images, text, program code, and a wide range of other formats.

Ferreira et al. (2019) present an early exploration of encoder-decoder mechanisms (Sutskever et al. 2014) in the news domain, translating structured facts represented as RDF triples into natural language news reports. Generative Adversarial Networks (GANs) are also being explored for generating texts (Hossam et al. 2021) in addition to images (Goodfellow et al. 2020). However, there are at least four current challenges with using LLMs for text generation. The first one is their lack of *explainability* (Ras et al. 2022, Samek 2019). The second is their tendency to *hallucinate* (Rohrbach et al. 2018), or to generate plausible-sounding nonsense (Thorp 2023): texts that contain elements not found in the input data (Ferreira et al. 2019). The third is their *bias* or ideas about the world learned from their training data, such as the superiority of particular cultures (Stokel-Walker and Van Noorden 2023), which might make already marginalised groups experience further misrepresentation in the news (Hutson 2021). The fourth is their *toxicity*, or tendency to reproduce racist, sexist, hateful, or otherwise problematic language use in the text corpora they are trained on (Stokel-Walker and Van Noorden 2023). These issues need to be solved before large language models can be effectively used in trustworthy news production. The supervised fine tuning (SFT) approach used to align ChatGPT's responses better with human expectations has been reported to reduce toxicity somewhat, but with little effect on bias (Ouyang 2022). The auto-regressive nature of the large-language models can also lead to a proliferation of formulaic, stereotypical news reports that are able to replicate well-established story formulae infused with new facts, but unable to create new types of stories that transcend established ways of reporting.

Template-, script-, and rule-based approaches to text generation provide alternatives to large language models, but they are not yet capable of generating texts with advanced journalistic narrative structures. An alternative approach is to use case-based narrative generation (Hervas & Gervas 2006) which replicate how past narratives have organised their content units. Techniques from case-based planning (Borrajo et al. 2015) can be used to

[20] OpenAI <https://openai.com/about/>, accessed 2022-11.
[21] The Guardian, 2020-09-08. <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>, accessed 2023-04.
[22] For example Whisper <https://openai.com/research/whisper>, accessed 2023-03.
[23] For example, Speechify <https://speechify.com/>, accessed 2023-03.
[24] For example DALL-E <https://labs.openai.com/>, accessed 2022-12.
[25] For example InVideo <https://invideo.io/>, accessed 2023-04.

select previous narration cases and to combine, adapt, and modify them as new content units arrive. As already explained, the inferred narrative can also include existing, certified, and self-contained semantic units (Caswell 2019). The journalist can take the proposed narrative structure, perhaps after reorganising it, as a starting point for generating or writing the final text

A case-based approach has several advantages with respect to trustworthy news presentation. The use of verified content units and the application of an already tested narrative structure, perhaps created by journalists, ensures coherence and validity. As with all narratives, the success of the generated structure can be gauged by analysing positive and negative feedback. Feedback helps to indicate how good a narrative structure is, perhaps even identifying which parts work best in a final presentation. Finally, in contrast to deep learning approaches, case-based narrative generation is easily explainable, as it is possible to trace both the origin of content units, the annotations and summaries created and used, and the narrative structure (or how the narrative organises its content units).

Contextual presentation: Media consumption has moved from one-dimensional linear content streams (such as linear TV, static HTML) onto multiple platforms that are capable of adaptation and interactivity (such as phones, smart speakers, smartwatches, tablets, etc). Journalists therefore need tools for creating and preparing flexible content presentations according to their users' profiles, current needs, and context (Zorilla et al, 2015) and on the most suitable platform. Generative multimodal representation models can be used to create *transmedial* narratives that can be presented across several platforms of different types. This development can be harnessed for trustworthiness. For example, while presenting a story on a TV screen, trustworthiness can be underlined by making deeper information, such as background facts, related social-media content, examples, links, and other information, available through the viewer's mobile phone at the same time. A prerequisite for developing AI tools that tell stories multimodally and transmedially in trustworthy ways is to understand which types of media that are perceived by users as trustworthy for which purposes. A danger of contextual presentation is that the news can become *too adapted* to the consumers' current situation, enclosing them in information bubbles that reflect their existing interests and values at the expense of providing challenging and potentially eye-opening information.

