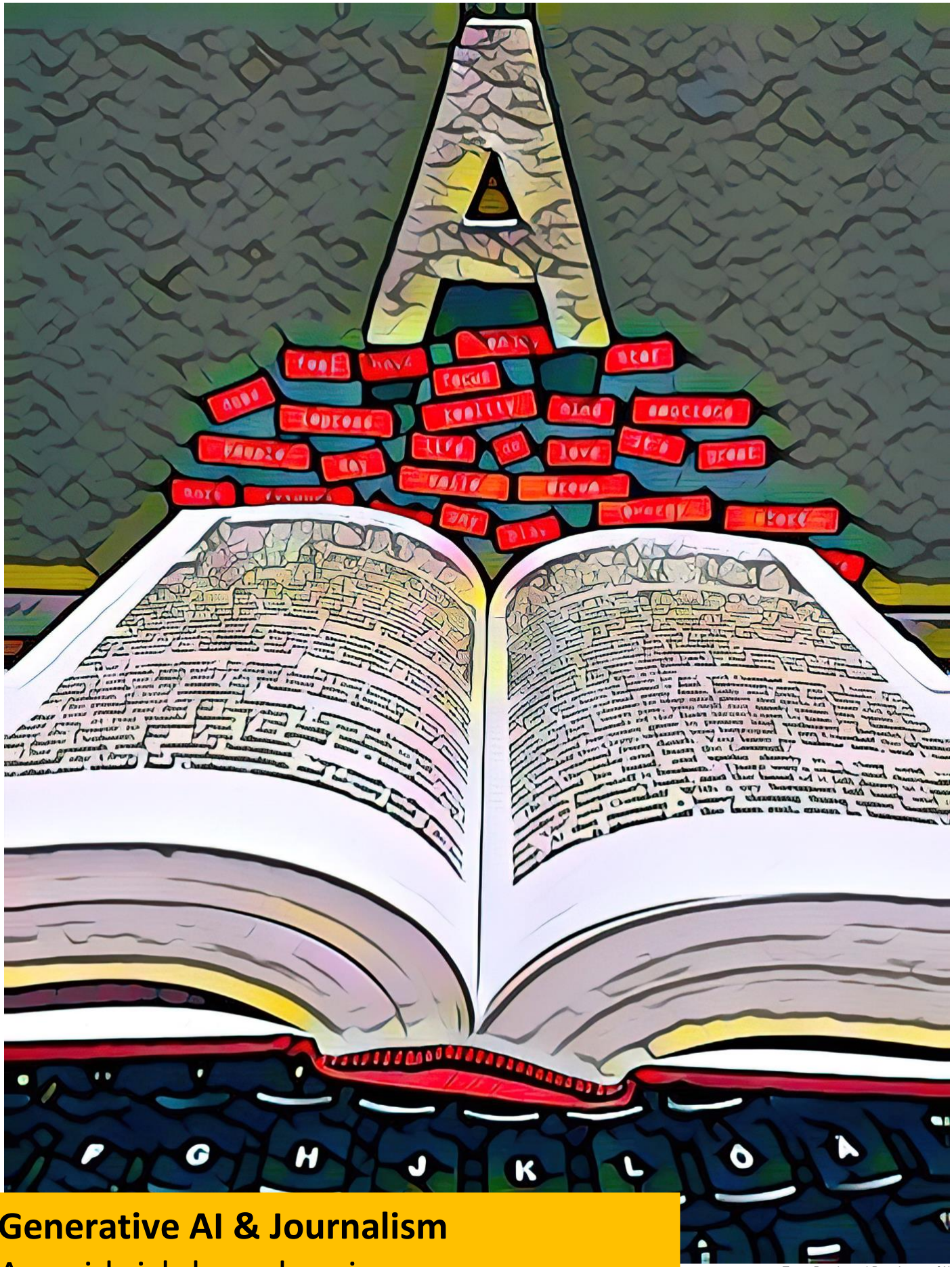
Generative AI & journalism



Teresa Berndtsson | Better Images of AI
Letter Word Text Taxonomy | CC-BY 4.0

Generative AI & Journalism

A rapid risk-based review

Dr Bronwyn Jones
Dr Rhianne Jones
Prof Ewa Luger

Generative AI & Journalism

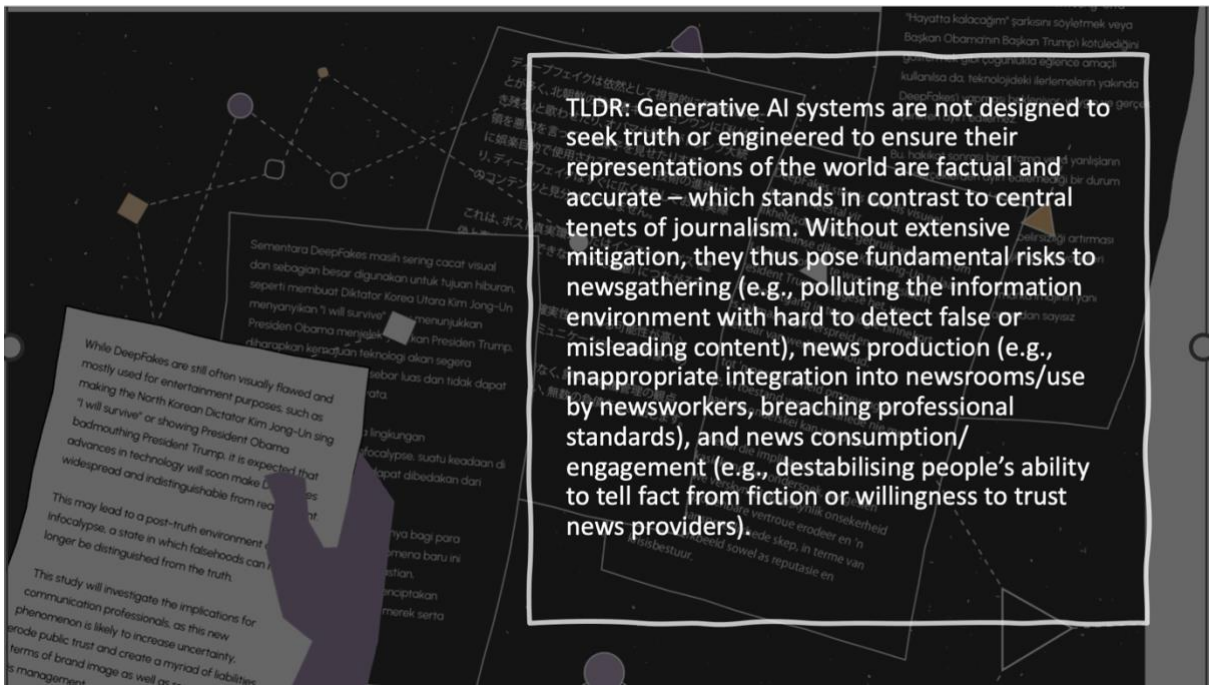
A rapid risk-based review

Authors: Dr Bronwyn Jones, Dr Rhianne Jones, Prof Ewa Luger

This **rapid review** outlines a range of existing and potential **risks** generative AI poses if incorporated into journalism, written with **newsroom leaders and journalists** in mind. It is intended as a quick entry point into live and rapidly evolving discussions of the issues, with links and references out to useful resources – some academic and peer-reviewed, some journalistic. It is *not* a comprehensive analysis or an exploration of applications or benefits (of which there are a growing number of resources e.g., here). For ease of navigation, the document is structured into three broad risk categories: editorial, legal, and societal *(see the following page for an overview).*

It was created as an output of collaboration between the University of Edinburgh and the BBC R&D Responsible Innovation team, as part of the PETRAS *Building Public Value via Intelligible AI* project. The work underpinning it includes: a review of existing research and grey literature, expert workshops with BBC staff, interviews and focus groups with BBC journalists.

Why have we produced this? Generative AI is a branch of general purpose AI (also referred to as foundation models) that can create media content of varied types, including text, images, audio and code (see here and here for further explanation). Generative AI systems such as Large Language Models (LLMs) have pushed the boundaries of what is possible in content generation and created new challenges and risks for society. They will likely have significant impacts on news organisations and journalists as well as audience members/news users, impacting how news is gathered, produced, distributed and consumed. However, the news media industry currently lacks an advanced understanding of exactly how they work, when and how they fail, and what mitigations are required to ensure they work in the public interest.



OVERVIEW

1. EDITORIAL RISKS 1

INACCURACY & FABRICATION 1
LACK OF SOURCE TRANSPARENCY 2
PLAGIARISM 2
BIAS & ASSOCIATED HARMS 2
UNDERMINING EDITORIAL VALUES 3
QUALITY DEGRADATION 4
BREACHING AUDIENCE EXPECTATIONS 4
+ COMPOUNDING FACTORS 4
INAPPROPRIATE USE 4
LACK OF TRANSPARENCY & EXPLAINABILITY 5

2. LEGAL & REGULATORY RISKS 5

COPYRIGHT INFRINGEMENT 5
LIBEL & DEFAMATION 6
DATA PRIVACY BREACHES 6
ONEROUS TERMS OF SERVICE/USE & CONTRACTS 6
DATA LEAKS 7
+ COMPOUNDING FACTORS 7
RAPIDLY CHANGING ENVIRONMENT & SLOW LEGAL RESPONSES 7

3. SOCIETAL RISKS 7

BUSINESS MODEL/MARKET DISRUPTION 7
POLLUTION OF INFORMATION ECOSYSTEM 8
UNDERMINING TRUST IN INFORMATION & MEDIA 8
DISRUPTION TO JOBS & VALUING OF HUMAN LABOUR 9
CENTRALISED POWER 9
REPRODUCING INEQUALITIES & REINFORCING SOCIAL HARMS 10
DEGRADING THE ENVIRONMENT & PERPETUATING UNSUSTAINABILITY 10
+ COMPOUNDING FACTORS 10
RUSH TO AUTOMATE 10
EMBEDDEDNESS & INTERCONNECTEDNESS 10

1. EDITORIAL RISKS

Breaching professional standards

INACCURACY & FABRICATION

RISK: Generative AI may provide false or misleading information which is difficult to detect and could make its way into published content either inadvertently or as a result of deliberate efforts by others to manipulate or dupe journalists.