Trustworthiness reports: The assessment phase generates secondary content in the form of justification or grounding of primary content. Explainable AI and related techniques can be used to automatically generate trustworthiness reports from this provenance information. Such reports provide *accountability* by presenting the origin of content units and explaining how and why they have been combined. They can also explain how the credibility (Flanagin & Metzger 2020) of a particular source on a particular topic has been assessed. This will both document trustworthiness, make the assessment available for criticism, and make the position and values behind the story explicit. For example, when applying case-based reasoning and adopting a narrative structure from an earlier report, one could explain how the structure matches the event and what rules are used to adopt the content to the structure. This is in line with current research on explainable case-based reasoning (Schoenborn et al. 2021). As for more elaborate chains of evidence, content should ideally be traceable all the way from the original sources to the claims made in the final report. Trustworthiness reports can be made available to news consumers on demand. When the

information has been corroborated by many independent sources and analyses, personalised verification can ground claims selectively in the types of sources and analyses that each user prefers, trusts, and understands best according to their profiles. The danger is an over-reliance on automatic assessment, letting news consumers rely too much on compelling explanations presented in appealing contexts, rather than on their own critical sense. Also, a trustworthiness report can only be as reliable as the provenance of the units it composes, of their sources, and of their context descriptions.

Actionability: Trustworthiness reports and their underlying evidence chains can even be made available in a computer-ready format, using the PROV-O standard for provenance information (Lebo et al. 2013), The reports thus become available for community fact checking. In this way, the news organisation demonstrates willingness to be audited by its audience and the general public, an important aspect of trustworthiness (Jones 2012). News users may for example be given links to sensors they can use to double check the data behind claims and to trusted sources that explain how conclusions are reached. A risk is that editorial resources become tied up defending attacks from well-resourced mischievous actors that challenge the veracity of the published news using untrustworthy information sources and dishonest argument styles.

News outsiders: The democratisation perspective on journalism emphasises making quality information available to all (Ward 2019). As explained in the introduction, reasons for avoiding mainstream news include media habits, distrust in authorities, and limited skills. To gain the trust of news outsiders - and to demonstrate social responsibility to the general audience - trustworthy journalism should not present content that can be considered sexist, racist, or attacking vulnerable groups (Ward 2019). Presentations must accommodate multiple viewpoints and follow narratives that are explicit about the positions, sentiments, and values they embed. For individuals with attention deficits and recent migrants with limited local language skills, summarisation and automatic translation are valuable techniques. An obvious ethical concern is models that are biassed due to unbalanced or tendentious training materials. Also, as before, not all alternative perspectives - such as ungrounded and malignant ones - deserve equal attention.

Privacy: Privacy is important to gain and maintain the trust of vulnerable and exposed informants (Jones 2012, Strömbäck et al. 2020). A challenge is to design tools that can detect personal and sensitive information and assess whether privacy and personal protection is maintained in a news presentation. This is not an easy problem to solve, because the journalist constantly has to assess who may need privacy or can demand it, and whether it is actually in the interest of society to be told about particular behaviours of a person. Techniques such as automatic face-blurring[26] can be used to protect the identity of participants in a political rally, but may decrease trustworthiness because the picture or video becomes easier to fake and harder to verify.

Monitoring reception: Monitoring and otherwise collecting information about how the presented content is perceived (Lee et al. 2020) will provide feedback to the computerised processes at work. In particular, perceived trustworthiness can be assessed through audience studies, user ratings, NLP analysis of commentary fields, and by monitoring news

[26] E.g., <https://www.theverge.com/2020/6/4/21280112/signal-face-blurring-tool-ios-android-update>.

sharing behaviours. These and other measures can be used to gauge both the trustworthiness of news items overall and of their individual units and aspects, such as content, background materials, sources, and presentation. Trustworthiness measures can then be fed back to all stages of the production process (Figure 1), optimising tasks that include source selection and balancing, presentation of tracing information, fabula generation, and the final presentation. A danger is that certain audience groups may influence the news disproportionately through their feedback. Monitoring can easily come to optimise news presentation for existing news consumers at the expense of news outsiders. Hendrickx (2022) warns about "taking an 'audience turn' without actually knowing how, if at all, audiences actually think, act and feel towards news and other forms of media content." Indeed, understanding audience reception is key not only to understand their trust, but also to avoid that AI- and ML-augmented news production only repeats past mistakes more effectively and efficiently, further alienating news outsiders and parts of the audience that are critical to journalism. A final risk is that malignant actors may exploit monitoring to influence news production to their advantage.

Table 4. *Presentation* challenges and potential AI solutions.

Problem area	AI techniques	AI opportunities	AI risks
Narrative generation	Large language models, NL generation, rule- and case-based reasoning, explainable AI	Create and explain narratives	Lack of explainability, hallucination, bias, toxicity, stereotypical/formulaic news reports
Contextual presentation	Profiling, representing and reasoning over context, multimodal representation learning	Multi-platform presentation, multi modality, transmediality, understanding trustworthiness of platforms	Over-adaptation to users, information bubbles
Trustworthiness reports	Representing and reasoning over provenance, explainable AI, case-based reasoning, profiling	Explain trustworthiness on demand, understand perception of trustworthiness	Over-reliance on automatic assessment, less focus on critical sense, optimising for believability over correctness
Actionability	Provenance representation	Computer-readable trustworthiness reports, audience verification	Defending against unfair trustworthiness attacks
News outsiders	Sentiment analysis, translation, summarisation	Represent news outsiders fairly, reach out to news outsiders	Bias, dealing with ungrounded and malignant positions
Privacy	Classification (of personally identifiable and sensitive information), automatic face blurring	Preserving the privacy of informants and others	Balancing societal needs and individual rights
Monitoring reception	NLP (of user feedback)	Understanding and representing how news is received	Optimising for particular audiences, not understanding audiences, repeating past mistakes more effectively and efficiently, optimising for the mainstream, gaming

Conclusions

In this paper we have discussed how AI can support quality journalism, a key pillar of a democratic society. Our vision is motivated by the need for quality journalistic outlets to counterbalance disinformation and mitigate polarisation, thus contributing to a progressive view of society and democracy. Although AI techniques have been exploited by rogue political actors, as in the Cambridge Analytica scandal, it is also essential that they are harnessed to support quality journalism and, more in general, to help tackle the major

societal challenges of our age. However, it goes without saying that AI (and technology, in general) is not a panacea. AI can support quality news production but other factors, such as sound business models, independence, lack of political interference, and strong ethical values must be in place to ensure that quality media can thrive. In other words, while we believe that the adoption of AI techniques is not necessarily sufficient to tackle the current issues in the media landscape, we also believe that no quality news outlet is likely to survive in the 21st century unless it is able to take full advantage of the opportunities created by AI and its related technologies.

Given the economic pressures on the modern newsroom, a central aim of introducing AI to support journalism is to cut costs and improve efficiency. For example, the JournalismAI[1] report (Beckett 2019), which is based on a survey of 71 news organisations in 32 different countries, emphasises efficiency as the key driver for introducing AI. Indeed, most of the challenges we have identified focus on *making journalists more efficient*, for example by automatically collecting and preparing background information; suggesting sources and informants; freeing journalists from tedious verification and fact-checking tasks; and assisting central creation and presentation tasks. On the *newsroom* level, we have discussed improving efficiency further by using AI to identify journalists that work on the same or related stories and to form teams with complementary backgrounds and competencies. However, *sound business models and editorial independence* are essential to ensure that these new opportunities for journalistic and organisational efficiency are used to *provide higher quality journalism*, and not misused to reduce the journalistic workforce.

We therefore believe that the central aim for AI-augmented journalism is to relieve journalists of their most tedious tasks in order to free up time for creativity and critical reflection - not only to increase trustworthiness, but in broad pursuit of high-quality journalism. For example, the Finnish Broadcasting Company (Yle) uses Voitto, an automated data-based journalistic bot, to write schematic articles in Finnish and Swedish about ice hockey results and statistics: “bots can save journalists' time, allowing them to use more time for considered journalism while bots take care of the mechanised tasks”.[27] Powerful large language models like OpenAI’s GPT-4 are likely to expand the boundary of robot journalism further, enabling generation of a broader variety of stories with more complex narratives on a wider range of subjects. GPT-4 and similar models from tech-giants such as Google and Meta will also push the limits of how stories are presented to users, generating dashboards, storyboards and other visualisations in addition to texts; presenting news in different languages and vernaculars on demand; and enabling interactive news through chatbots and interactive storytelling. Nevertheless, the challenges mentioned earlier - explainability, hallucination, bias, and toxicity - may not become sufficiently manageable in the foreseeable future. Additional challenges that must be overcome include the ownership of training materials and model outputs (van Dis et al. 2023); provenance and reliable accreditation of sources (Stokel-Walker and Van Noorden 2023); responsibility for the content and integrity of the final product (Stokel-Walker 2023); concerns about ecological footprint (Stokel-Walker and Van Noorden 2023); and public trust in large language models.