WHY: These models are trained by scraping, analysing, and processing massive amounts of publicly available data from the internet, then probabilistically reproducing the patterns they observe to mimic the desired output (e.g., natural language, paintings or photographs). They are not designed to seek truth or engineered to ensure their representations of the world are factual, accurate or up-to-date. For instance, image models create representations of people and things that can be highly realistic but are entirely synthetic. ChatGPT produces text that is authoritative and plausible but often subtly incorrect, generating responses to user prompts that are not always wrong… but not always right. LLMs are not designed to understand, reason, or express underlying information in natural language - they are only manipulating the form of language. The low barriers to use (cheap, little expertise needed etc.) mean they can be used to create mis- and dis-information at scale or to target journalists with tailored false information or personal attacks, thus exacerbating existing threats.

IMPLICATIONS: Without robust new editorial rules, processes, and conventions for ensuring responsible use that aligns to professional standards, generative models could open up risks for the inadvertent publication of incorrect information. This in turn raises legal risks, including claims of libel, defamation, and breach of privacy (see 2.) – all well-worn legal considerations for journalists, but which could be exacerbated or present in new ways. Journalists will likely have to deal with increasing amounts of auto-generated source material provided to news organisations, placing an increased emphasis on fact-checking and AI literacy in the newsroom. Additionally, newsroom developers will need to be wary of inaccuracies in AI-generated code. Moreover, it remains unclear how susceptible these systems are to forms of hacking and manipulation e.g., via data poisoning using methods such as prompt injection (where bad actors subvert training data) and whether the growing trend of connecting generative models to other applications, plug-ins etc., will open up new vectors for attack. Beyond the newsroom, these systems could generate false information about news organisations and journalists that could be damaging to their reputations.

RELATED EXAMPLES: German public broadcaster Bayerischer Rundfunk tried testing GPT-3 for fact box generation and found "fictional numbers, wrongly connected facts, and hallucinations" which ended up making sub-editing more time-consuming than for manual creation. Tech publisher CNET had to issue multiple corrections after publishing AI-written articles containing numerous errors. ChatGPT made up a Guardian article citation in response to a researcher's query, which the journalist who supposedly wrote it found plausible enough to prompt him to check if it existed in the newspaper's archive. A hoaxer used ChatGPT and DALL-E to dupe the Irish Times into publishing a supposed opinion piece, leading the publisher to apologise and review pre-publication procedures to make them more robust in the face of generative AI. In non-journalistic contexts, a song featuring AI-generated vocals purporting to be Drake and the Weeknd was pulled from streaming services, a photograph that won an international photography competition was withdrawn after its creator said it was made with the aid of AI, and images that went viral of Pope Francis wearing a Balenciaga bubble jacket were shown to be fake.

LACK OF SOURCE TRANSPARENCY

RISK: Generative AI systems do not provide source transparency, which makes outputs difficult to check and breaks the chain of attribution necessary for journalists to justify their claims to knowledge and be held accountable.

WHY: Most generative AI tools are not able to, or not enabled to, show the origins of the information they create and are intended to generate outputs that are persuasive and pleasing to humans, but not traceable to their origins or adhering to established citation practices. For instance, ChatGPT will make up fictitious citations and reference non-existent work, sometimes merging real people, organisations, titles etc. with completely fabricated information. Even when asked what source material it has drawn from, existing LLM-based models cannot reliably provide sources or citations. Additionally, proprietors of (non-open-source) systems can alter their product or training data at any time, which can change outputs and means they may not be reproducible.

IMPLICATIONS: This blocks journalists' ability to locate information in context and hinders attempts to make well-informed decisions about its veracity – even though they and their editors will be the ones held responsible for any errors. If journalists were to rely on generative AI systems to make claims to knowledge, they would have to independently check all information beyond accepted knowledge, often without having access to the source information. The onus is on the journalists and their news organisations to do due diligence to ensure they are relying on sound sources and can justify their claims to knowledge. Both surreptitious and explicit use of tools like ChatGPT could pose a challenge for sub-editors/editors who must check output but whose processes rely on journalists' skills in sourcing and assessing claims and who are attuned to human, rather than machine error. Generative AI's tendency to invent sources and citations, and the risk these leak out into the wider information ecosystem further compounds this situation.

RELATED EXAMPLES: The public demo of Meta's Galactica was taken down after just three days when, despite being trained on scientific material, it made up facts and citations. A lawyer in the US is facing sanctions after using ChatGPT for his legal filings after it manufactured previous cases, which he cited.

PLAGIARISM

RISK: Generative AI may plagiarise existing works without recognition of the source material, which could make its way into news output or conversely, it could regurgitate proprietary news output without accreditation. This in turn raises legal risks, including claims of copyright infringement (see COPYRIGHT INFRINGEMENT).

WHY: Systems are trained on masses of training data - and are opaque about how exactly they draw from that data – so risk reproducing parts of it wholesale.

IMPLICATIONS: Since the efficacy of detection tools for AI generated materials (like GPT Zero) and of plagiarism detectors (like Turnitin) are disputed, and the field moves so quickly that they become rapidly outdated, journalists may have trouble identifying plagiarised material.