We therefore believe that complete automation of news production will remain unrealistic and unwanted outside of restricted journalistic enclaves, in which the reliance on AI will need

[27] Jukka Niva, Yle Labs <https://yle.fi/a/3-10126261>, accessed 2022-11.

to be flagged. Notably, Yle has begun to self-label its robot articles as “made by Voitto” to separate them from human-in-the-loop content. This separation will become increasingly blurred as AI becomes embedded in standard writing, searching, and reading tools (van Dis et al. 2023). But, although the work practices of journalists will change, they will still hold responsibility for journalistic processes and products that abide by journalism’s ethical standards. To proliferate in the AI age, journalists and their editors must therefore be encouraged to learn new AI-powered tools and to apply them in their daily work. Because the tools and services are used for different purposes, at different levels, and on different types and sources of information, they must facilitate collaboration, not only among journalists, but also between journalists, editors, external fact-checkers, and the general audience. Transforming newsrooms digitally through new AI-based tools and practices also requires institutional work. Established ways of working must be revealed and revised. The media organisation must also manage its legitimacy to both insiders and outsiders, for example by exposing its adoption of "best AI practices" (Bitektine and Haack 2015, Harmon et al. 2019).

A final challenge in building advanced AI platforms for journalism is that tools designed for good can be exploited by malevolent actors, such as rogue media sources, political groups, and governments. We would argue that our vision for *rigorous evidence-based journalism that is based on robust democratic values and that is open about its positions and values* makes malicious exploitation harder in two ways: When malicious content is explicitly grounded in evidence that is untrustworthy it should become easier to identify and expose. And when malicious content is not explicitly grounded, this in itself should raise suspicion.

In future work, we plan to investigate these and related challenges, opportunities, and risks in the *MediaFutures: Research Centre for Responsible Media Technology & Innovation*[28] , which is hosted by the University of Bergen in Norway where it is co-located with major media houses and technology providers. MediaFuture’s main objective is to develop new responsible AI-based media technology, for better audience understanding and for effective media user engagement, content production, interaction, and accessibility.

Acknowledgement

This work was supported by the industry partners and the Research Council of Norway with funding to *MediaFutures: Research Centre for Responsible Media Technology and Innovation* through the centres for Research-based Innovation scheme (project number 309339) and by the *News Angler* project funded by the Research Council of Norway (grant number 275872). The authors would like to thank our anonymous reviewers deeply for their highly useful insights and suggestions.

[28] MediaFutures - Research Centre for Responsible Media Technology & Innovation <https://mediafutures.no/>, accessed 2021-10.

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, *7*(1), 39-59.

Aneja, S., Midoglu, C., Dang-Nguyen, D. T., Riegler, M. A., Halvorsen, P., Niessner, M., Adsumilli, B. & Bregler, C. (2021). MMSys' 21 grand challenge on detecting cheapfakes. arXiv preprint arXiv:2107.05297.

Atzori, L., Iera, A. & Morabito, G. (2010). The internet of things: A survey. Computer networks 54(15), 2787-2805.

Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C. & Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. arXiv preprint arXiv:1909.03242.

Beckett, C. (2019). New powers, new responsibilities: A global survey of journalism and artificial intelligence. Polis, London School of Economics and Political Science. https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities.

Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F. & Cui, X. (2012). Modeling the impact of short-and long-term behavior on search personalization. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 185-194).

Bitektine, A., & Haack, P. (2015). The "macro" and the "micro" of legitimacy: Toward a multilevel theory of the legitimacy process. *Academy of Management Review, 40(1):* 49-75.

Boididou, C., Middleton, S. E., Jin, Z., Papadopoulos, S., Dang-Nguyen, D. T., Boato, G. & Kompatsiaris, Y. (2018). Verifying information with multimedia content on twitter. Multimedia tools and applications 77(12), 15545-15571.

Borrajo, D., Roubíčková, A. & Serina, I. (2015). Progress in Case-Based Planning. ACM Computing Surveys 47(2).

Bouquet, P., Ghidini, C., Giunchiglia, F. & Blanzieri, E. (2003). Theories and uses of context in knowledge representation and reasoning. Journal of pragmatics 35(3), 455-484.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., and others (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Bryman, A. (2016). Social research methods. Oxford University Press.

Burel, G., Farrell, T. & Alani, H. (2021). Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter. Information Processing & Management, 58(6).

Burrows, S., Potthast, M. & Stein, B. (2013). Paraphrase acquisition via crowdsourcing and machine learning. ACM Trans. Intell. Syst. Technol. 4, 3, Article 43 (June 2013), 21 pages. https://doi.org/10.1145/2483669.2483676

Caldelli, R., Becarelli, R. & Amerini, I. (2017). Image origin classification based on social network provenance. IEEE Transactions on Information Forensics and Security 12(6), 1299-1308.

Caswell, D. (2019). Structured journalism and the semantic units of news. Digital Journalism 7(8), 1134-1156.

Ceolin, D., Groth, P., Van Hage, W. R., Nottamkandath, A., & Fokkink, W. J. (2012). Trust Evaluation through User Reputation and Provenance Analysis. *URSW*, *900*, 15-26.

Chyi, H. I. & Ng, Yee, M. M. (2020). Still Unwilling to Pay: An Empirical Analysis of 50 U. S. Newspapers' Digital Subscription Results, Digital Journalism, 8:4, 526-547, doi:10.1080/21670811.2020.1732831

De Nies, T., Coppens, S., Van Deursen, D., Mannens, E. & Van de Walle, R. (2012). Automatic discovery of high-level provenance using semantic similarity. In International Provenance and Annotation Workshop (pp. 97-110). Springer, Berlin, Heidelberg.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G.,Stanley, H. E. & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554-559.

Dong, X. L. & Rekatsinas, T. (2018). Data integration and machine learning: A natural synergy. In Proceedings of the 2018 international conference on management of data (pp. 1645-1650).

Farid, H. (2016). Photo forensics. MIT press.

Ferreira, T. C., van der Lee, C., van Miltenburg, E. & Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. arXiv:1908.09022 [cs]. http://arxiv.org/abs/1908.09022. Accessed 20 October 2021.

Flanagin, A. J. & Metzger, M. J. (2020). Source Credibility. The International Encyclopedia of Media Psychology 1-5.

Gaillard, S., Oláh, Z. A., Venmans, S. & Burke, M. (2021). Countering the Cognitive, Linguistic, and Psychological Underpinnings Behind Susceptibility to Fake News: A Review of Current Literature With Special Focus on the Role of Age and Digital Literacy. Front. Commun 6:661801, doi: 10.3389/fcomm.2021.661801

Galily, Y. (2018). Artificial intelligence and sports journalism: Is it a sweeping change? Technology in Society 54, 47–51. https://doi.org/10.1016/j.techsoc.2018.03.001

Gartner, R. & Gartner, R. (2016). Metadata. Springer.

Gervas, P. (2009). Computational Approaches to Storytelling and Creativity. AI Magazine 30(3), 49. https://doi.org/10.1609/aimag.v30i3.2250

Giachanou, A., Ghanem, B., Ríssola, E. A., Rosso, P., Crestani, F., & Oberski, D. (2022). The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. *Data & Knowledge Engineering*, *138*, 101960.

Goodfellow, I. J., Bengio, Y. & Courville, A. (2016). Deep Learning. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative Adversarial Networks. Communications of the ACM 63(11), 139–144. https://doi.org/10.1145/3422622

Grieco, E. (2020). U. S. newsroom employment has fallen 26% since 2008. Pew Research Center. https://www.pewresearch.org/fact-tank/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/, accessed 2021-10.

Google (2021). People + AI Playbook. Chapter on “User Needs + Defining Success”. https://pair.withgoogle.com/chapter/user-needs/ .

Guo, W., Wang, J. & Wang, S. (2019). Deep multimodal representation learning: A survey. IEEE Access 7, 63373-63394.