RELATED EXAMPLES: Stock image watermarks have appeared on DALL-E generated images, and some artists' work has been replicated almost exactly. CNET's AI-generated articles included substantial plagiarism with deep structural and phrasing similarities in the text to articles others had published.

BIAS & ASSOCIATED HARMS

RISK: Generative AI can discriminate unfairly, perpetuate stereotypes and social biases, and overrepresent hegemonic viewpoints, which could feed downstream into journalistic outputs and cause representational harms that may be damaging to marginalised populations (see REPRODUCING INEQUALITIES & REINFORCING SOCIAL HARMS).

WHY: Generative models can pick up on subtle biases and overtly abusive patterns in training data and incorporate them into generated material. The ever-increasing size of the data sets on which language models are trained necessitate incorporation of non-curated and often undesirable material (including for instance pornographic and explicit imagery, and abusive language), which skews the model's likelihood of replicating these associations. Companies make decisions about whether and how to try to mitigate these biases but in a rush to commercialise, many do not prioritise mitigation (e.g., via data set curation, model choices, and filters).

IMPLICATIONS: This could cause harms to news audiences if integrated into output in ways that allow for the reproduction of racist, sexist, ableist, extremist or other harmful ideologies or inclusion of derogatory or toxic language, e.g., inciting hate or violence. There is a particular risk of compounding representational harms that for instance, reinforce subordination of certain groups, erase or fail to recognise certain groups or perspectives, stereotype, demean or stigmatise certain people. These risks can be amplified for different language communities and for different global communities. This is compounded when the view of the world presented by these systems is presented as, and perceived by audiences as, objective.

RELATED EXAMPLES: A female journalist of Asian heritage found that the Lensa AI avatar generator app created cartoonishly pornified representations of her while her male colleagues were portrayed as astronauts, explorers, and inventors and white female colleagues were not overly sexualised in the same way.

UNDERMINING EDITORIAL VALUES

RISK: News organisations have editorial commitments to particular values, which could be compromised as a result of incorporation of generative AI tools or outputs. For instance, public service media have devised processes and conventions for ensuring impartiality, fairness, balance, diversity, and universality and it is not yet clear how generative models may pose challenges to upholding these values in the near or long-term.

WHY: News organisations are not in control of the ways in which generative AI systems are developed, and for the most part, are not able to access information necessary to be able to make assessments or judgments on the appropriateness or suitability of these tools with regard to specific normative values they seek to uphold. While professional standards are made explicit, values are higher order statements about what people see as desirable and that guide their conduct - they are not easily distilled into simple rules of conduct and are subject to interpretation and change over time. Rich human practices have developed over time and are enacted by journalists who have accumulated experience and expertise through ongoing processes of learning, discussion, and deliberation. The inability to interrogate and influence decisions affecting the development of these systems, may result in value misalignment and risk undermining or compromising organisational/professional integrity.

IMPLICATIONS: Frequent and distributed use of generative AI tools could inadvertently undermine important values that could over time result in reputational or representational harms, for example, use of generative AI could result in groups being treated differently, or forms of partiality creeping into output. Other implications could include the independence of news organisations coming into question if they become reliant on third-party systems like ChatGPT to underpin their products and services, and in this case to inform the construction of meaning through language in their communication with audiences. In this way, impartiality could also be impacted if political biases absorbed into datasets and models find expression in organisational output.

RELATED EXAMPLES: Previously cited examples of breaches to professional standards in the form of biased, plagiarised, uncited, and inaccurate or fabricated material indicate how editorial values such as diversity, honesty, transparency, and truth may be impacted and potentially undermined.

QUALITY DEGRADATION

RISK: Limitations of control over output or underqualified people using generative AI tools to perform tasks they otherwise don't have the skills to do, could lead to lowered quality standards, and deviation in for example consistency, tone, or brand continuity.
WHY: Generative AI tools draw from the same dataset for each user and so have been noted for producing similar and standardised outputs to user prompts. Furthermore, though language models can produce text in a range of particular styles if prompted, the general style of writing tends to be bland as it reproduces the most likely composition of words.
IMPLICATIONS: If all news outlets are using the same tools to aid reporting, there is a risk of producing similar, dull and homogenous writing.
RELATED EXAMPLES: CNET's auto-generated copy was found to be full of mistakes and compromised the perceived quality of the outlet, which had to issue multiple corrections.

BREACHING AUDIENCE EXPECTATIONS

RISK: Both disclosed use, and a lack of transparency when using generative AI could undermine public trust by breaching their expectations of news publishers.
WHY: Audience expectations concerning news production and journalistic output have evolved in an era in which AI was not present and so conceptions of authenticity and journalistic integrity are highly based on human creation of news. Though AI applications in news have grown in recent years, it is not yet clear what audience reactions will be to incorporation of generative AI or how their expectations will change. As such, there is a lack of clarity currently over best practice concerning disclosure and disclaimers on content generated with the use of generative AI, particularly as the extent and nature of use varies. How different levels of, and approaches to transparency will impact trust is unclear.
IMPLICATIONS: If there are widely differing approaches across news organisations regarding whether and how they disclose use of generative AI, and explain its role in news production and particular use cases, this could prompt unease or mistrust among audiences towards journalism more broadly and towards specific providers. This also has implications for the public's ability to hold journalists and organisations responsible for the news they produce, as tracing human accountability through AI-human infrastructures is complex and auditing process are not yet mature.
RELATED EXAMPLES: CNET changed the byline they use for articles created with an AI engine to always include 'CNET Money' as an identifier of the AI system, with the disclaimer: "This article was assisted by an AI engine and reviewed, fact-checked and edited by our editorial staff." It also added an editors' note to stories *about* generative AI to give context to their relationship with the tools they're writing about. It says: "CNET is using an AI engine to create some personal finance explainers that are edited and fact-checked by our editors. For more, see this post." Wired was one of the first publishers to outline their "ground rules" for generative AI followed by others like the Financial Times, which said it will "explore using AI-augmented visuals (infographics, diagrams, photos)" and "make that clear to the reader".

+ COMPOUNDING FACTORS

INAPPROPRIATE USE

RISK: Inappropriate use stemming from lack of literacy amongst journalists, editors and decision-makers and/or a lack of or unsuccessful mitigation, could lead to negative impacts on output, workforce and audiences. For instance, uncritical use and anthropomorphising of generative AI systems by newsworkers leads to an overestimation of their capabilities, overreliance (automation bias), complacency, and unsafe/inappropriate use. The opposite extreme of this is lack of understanding leading to algorithmic aversion and fear or blanket mistrust of AI tools, hampering efforts to responsibly integrate AI in safe, secure and helpful ways. This risk is heightened by recent

moves towards integrating generative AI into widely used software (e.g., Google's LaMDA into Gmail, Docs, Drive), which raises questions of knowing where and when generative AI is at play and concomitant issues of newsrooms recognising when and where it is impacting their staff and work. The capabilities of these systems increasingly resemble those of humans but they work in ways that are fundamentally different from how humans work. This important distinction can get lost when we use terms previously reserved for describing people, such as "knows", "believes", and "thinks", to describe AI. This is important because journalists using for instance ChatGPT assume responsibility for the output accuracy according to Open AI's terms, which releases the company from accountability.

LACK OF TRANSPARENCY & EXPLAINABILITY

RISK: The vast majority of the companies behind these generative AI systems do not share details of how their systems work or allow scrutiny and auditing of them. Additionally, many of the techniques they use are not interpretable and create 'black box' systems that even experts find difficult to accurately explain. This makes it difficult for news organisations to fully understand the tools they are using or interrogate the underlying training data, models, or algorithms. This then pushes the responsibility for undertaking checks and balances to understand and mitigate risk onto news organisations and creates a challenge to accurately explaining how the systems work to journalists and the audience, how/why exactly certain outputs were generated, or creating mechanisms for recourse following complaints. Additionally, it restricts their ability to ensure legal and regulatory compliance (see below).

2. LEGAL & REGULATORY RISKS

Breaching laws and regulations

COPYRIGHT INFRINGEMENT

RISK: Generative AI can infringe on intellectual property rights and distribute copyrighted materials without permission, placing journalists and news organisations at risk of legal claims if they use these systems and their outputs. It is also unclear if outputs created when using these tools can be copyrighted.

WHY: Models are trained on masses of data scraped from the web without permission of the creators of that data.

IMPLICATIONS: Newsrooms may in future face claims based on their use of these systems and may damage relationships with people and organisations in the creative sector by openly using them despite them being seen as exploiting people's copyrighted work. It is not clear currently to what extent the output of generative AI models is protected by copyright and this may be different across jurisdictions. Numerous lawsuits are in progress – e.g., alleging infringement of images in training data. Rights holders are suing after AI providers have used their data to train models. This primer and FAQ outlines the main issues from an editorial perspective. In the European Union, the recent EU draft AI Act would oblige companies deploying generative AI to disclose any copyrighted material used to develop their systems – but this is not yet a regulatory requirement.

EXAMPLES: Getty Images is suing Stability AI for copyright infringement "on a staggering scale" and individual artists are suing to have their work removed from training data sets. Adobe Firefly is an example of attempted mitigation by using dedicated AI models trained only on data that is legally obtained with appropriate licenses in place and recompense to creators.

LIBEL & DEFAMATION

RISK: Libelous or defamatory content could be published by a news provider, leaving them open to legal claims and pecuniary and reputational damage.
WHY: Generative AI tools are not designed to respect truth or to evidence claims. LLMs for instance have no way of 'knowing' what is accurate or of reasoning about what is true.
IMPLICATIONS: News organisations have strict rules and checks relating to the legality and likely risk of what they publish and are continually engaged in balancing the public interest of publishing information against the potential harms it could cause. If automating elements of production using generative AI, there would be a need for new or different mechanisms, checks, and balances to prevent the tendency of these systems to fabricate having knock-on effects on published news output. Other areas of legal risk include contempt of court (particularly through triangulation of data) and ethical risks include breach of confidentiality and source disclosure.
EXAMPLES: A mayor in Australia planned to sue OpenAI for defamation after ChatGPT falsely claimed he had served time in prison for bribery. A law professor found it had concocted an accusation of sexual assault against him and cited a non-existent Washington Post article.