Harmon, D. J., Haack, P., & Roulet, T. J. (2019). Microfoundations of institutions: A matter of structure versus agency or level of analysis? *Academy of Management Review, 44(2):* 464-467.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., Sable, V., Li, C. & Tremayne, M. (2017). Claimbuster: The first-ever end-to-end fact-checking system. Proceedings of the VLDB Endowment, 10(12), 1945-1948.

Hendler, J., Gandon, F. & Allemang, D. (2020). Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL. ACM.

Hendrickx, J. (2022). Power to the People? Conceptualising Audience Agency for the Digital Journalism Era. *Digital Journalism,* 1-9.

Hervás, R. & Gervás, P. (2006). Case-Based Reasoning for Knowledge-Intensive Template Selection During Text Generation. In T. R. Roth-Berghofer, M. H. Göker & H. A. Güvenir (eds.), Advances in Case-Based Reasoning (pp. 151–165). Berlin, Heidelberg: Springer. https://doi.org/10.1007/11805816_13

Hossam, M., Le, T., Papasimeon, M., Huynh, V. & Phung, D. (2021). Text Generation with Deep Variational GAN. arXiv:2104. 13488 [cs]. http://arxiv.org/abs/2104.13488. Accessed 20 October 2021

Hutson, M. (2021). Robo-writers: the rise and risks of language-generating AI. Nature, 591(7848), 22-25.

Jones, K. (2012). Trustworthiness. Ethics 123(1), 61-85.

Jones, R. & Jones, B. (2019). Atomising the News: The (In)Flexibility of Structured Journalism. Automation, Algorithms and News: An International Conference, Ludwig Maximilians Universität, Munich.

Kazai, G., Yusof, I. & Clarke, D. (2016). Personalised news and blog recommendations based on user location, Facebook and Twitter user profiling. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 1129-1132).

Khan, S. A., Sheikhi, G., Opdahl, A. L., Rabbi, F., Stoppel, S., Trattner, C., & Dang-Nguyen, D. T. (2023). Visual User-Generated Content Verification in Journalism: An Overview. *IEEE Access*.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Kotonya, N. & Toni, F. (2020). Explainable automated fact-checking: A survey. arXiv preprint arXiv:2011.03870.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S. & Zhao, J. (2013). PROV-O: The PROV Ontology, World Wide Web Consortium (W3C).

Lee, D., Hosanagar, K. & Nair, H. S. (2018). Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. Management Science 64(11):5105-5131. https://doi org/10.1287/mnsc.2017.2902

Lee, S., Nah, S., Chung, D. S. & Kim, J. (2020). Predicting AI News Credibility: Communicative or Social Capital or Both? Communication Studies 71(3), 428–447. https://doi.org/10.1080/10510974.2020.1779769

Leppänen, L., Munezero, M., Granroth-Wilding, M. & Toivonen, H. (2017). Data-driven news generation for automated journalism. In Proceedings of the 10th International Conference on Natural Language Generation (pp. 188-197).

Lewis, S. C., Sanders, A. K. & Carmody, C. (2019). Libel by Algorithm? Automated Journalism and the Threat of Legal Liability. Journalism & Mass Communication Quarterly 96(1), 60–81. https://doi.org/10.1177/1077699018755983

Lin, B., & Lewis, S. C. (2022). The One Thing Journalistic AI Just Might Do for Democracy. *Digital Journalism,* 1-23.

Liu, X., Nourbakhsh, A., Li, Q., Shah, S., Martin, R. & Duprey, J. (2017). Reuters Tracer: Toward automated news production using large scale social media data. In 2017 IEEE International Conference on Big Data (pp. 1483-1493). IEEE.

Maiden, N., Zachos, K., Brown, A., Brock, G., Nyre, L., Tonheim, A. N., Apsotolou, D. & Evans, J. (2018). Making the news: Digital creativity support for journalists. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-11).

Marconi, F. (2020). Newsmakers. Columbia University Press.

Marconi, F., Siegman, A. & Machine Journalist (2017). The future of augmented journalism: A guide for newsrooms in the age of smart machines. New York: AP Insights. https://broadcastnewsabrahamsen.files.wordpress.com/2017/09/ap_insights_the_future_of_augmented_journalism.pdf , accessed 2023-04.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. & Gao, J. (2021). Deep Learning-based Text Classification: A Comprehensive Review. ACM Computing Surveys (CSUR) 54(3), 1-40.