DATA PRIVACY BREACHES

RISK: Journalists could input private or personal identifiable information, which could then be included in downstream uses for training or be leaked, e.g., others getting a response from ChatGPT that includes other peoples' private data.
WHY: Training data for proprietary models is unknown which (for many models) makes it unclear whether personal data used to train models was used lawfully – and then whether any potential data leakages are occurring that could amount to a data breach.
IMPLICATIONS: This raises questions about whether anyone can use the tools legally and in compliance with local data protection laws. For example, LLM-based text generators could include private messages or texts written by children who cannot legally consent to use of their data. News organisations will have to consider whether their data protection obligations are compatible with any planned use.
EXAMPLES: Italy suspects ChatGPT is breaching GDPR and issued temporary demand for OpenAI to stop using Italians' personal information. Open AI relies upon "legitimate interests" when it "develops" its services but there are claims it does not have lawful grounds for data processing. The company has now provided ways for users in Europe to opt out and request deletion of their personal data in response.

ONEROUS TERMS OF SERVICE/USE & CONTRACTS

RISK: By using generative AI tools, journalists and news organisations are accepting the tool provider's terms of use, which may incur significant legal obligations and grant rights to the tool provider that could lead to situations that cause financial and/or reputational damage.
WHY: Organisations making these tools available are looking to gain from people's interaction with them beyond directly paying for the service they provide. It is often in the provider's favour to construct terms by which the user owns input and output, which means they also take responsibility and own liability – and indemnify the provider for losses arising from use. However, the provider may also have the right to use such input and output themselves e.g., in training data, and companies are not transparent about what they do with this data.
IMPLICATIONS: Journalists and news organisations will be held responsible for their inputs (prompts) and outputs and claims could be made against them. Additionally, they will be giving providers broad

rights to anything used as a prompt or other input which could breach non-disclosure agreements, accidentally share private and proprietary information, and trigger serious liability risks.
EXAMPLES: See above examples regarding data privacy, copyright infringement.

DATA LEAKS

RISK: People (workers) inputting sensitive or confidential data, which is then impossible to retrieve as it is stored on servers belonging to the company in question and could be leaked to others.
WHY: It is known that companies use this data to further train models and unknown what else they use it for.
IMPLICATIONS: The data could be shared or reconstituted by the generative AI tool in future outputs.
EXAMPLES: Samsung's semiconductor arm leaking source code for a new program, internal meeting notes data relating to their hardware after workers used them as inputs. A bug in ChatGPT temporarily exposed AI chat histories to other users.

+ COMPOUNDING FACTORS

RAPIDLY CHANGING ENVIRONMENT & SLOW LEGAL RESPONSES

RISK: The rapid pace of technological change stands in contrast to the slow legal and regulatory environment. Test cases have yet to indicate how exactly existing law will deal with the aforementioned issues raised and regulation takes significant time to develop and be approved. The specific characteristics of emerging issues are unlikely to be known for some time but if use of generative AI becomes widespread in newsrooms, this could cause legal, regulatory and reputational repercussions down the line. Furthermore, there will be legal and regulatory divergences across nations and blocs globally.

3. SOCIETAL RISKS

The term societal risk is used here to refer to negative or undesirable consequences that could affect groups of people in society or society at large – they can be social, economic, political, and environmental. We consider the frequency, number of people affected, and significance of the consequences to determine what qualifies as a societal risk. These tend to be longer-term risks that occur as a result of incremental and aggregate change over time with multiple contributory factors that interrelate. The following list is not comprehensive but points to a selection of societal risks that hold significant implications for journalism.

BUSINESS MODEL/MARKET DISRUPTION

RISK: Rapid and widespread change to the environment in which news is produced could undermine the (already struggling) existing business model for journalism and precipitate the shrinking or closure of reputable providers (without replacement by new but similarly reputable entrants).
WHY: Generative AI lowers the cost and barriers for content generation and news production, enabling new competitors to enter the market which may not share the goals, ethics, practices etc., of public interest journalism (e.g., content mills, misinformation operations) but which may look like/mimic the genre (via design, format, structure, linguistic conventions etc.) A growing number of content farms are already churning out clickbait articles. There is also a chance that if generative AI is fully integrated into search technologies, the model of linking to web pages and driving traffic to news providers gets

broken and the news industry struggles to retain audience numbers, impacting revenue. Additionally, the sheer amount of cheaply produced, low quality information that could flood the internet would compete for people's attention and could reduce their engagement with news. Some news organisations are already requesting compensation for use of their content as a way to find some financial recompense in this altered environment.

POLLUTION OF INFORMATION ECOSYSTEM

RISK: The internet could become flooded with false and misleading information and people could find it increasingly hard to know what is true and trustworthy.

WHY: Increasingly easy access to generative AI designed for non-expert users makes it easier to generate synthetic content which can be distributed at scale. This affordance can be weaponised by bad actors and used to target populations, organisations or individuals and increase the efficacy of mis- and dis-information campaigns. It could be used to add noise around a topic in an attempt to drown it out or distort the narrative and powerful people and state actors could use the situation to profit from seemingly plausible denial of real wrongdoing, profiting from the 'liar's dividend'. This could also be an inadvertent impact of general use e.g., automating content production jobs such as public relations, advertising, and (non-news) information dissemination, and of content farms making money by imitating news outlets whilst distributing unverified and false information. A situation in which such poor quality information abounds could exacerbate existing mistrust among certain populations of news media and lead to an 'authenticity crisis' in which broad swathes of people lose trust in media and potentially other institutions in society. It could destabilise publics' abilities to assess and understand the news and worsen existing inequalities amongst publics, including digital and generational divides. Efforts to develop mechanisms to watermark AI-generated material and to track and register provenance of news are underway but it is unclear if they can be effective. Moreover, if a high volume of mis- and dis-information is then used to train further generative systems, this could compound the problem and create increasingly incorrect outputs.

UNDERMINING TRUST IN INFORMATION & MEDIA

RISK: If journalism fails to live up to public expectations and professional standards are lowered as a result of industry and widespread use of generative AI, there is a risk that people's trust in terms of their relation to information and/or with those organisations providing it is disrupted. The implications of this are complex and manifold but include growing disengagement with traditional news media and increasing levels of skepticism and/or distrust in news organisations, which in turn risk undermining news providers' legitimacy and viability in society. This in turn could result in wider negative impacts such as diminished levels of informed public debate, which could in undermine engagement with public issues and importantly, with democratic processes.

WHY: Journalism comes in many forms and plays many roles in society but informing the public of important, relevant, and up-to-date information about the societies in which they live in order that they can engage in civil society and democratic activities is a central function. The success of this endeavour relies on a degree of trust on the part of those receiving and engaging with news that it meets certain professional standards, including: accuracy, pursuit of truth, and timeliness. Trust here is relational and refers to the belief one has in the honesty, reliability and integrity of another. Widespread use of generative AI and the creation of synthetic forms of news media embody multiple risks as previously discussed, from the circulation of inaccurate, misleading or fabricated information, to reproducibility of existing biases in reporting, to privacy breaches. These risk destabilising people's trust and undermining the 'social contract' they have with news providers. Inadequate, uncritical or

alarmist coverage of AI, much of which is dominated by industry and business interests, is a separate but compounding issue.

DISRUPTION TO JOBS & VALUING OF HUMAN LABOUR

RISK: Rapid disruption to the labour market could result in job losses and/or changes the nature of work in undesirable ways and lead to human creativity being undervalued and poorly recompensed.
WHY: Generative AI tools may be used to automate specific tasks, decisions, or aspects of workflows previously performed by journalists, which would likely impact the availability and nature of jobs and the way we value human labour associated with journalism. Whether this is in fact problematic and results in undesirable outcomes for journalists and journalism depends on how the process is managed, including which values and outcomes are prioritised by those making decisions about resource allocation, quality assurance, and the editorial and ethical direction of news production. In recent years, synthetic newsreader avatars have been deployed in China and Reuters began prototyping automated sports. News organisations are already using AI-generated art and design for certain products (it unclear whether they would have otherwise paid artists, designers and photographers for this work) and Buzzfeed is using ChatGPT to help create quizzes and to personalise some editorial content, while at the same time we are seeing newsroom closure and job losses. There are diverse ways this risk could manifest, for example entry level news jobs that are typically made up of tasks that could be automated (e.g., reading and summarising documents, writing briefs) may be reduced if generative AI tools can speed up or make the process more efficient. This could result in barriers to entry into the newsroom via the more traditional development/progression paths for younger people – particularly those from non-elite or traditional backgrounds, or a deskilling and undervaluing of certain types of journalistic work and skills which are important for the role. Use of generative AI may lead to closure of roles or reduce roles to mundane or repetitive 'factory line' functions of checking or reviewing generative AI output. For example, IBM said it will pause hiring for roles it thinks could be replaced with AI in the coming years and this may now include previously hard-to-automate knowledge work. Additionally, further up the value chain, concerns have been raised around exploitation and the working conditions of workers involved in aspects of the creation of these tools, for example with regard to content labelling, filtering and moderation. This could breach news organisations' commitments to worker rights and tarnish their reputation if perceived as contributing to such harmful practices through business relationships.

CENTRALISED POWER

RISK: Use of generative AI may centralise power in the hands of a few companies and shift it away from workers and news organisations, which could compromise their ability to act independently and with autonomy.
WHY: The vast costs associated with developing generative AI, and particularly LLMs, means only a small number of companies have access to the immense resources and capital needed to develop them. This could result in large shifts in power which results in the wider generative AI research and infrastructure being concentrated and centralised it in the hands of a few (private) organisations who determine terms of use and monopolise necessary infrastructure. If the generative AI landscape is dominated by only a few players occupying dominant market positions news media organisations risk over relying on a few companies with regards to proprietary models and tools. There is a risk of vendor lock in and news organisations – particularly smaller organisations – lacking bargaining power around terms of use or being subject to pricing structures which could result in newsrooms being trapped in expensive contracts. When taken alongside wider issues such as the lack of visibility with regard to details of the model and limited power to inspect, audit, challenge, negotiate, this could compromise journalistic/professional values and ability to act independently and exert autonomy, which is

particularly important for public service media. This could exacerbate an already existing power inequality in the news industry.

REPRODUCING INEQUALITIES & REINFORCING SOCIAL HARMS

RISK: The aggregate impact of representational harms writ large through widespread use of generative AI in news media risks reproducing social inequalities across society.
WHY: Generative AI outputs reflect the value systems of the people who developed them and the prejudices, biases, and presuppositions present in the training data, which skews to western hegemonic positionality and risks erasing and deprioritising the cultural identity of non-western views. This and the aforementioned risks of BIAS & ASSOCIATED HARMS, if repeated across different generative AI use cases and not mitigated, risk over/under/mis-representing social groups, stereotyping or stigmatising them, and causing representational harms at scale. There is also a risk of exacerbating differential quality of service, since LLMs work better in some languages than others, and over/under-serve groups, cultures and societies. There could also be a worsening of differential access and accentuation of digital divides, whereby AI tools may be limited or restricted to those that can afford them, have the technical and computing resource needed to make use of them, and skills and opportunities to benefit from them, leading to the reproduction or exacerbation of social inequalities (conversely, given the right conditions, they could have democratising effects or help or assist some formerly disadvantaged users). Similarly, if the POLLUTION OF INFORMATION ECOSYSTEM comes to pass, there could be a further division between those who can afford access to high quality and reliable news and those who cannot.

DEGRADING THE ENVIRONMENT & PERPETUATING UNSUSTAINABILITY

RISK: The use of generative AI models could lead to undesirable environmental impacts due to their large carbon and water footprints, which contravenes news organisations' stated commitments and audiences' expectations/wishes.
WHY: Generative AI models come at a cost to the planet, due to factors including the environmental toll of energy intensive training and the mining of rare minerals for components. Recent advances in the field have been underpinned by large neural network models which necessitate exceptionally large computational resources and similarly substantial energy consumption to train and to deploy in order to service user queries.

+ COMPOUNDING FACTORS

RUSH TO AUTOMATE

There are financial incentives to automate more activities, more quickly, and with goals of deploying more fully autonomous systems without slowing down to adequately think about, consult on, and mitigate risks. Big tech companies are engaged in a competitive 'arms race' and 'land grab' to establish dominance and market supremacy, which is leading to an environment in which hyperbole prevails, developments are rushed to market, and concerns, checks and balances are ignored or deprioritised.

EMBEDDEDNESS & INTERCONNECTEDNESS

Generative AI models are being increasingly integrated into multiple existing enterprise systems through business collaborations, for instance Microsoft and Google embedding models into products such as calendars and email. Moreover, the app ecosystem enables integrating of generative-AI driven elements into apps, which interact with each other. This makes it difficult to know where and when generative AI is being used and what the implications are for security, privacy, and editorial purposes. Moreover, the defects of the model in question are then inherited by all adapted models downstream.

Dr Bronwyn Jones
Translational Fellow, University of Edinburgh
Journalist, BBC

Bronwyn Jones is a scholar and practitioner; she splits her time between working as a journalist for the BBC and researching the role of AI, algorithms, and automation in news production at the University of Edinburgh. Her research focuses on the implications of data-driven systems for public service news in democratic societies. She is Translational Fellow on the AHRC Bridging Responsible AI Divides (BRAID) programme exploring how the arts and humanities can contribute to enabling responsible AI innovation.

Dr Rhianne Jones
Lead, Responsible Innovation & Society, BBC Research and Development

Rhianne Jones is the BBC R&D lead for Responsible Innovation. She has over 10 years' experience researching and developing emerging technologies in the media industry. Rhianne's work helps to ensure that technology and innovation benefits people and society. Her research spans questions concerning responsible use of data and machine learning to how changes in technology impact media access, participation, and inclusion.

Prof Ewa Luger
Professor of Human Data Interaction, University of Edinburgh

Ewa Luger is co-director of the Institute for Design Informatics, Fellow of the Alan Turing Institute, and Director of Research Innovation at Edinburgh College of Art. She is co-director of the AHRC programme BRAID (Bridging Responsible AI Divides) and has been investigator on over £17m of externally-funded projects (EPSRC, ESRC, AHRC, Centre for Digital Built Britain, DCMS). Ewa was a founding member of the Scottish Government's AI Alliance and an expert adviser to DCMS.

Generative AI & Journalism

A rapid risk-based review

Dr Bronwyn Jones
Dr Rhianne Jones
Prof Ewa Luger

Cite as: Bronwyn Jones, Rhianne Jones, and Ewa Luger, “Generative AI & Journalism: A rapid risk-based review”, June 6, 2023.

This report is necessarily reflective of the time of writing. The authors would appreciate any feedback and suggestions for additions or amendments.

Please contact Bronwyn.jones@ed.ac.uk or Rhia.jones@bbc.co.uk

BBC R&D

THE UNIVERSITY of EDINBURGH

This work has been supported by the PETRAS National Centre of Excellence for IoT Systems Cybersecurity, which has been funded by the UK EPSRC under grant number EP/S035362/1