Miroshnichenko, A. (2018). AI to bypass creativity. Will robots replace journalists? (The answer is “yes”). Information 9(7), 183.

Mishra, R. & Setty, V. (2019). Sadhan: Hierarchical attention networks to learn latent aspect embeddings for fake news detection. In Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (pp. 197-204).

Moran, R. E., & Shaikh, S. J. (2022). Robots in the News and Newsrooms: Unpacking Meta-Journalistic Discourse on the Use of Artificial Intelligence in Journalism. *Digital Journalism,* 1-19.

Motta, E., Daga, E., Opdahl, A. L. & Tessem, B. (2020). Analysis and Design of Computational News Angles. IEEE Access 8, pp. 120613-120626, doi:10.1109/ACCESS.2020.3005513.

Oh, A., Lee, H., and Kim, Y. (2009). User evaluation of a system for classifying and displaying political viewpoints of weblogs. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 3, No. 1).

Opdahl, A. L., Al-Moslmi, T., Dang-Nguyen, D. T., Gallofré Ocaña, M., Tessem, B., & Veres, C. (2022). Semantic Knowledge Graphs for the News: A Review. ACM Computing Surveys, 55(7), 1-38.

Opdahl, A. L. & Tessem, B. (2021). Ontologies for finding journalistic angles. Software & Systems Modeling 20, 71–87 (2021). https://doi.org/10.1007/s10270-020-00801-w .

OpenAI. (2023). GPT-4 Technical Report. arXiv:submit/4812508. https://cdn.openai.com/papers/gpt-4.pdf .

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

Paris, B. & Donovan, J. (2019). Deepfakes and cheap fakes: The manipulation of audio and visual evidence. Data & Society, September 18.

Park, S., Lee, S. and Song, J. (2010). Aspect-level news browsing: understanding news events from multiple viewpoints. Proceedings of the 15th International Conference on Intelligent User Interfaces, pp. 41-50.

Pasquini, C., Amerini, I. & Boato, G. (2021). Media forensics on social media platforms: a survey. EURASIP Journal on Information Security 2021(1), 1-19.

Pellegrini, T. (2012). Semantic metadata in the news production process: Achievements and challenges. In *Proceedings of the 16th international academic mindtrek conference* (pp. 125-133).

Plaisance, P. L. (2014). Trustworthiness in Digital Journalism, Psychology Today, October 30, 2014. https://www.psychologytoday.com/us/blog/virtue-in-the-media-world/201410/trustworthiness-in-digital-journalism , accessed 2021-11-15.

Pomerantz, J. (2015). Metadata. MIT Press.

Popat, K., Mukherjee, S., Yates, A. & Weikum, G. (2018). Declare: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2018). Language models are unsupervised multitask learners. Technical Report, OpenAI.

Ras, G., Xie, N., van Gerven, M., & Doran, D. (2022). Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329-397.

Resnick, P., Garrett, R. K., Kriplean, T., Munson, S. A. & Stroud, N. J. (2013). Bursting your (filter) bubble: strategies for promoting diverse exposure. In Proceedings of the 2013 conference on Computer supported cooperative work companion (pp. 95-100).

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T. & Saenko, K. (2018). Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Nature.

Schiffrin, A. (2019). Credibility and Trust in Journalism. Communication. *Oxford Research Encyclopedias: Communication,* Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.794

Schoenborn, J., Weber, R., Aha, W., Cassens, J. & Althoff, K.-D. (2021). Explainable Case-Based Reasoning: A Survey. In AAAI-21 Workshop Proceedings.

Seo, H., Blomberg, M., Altschwager, D. & Vu, H. T. (2021). Vulnerable populations and misinformation: A mixed-methods approach to underserved older adults' online information assessment. New Media & Society 23(7), 2012-2033.

Siles, I. & Boczkowski, P. J. (2012). Making sense of the newspaper crisis: A critical assessment of existing research and an agenda for future work. New media & society 14(8), 1375-1394.

Simon, F. M. (2022). Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism,* 1-23.

Skovsgaard, M. & Andersen, K. (2020). Conceptualizing News Avoidance: Towards a Shared Understanding of Different Causes and Potential Solutions, Journalism Studies 21:4, 459-476, doi:10.1080/1461670X.2019.1686410

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. Business Information Review, 34(3), 150-160.

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, 613:620-621.

Stokel-Walker, C. & Van Noorden, R. (2023). The promise and peril of generative AI. *Nature*, 614:215-216.

Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism* 7(8), 1076-1097.

Strömbäck, J., Tsfati, Y., Boomgaarden, H., Damstra, A., Lindgren, E., Vliegenthart, R. & Lindholm, T. (2020). News media trust and its impact on media use: Toward a framework for future research. Annals of the International Communication Association 44(2), 139-156.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of management review*, *20*(3), 571-610.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215 [cs]. http://arxiv.org/abs/1409.3215. Accessed 20 October 2021.

Taddeo, M. & Floridi, L. (2018). How AI can be a force for good. *Science* 361(6404), pp. 751-752.

Thorne, J. & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. arXiv preprint arXiv:1806.07687.

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313-313.

Trabelsi, A., & Zaïane, O. R. (2015). Extraction and clustering of arguing expressions in contentious text. *Data & Knowledge Engineering*,

Troncy, R. (2008). Bringing the IPTC news architecture into the semantic web. In International Semantic Web Conference (pp. 483-498). Springer, Berlin, Heidelberg.

Van Der Aalst, W. (2012). Process mining. Communications of the ACM 55(8), 76-83.

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, 614(7947), 224-226.

Verdoliva, L. (2020). Media forensics and deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing 14(5), 910-932.

Viazovetskyi, Y., Ivashkin, V. & Kashin, E. (2020). Stylegan2 distillation for feed-forward image manipulation. In European Conference on Computer Vision (pp. 170-186). Springer, Cham.

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. Science 359(6380), 1146-1151.

Ward, S. J. (2019). Journalism ethics. In The handbook of journalism studies (pp. 307-323). Routledge.

Wintterlin, F., T. Schatto-Eckrodt, L. Frischlich, S. Boberg, and T. Quandt. (2020). “How to Cope with Dark Participation: Moderation Practices in German Newsrooms.” Digital Journalism 8 (7):904–924.

Zheng, Z., Xie, S., Dai, H. N., Chen, X. & Wang, H. (2018). Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services 14(4), 352-375.

Zhuhadar, L. P. & Ciampa, M. (2021). Novel findings of hidden relationships in offshore tax-sheltered firms: a semantically enriched decision support system. Journal of Ambient Intelligence and Humanized Computing 12(4), 4377-4394.

Zorrilla, M., Borch, N., Daoust, F. et al. (2015). A Web-based distributed architecture for multi-device adaptation in media applications. *Personal and Ubiquitous Computing* ***19(5)*****,** 803–820.

ANDREAS L. OPDAHL is Professor of Information Systems Development at the University of Bergen, Norway, where he heads the Research Group for Intelligent Information Systems (I2S). Opdahl received his Ph.D. from the Norwegian University of Science and Technology (NTNU) in 1992. His research interests include ontologies and knowledge graphs, enterprise and IS modelling, as well as safety and security requirements. He is the author, co-author or co-editor of more than a hundred peer-reviewed research papers that have been cited many thousand times. He is a member of IFIP WG5.8 on Enterprise Interoperability and WG8.1 on Design and Evaluation of Information Systems. He serves regularly as a reviewer for premier international journals and on the program committees and as an organizer of renowned international conferences and workshops.

Andreas L Opdahl: Funding acquisition, Conceptualization, Writing - Original draft preparation. Writing - Review & Editing, Visualization; **Bjørnar Tessem:** Conceptualization, Writing - Original draft preparation. Writing - Review & Editing; **Duc-Tien Dang-Nguyen:** Conceptualization, Writing - Original draft preparation. Writing - Review & Editing; **Enrico Motta:** Conceptualization, Writing - Original draft preparation. Writing - Review & Editing; **Vinay Setty:** Conceptualization, Writing - Original draft preparation. Writing - Review & Editing; **Eivind Throndsen:** Writing - Reviewing and Editing; **Are Tverberg:** Writing - Reviewing and Editing; **Christoph Trattner:** Funding acquisition, Writing - Reviewing and Editing, Visualization.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: