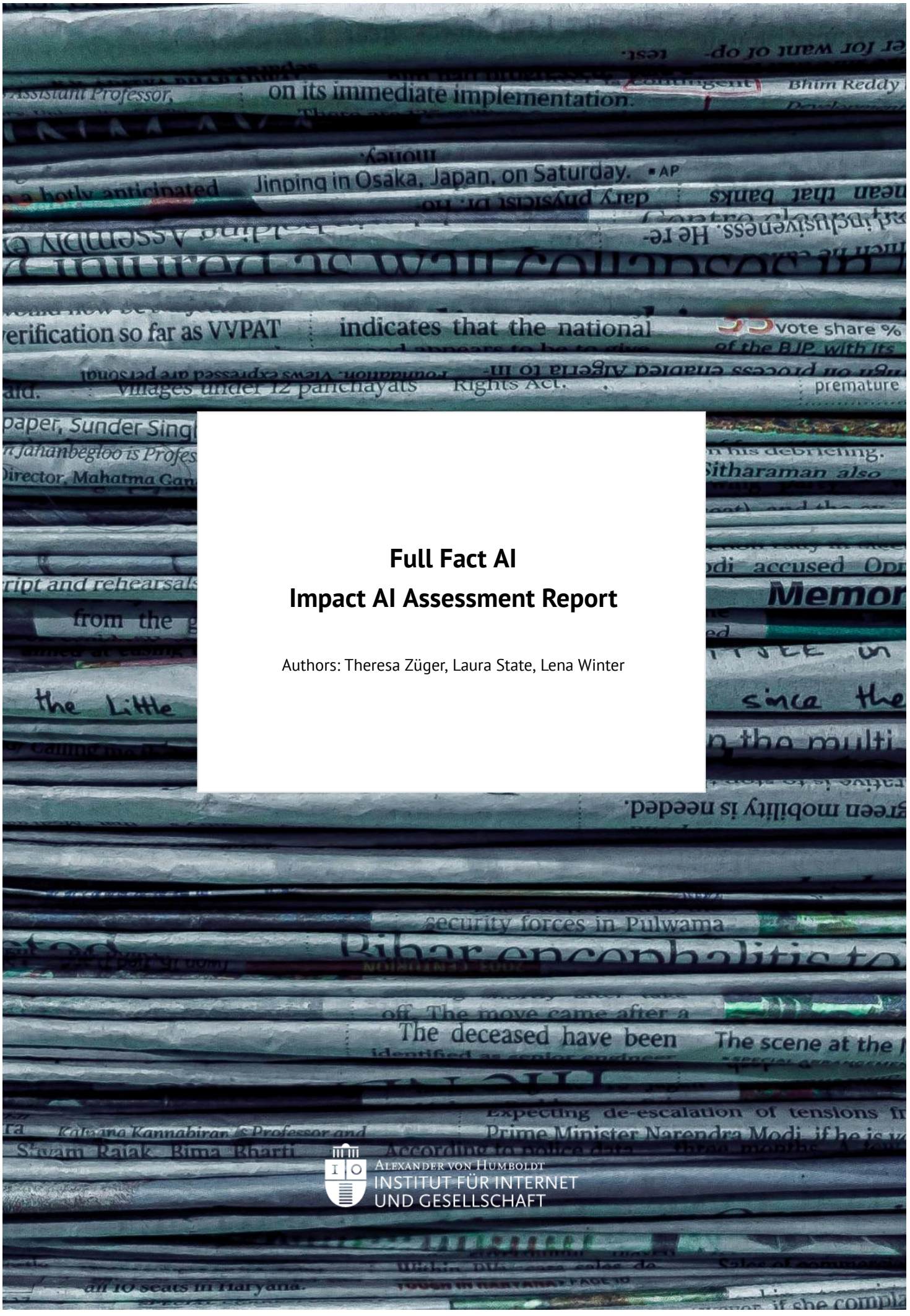


Full Fact AI
Impact AI Assessment Report

Authors: Theresa Züger, Laura State, Lena Winter

Alexander von Humboldt
INSTITUT FÜR INTERNET
UND GESELLSCHAFT

Full Fact AI

Impact Assessment Report

Dr. Theresa Züger
Alexander von Humboldt Institute for Internet and Society (HIIG)

Dr. Laura State
Alexander von Humboldt Institute for Internet and Society (HIIG)

Lena Winter
Alexander von Humboldt Institute for Internet and Society (HIIG)

Citation

Züger, T., State, L., & Winter, L. (2026). *Full Fact AI. Impact AI Assessment Report.*
https://doi.org/10.5281/zenodo.23884929

Acknowledgement

The Impact-AI project is a transdisciplinary research project carried out by the Alexander von Humboldt Institute for the Internet and Society in collaboration with the civil society partners Greenpeace Germany and Economy for the Common Good. The project is funded by the Volkswagen Foundation as part of the Change! Fellowship Programme. The authors would like to express their gratitude to the members of the Impact-AI Project Council for their valuable feedback during the methodology development process, as well as to our project partners. We would also like to extend our sincere thanks to Sadiem Youssef who supported this report as a student assistant and created the visualisations contained here. Finally, we would also like to thank Dr. Sami Nenno for his feedback and expertise in the field of fact checking.
Cover Image by Utsav Srestha on Unsplash.

Full Fact AI

Impact Assessment Report

Berlin, November 2026

Table of Contents

1 Introduction **4**

2 Description of Full Fact AI **5**

3 Assessment of Full Fact AI **10**

3.1 How does the assessment work? 10

3.2 Overview of the assessment results 12

3.3 Cluster 1: Impact 14

3.4 Cluster 2: Governance 27

3.5 Cluster 3: Social sustainability 33

3.6 Cluster 4: Environmental sustainability 44

3.7 Cluster 5: Economic sustainability 47

4 Conclusion and recommendations **50**

References **54**

Appendix **56**

1 Introduction

The transdisciplinary research project Impact AI analyses projects, organisations and initiatives that use AI with the goal of contributing to sustainability and the public interest. The research is conducted by the Alexander von Humboldt Institute for Internet and Society (HIIG) in a transdisciplinary partnership with Greenpeace and Economy for the Common Good.
To assess AI use cases, the Impact AI project developed the Impact AI Method, an impact assessment method to systematically evaluate the actual contribution of AI use to sustainability and the public interest. At the core, the method consists of a survey of technical details of the AI system, as well as several interviews with different project leads and, if possible, additional external perspectives. More information on the project and method can be found in the Impact AI FAQ. Additionally, the templates and materials of the Impact AI Method are available here. The assessment depends on 26 evaluation criteria to analyze the findings gathered through the interviews and survey. Participation in the impact assessment is voluntary and no label will be awarded for it.
In this report, we present our findings on the "Full Fact AI" project, developed by the UK-based organization Full Fact. The report is based on four interviews conducted between October 2025 and August 2026 with Andrew Dudfield, then Head of AI as organizational lead, as well as Dr. David Corney, Senior Data Scientist, as technical lead of Full Fact AI who have both agreed to be identified as key interviewees for the impact assessment. Documents and documentation provided by Full Fact, the website of Full Fact and Full Fact AI, as well as the analysis of a demo version of the Full Fact AI software were drawn on as additional sources. Additionally, the assessment was complemented by one interview with a fact checker from the fact checking organisation Africa Check, that uses Full Fact AI.

When reading this report, there are a few limitations to keep in mind. While we put emphasis on including additional media reports, documentation and existing research, many of our evaluations rely on self-reporting and self-descriptions of projects, which might be a source of bias. Also, despite our attempts to involve more feedback from external Full Fact AI users, we received feedback from only one other fact checking organization and conducted an interview with them.
Further, comprehensive evaluation of the models and datasets of Full Fact AI was not possible for our impact assessment, since Full Fact did not provide access to their models and datasets in use. While we understand this decision–which will also be discussed in this report–it limits the complete evaluation of the AI software in these regards. However, due to the incorporation of different variables, dimensions and methods into the assessment approach, this technical opacity is only a minor constraint for the comprehensiveness of the impact assessment framework.

2 Description of Full Fact AI

Profile	
AI system	Full Fact AI
Organisation	Full Fact (charity and non-profit)
Registered in	England, Wales
AI tool deployed since	2016
Number of users	45 fact checking organisations (as of 2026)
Funding/Partnership of AI tool	Diverse funding (Full Fact - Funding), incl. Google.org Charitable Giving Fund, Meta (Third Party Fact Checking Programme)

Full Fact AI is developed by the non-profit company and registered charity Full Fact, based in London, UK. Full Fact AI constitutes a software of "*AI powered fact checking tools*" (Website Full Fact AI), which was created in 2016 and is designed for fact checkers to support the identification of misinformation or potentially harmful information. The stated goal is to contribute to a "*better information environment*" (Andrew Dudfield, Head of AI, Full Fact). According to Full Fact, this is important, as "*[s]haring accurate information helps to build a society where better decisions are made, trust is earned, and democracy thrives*" (Website Full Fact).
However, Full Fact does not seek to automatize fact checks through Full Fact AI. Eventually, the final decision about whether a piece of information should be considered misinformation or harmful is always with a human fact checker.
According to Full Fact, a fact check at Full Fact involves the following steps:
1) A decision is made on what information the fact checkers will check. Here, the Full Fact AI software plays its crucial role. It supports fact checkers in monitoring media and identifying checkworthy claims made by public figures, politicians, as well as in viral online content. 2) Full Fact gives the claimant the 'right to reply', meaning they reach out to them and ask for their sources and information. In the next step, 3) identified potential misinformation is verified. This includes journalistic work and, depending on the complexity of the issue, consulting with experts and conducting interviews. The outcome is a conclusion about the validity of the checked claim. These conclusions come in the form of a fact check, which is usually published on the Full Fact website and social media channels. Writing a fact check can take some hours to several days, depending on the complexity of the issue. 4) Fact checks are reviewed by a number of other fact checkers and the editor of Full Fact and then 5) published. However, the work of Full Fact is not finished here: 6) Full Fact contacts the claimant again with the request to correct themselves.
It is worth noting that these steps are specific to Full Fact, at other organizations the process may look different.
As emphasized in the relevant sections and assessments in this report, the work of Full Fact goes beyond producing fact checks. Not only does it, among other things, include campaigning

work. Full Fact also offers training for various target groups, such as journalists, public bodies, or citizens, on a range of different topics, which are broadly categorizable under media literacy. With these actions, Full Fact aims to contribute to their stated goal of a better information environment and to encourage well-informed public debate.

Functionalities of Full Fact AI

The Full Fact AI software consists of four elements: 1) the home page; 2) the search functionality; 3) the matching tool ("alerts"); and 4) the live tool. The search, matching, and live functionalities utilize Gemini, they work in all languages Gemini supports. The front end of the Full Fact AI interface on the home page is currently available in English, French, and Arabic.
Input media can assume the form of text or audio (including the audio of a video). According to the interviews conducted, the type of data covered are from newspapers, podcasts, television, radio as well as platforms such as YouTube, X, Telegram, Instagram or Facebook. Which platforms Full Fact AI can monitor depends on the data-sharing and access granted by the respective platforms. API access is often limited. Generally, access usually requires either cooperation with the platforms or the payment of fees. In recent years, for example, accessing data from the platform X has become increasingly difficult, as the open API provided by the platform has been changed and access is now subject to high fees (Full Fact, 2025). Full Fact problematizes this limited access for fact checkers and researchers: "*If there's no API to access [the platform] or a limited API, that can cause problems. [...] Some platforms, for example, put a limit on how much we can access data from their platform. So we don't get everything.*"[1] (David Corney, Senior Data Scientist, Full Fact).

Beyond this access limitation, users of the tool can decide which media should be monitored. Each of the articles[2] monitored by the software is first classified by topic. Then, subsequently, the article is split into individual sentences and classified by claim type. This is done for two reasons: On the one hand, it enables users to select claims by topic, e.g. claims on

[1] As data access is a crucial issue for fact checkers for a balanced overview of relevant potentially misleading claims, Full Fact problematises the current status and demands: "Platforms must provide relevant data access to make fact checks effective and worthwhile. For fact checkers to target the most harmful and far-reaching claims effectively, they require cooperation and data-sharing from technology platforms. Access to back-end data helps us understand how misinformation spreads between individuals, communities and across borders. An agreement to provide researcher access to this real-time, proprietary data would help fact checkers make smarter and faster assessments on the dissemination of false and harmful information. [...] The more access we have, the more effectively we can identify and prioritise the most harmful content. Full Fact has long called for this level of data access. [...] The Data (Use and Access) Bill [...] is the [UK] government's vehicle for enabling researcher access to platform data. But who qualifies as a researcher is still undecided and it remains unclear whether fact checkers will be included under this definition. [...] Full Fact has consistently urged that fact checkers be explicitly included, ideally verified through recognised accreditation bodies such as the International Fact Checking Network (IFCN) and the European Fact Checking Standards Network (EFCSN). Without a clear legal definition, there is a risk that platforms will arbitrarily block fact checkers from accessing their platforms." (Full Fact, 2025).

[2] In the following, we use the term "article" to refer to any written text that may be checked. That also includes posts from social media.

"environment" or "migration", as well as by claim type, e.g. "quantitative" claims. On the other hand, not all types of claims can be checked, e.g. personal opinions. Accordingly, the classification helps to sort out such claims.

After the incoming claims have been classified, they are displayed on the **home page** of Full Fact AI. Here, users can also specify for which time frame these claims should be shown on the website. Additionally, claims are sorted by their "checkworthiness". Further, each claim is displayed with a link to the original data source so that the original article can be retrieved directly.
To determine the checkworthiness of a claim, Full Fact has built a system consisting of a query to an LLM and a linear regression. The query contains ten "yes" or "no" questions, assessing the checkworthiness of a claim (e.g. Is the claim about health? Is the claim a prediction about the future?). Along with the claim, the query is then sent to the LLM. The results are used as an input to a linear model that outputs a score (the checkworthiness, a number between one and five). In addition, there is a separate dataset behind the linear model that consists of 200 labelled sentences.

Using the **search functionality** of Full Fact AI, users can search among all claims extracted from the monitored media. By default, this includes all claims from the previous 24h. However, the timeline can also be extended to two days, a week, or up to a month. The search itself consists of a standard keyword search, that can be further refined (advanced search). If Full Fact AI can identify the person who made a statement, their name will also be displayed.

In addition to that, the **matching tool** checks whether any of the newly identified claims match an already existing fact check, i.e. whether this claim has already been checked in the past. The input to this component of Full Fact AI runs through the same pipeline as claims displayed on the home page, i.e. the checkworthiness component, topic and claim type classifier. If a potential match is identified, the pair of claim and fact check will be displayed to the users of the matching tool. Users can either accept or reject the match: Accepting means that they agree that the content of the newly identified claim indeed matches the fact check, while rejecting means that they do not agree to this match. The data created through this feedback is used for retraining and improving the system.
Claims that are deemed a potential match by the matching tool are only displayed on the page of the matching tool but not on the homepage. Also, claims are displayed with a link to the original data source, so that users of the system can easily retrieve the original article.

Finally, the **live tool** allows users of Full Fact AI to check audio sources (e.g. audio of parliamentary debates, podcasts, or youtube videos), which can even be applied to live media broadcasts (hence the name "live tool"). In the first step, Full Fact AI transcribes the audio source, utilizing Gemini. Then, the same pipeline as before is applied, i.e. the transliterated texts run through the checkworthiness component, and are classified by topic as well as by claim type.

On the interface of the live tool, the transliterated text is displayed alongside the original media, so that users can go back to the original source to check it.

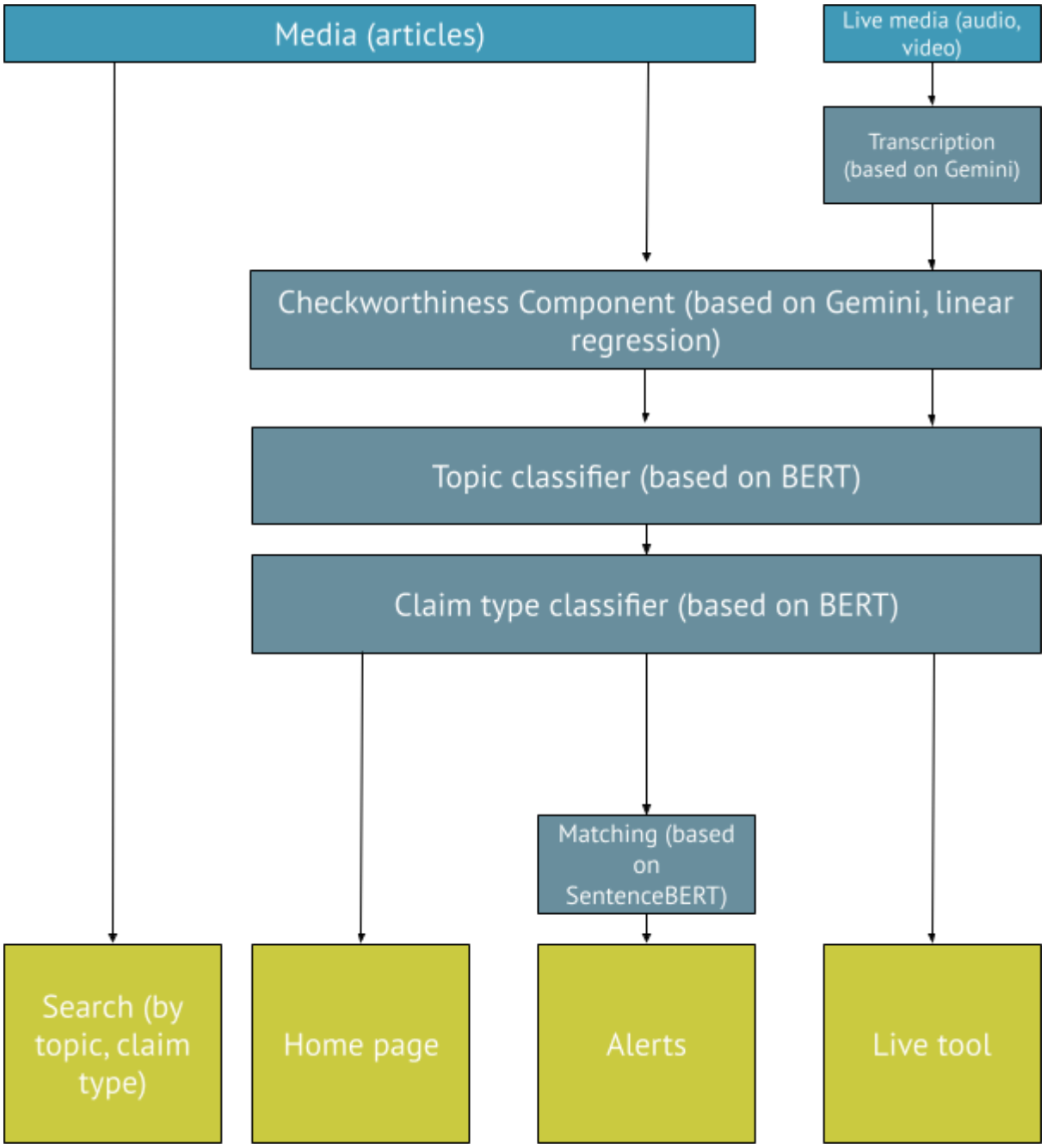


Figure 1: This figure provides an overview of the main components of the Full Fact AI system and how information is processed by the system. The figure should be read from top to bottom. Incoming media (articles, audio, or videos) are shown at the top while the four functionalities of Full Fact AI (search, home page, alerts, and live tool) are shown at the bottom. Components that involve an ML model are in dark gray color (transcription, checkworthiness component, topic and claim type classifier, matching)

Technical details of Full Fact AI

Topic classifier and dataset: The topic classifier is based on the BERT language model, which is short for Bidirectional Encoder Representations from Transformers and was introduced in 2018 by researchers at Google[3] (Devlin et al., 2019). Following the NewsCodes Media Topic taxonomy (source here, additional info here), the topic classifier classifies articles according to 17 categories[4]. The **dataset** for training and testing the classifier is a public dataset containing 10,917 news articles from January 2019 to December 2019, and was manually labelled. The dataset is available here, along with an accompanying paper and code.
No quantitative data on the **evaluation** of the topic classifier were provided to us.

Claim type classifier and dataset: Similar to the topic classifier, the claim type classifier is based on BERT and classifies articles according to 10 different claim types.[5] Each claim can potentially be assigned to multiple labels (claim types). However, most claims are only assigned to one. The **dataset** which was used to train and evaluate the claim type classifier was created by Full Fact internally. To this end, 3,000 sentences were collected from UK media and manually labelled by different individuals in the organization. To facilitate internal labelling, Full Fact created a tool:

> *"So we created a small tool for tentative, internal use to label sentences. And then, we gave this to everyone else in the organisation and just said, 'Can you spend half an hour going through every sentence we show you and just say, is this a quantity, is it personal, [or] is it a prediction?'." (David Corney, Senior Data Scientist, Full Fact)*

Matching tool and dataset: The matching tool is based on *sentence*BERT, which is a library, combined with a self-trained matching model. The input are two sentences which are converted into vectors, the output is the angle between these two vectors which is a measurement of their similarity. Some **evaluation** of the matching tool was provided to us which is discussed under 2. Impact achievements. The **dataset** used to train the matching tool was collected using crowdsourcing. For this purpose, Full Fact created a web interface through which participants could label the data. Participants were presented with pairs of sentences, i.e. a claim from the media and an existing fact check, and asked to indicate whether these two statements were similar or not. If a sufficient number of people agreed on the same label for a pair of sentences, this label would become the final label. Thus, the corresponding pair of sentences was then taken out of the exercise. The pairs of sentences and their labels that were selected for the crowdsourcing were taken from an earlier version of the AI tool. Originally, the data came from UK media. The crowdsourcing exercise took place for about two months.

[3] See also: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Accessed: 2026-07-13.
[4] See table 3 in the appendix displaying the categories.
[5] See table 4 in the appendix displaying the different claim types.

Live tool: The live monitoring tool uses Gemini to transcribe audio or video materials. No information on transcription quality or its evaluation was provided to us. The Google Cloud documentation does not list any information on the performance of the Speech-to-Text function either (Chirp models, see here). However, some external numbers can be found here.

3 Assessment of Full Fact AI

This section presents the results of the Impact-AI-Method applied to Full Fact AI. We present our analysis based on the 26 assessment criteria, highlighting areas in which the AI project performs well and also potential for improvement.

In 3.1, we explain how our assessment works, while we provide a summary of our results including a visualization in 3.2. Details of the assessment are given in 3.3-3.7.

3.1 How does the assessment work?

In total, our assessment framework contains 26 criteria. 24 of these criteria are evaluated on a scale of five while two criteria were assessed by drawing on and combining the other criteria, making them two meta-level criteria.

For the assessment of these two meta-level criteria, the following assessment questions are answered:

- **Defined goal in connection to public interest and sustainability**: Does the described goal of the project create a comprehensible connection to conceptions of public interest and/or dimensions of sustainability?
- **Impact for sustainability and/ or public interest**: What is the relation between the AI project and sustainability and PI from a holistic perspective? Does the overall impact of the project justify the claim to support public interests and/or sustainability?

For these two criteria, the assessment is written only and is therefore not rated on a scale or included in the visualization.

The other 24 criteria are assessed on a scale of five. Each criteria is matched with a corresponding assessment question. The answer to the assessment question is the rating on a scale on which the following five ratings are possible:

Insufficient	Weak	Ambivalent	Satisfactory	Exemplary

Table 1 displays the 24 criteria which are part of the visualization below and their corresponding assessment question. The answer of the assessment question is the rating on the scale described above.

Table 1: Catalogue of criteria with corresponding assessment questions.	
Criterion	**Assessment question**
Cluster: Impact	
1. Objective and realisation	Can goals, outcomes and outputs of the AI project be defined clearly and plausibly following a clear impact logic?
2. Impact achievements	Is there convincing evidence supporting the claim that the project has achieved envisioned outcomes?
3. Relevance of impact achievements	Are these outcomes relevant achievements in the given targeted problem area?
4. Potential harm and mitigation	Does the AI system in any way potentially cause or exacerbate harm, whether individual or collective, and including harm to social, cultural, economic, natural, and political environments and how is this harm prevented or mitigated?
5. Risk and harm management	Are there appropriate procedures to assess and monitor potential harms?
Cluster: Governance	
6. Necessity and proportionality	Is the use of an AI system, including the specific AI method(s) deployed, justified, necessary and proportionate in the given context and does it not exceed what is necessary and proportionate to achieve envisioned outcomes?
7. Accountability	Are there clear roles of responsibilities to ensure accountability for the AI system, its impacts and liabilities?
8. Decision processes and design	Is the design and decision-making process for the AI system reflecting principles of public interest and sustainability?
9. Inclusion, participation and co-design	When designing, deploying and using AI systems, does the project take an inclusive, interdisciplinary and participatory approach?
Cluster: Social sustainability	
10. IT Security	Are there organizational responsibilities and measures to identify, address and mitigate IT security risks throughout the AI system lifecycle to prevent where possible, and/or limit any potential or actual harm?
11. Sovereignty and power relations	Does the project contribute to shifting power in a democratic way (e.g. to civil society or marginalized groups) and support inclusive use and equal distribution of benefits and harms of the AI system?
12. Accessibility	Does the project convincingly describe efforts or create measures to create accessibility?
13. Non-discrimination	Does the project take appropriate measures to prevent bias, discrimination and stigmatization of any kind, in alignment with fundamental rights?
14. Data protection	Are the rights of data subjects respected, protected and promoted throughout the lifecycle of the AI system according to applicable laws?
15. Human oversight	Does the design of the system ensure human oversight and human capability to oversee the overall activity of the AI system?

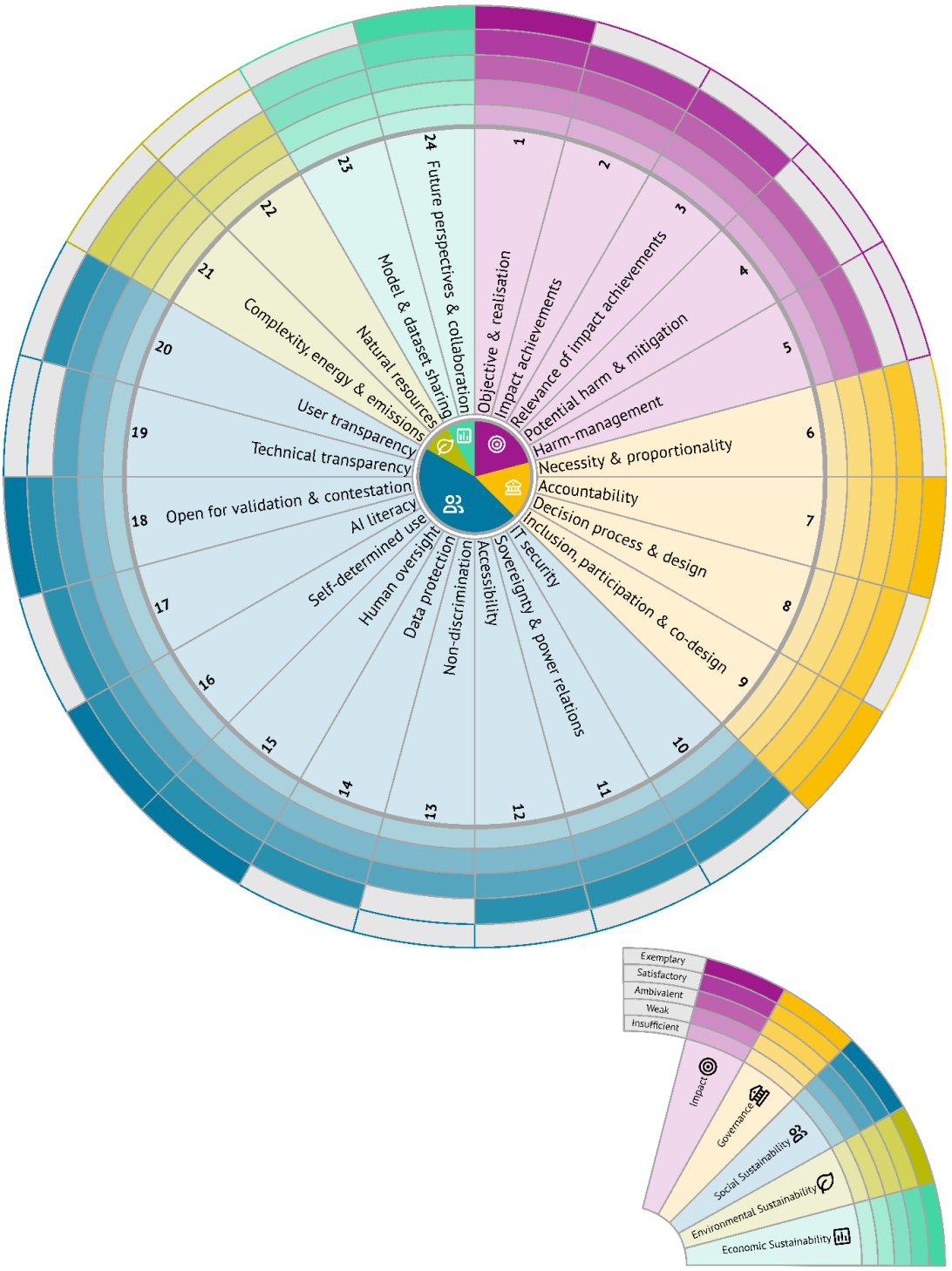
Table 1: Catalogue of criteria with corresponding assessment questions.	
16. Self-determined use	Do users make a self-determined choice when and how to use the system?
17. AI Literacy	Does the project describe efforts or concrete measures to improve AI literacy of people directly or indirectly impacted by the AI system?
18. Open for validation and contestation	Does the project provide sufficient transparency and openness for evaluation?
19. Technical explainability	Is technical explainability ensured, meaning that outputs by an AI system can be understood and traced by human beings?
20. User transparency	Is there public information about the AI system that is understandable to the user of this system? Are individuals meaningfully informed about the AI system that they use, when a decision is informed by or made based on AI algorithms?
Cluster: Environmental sustainability	
21. Complexity, energy and emissions	Does the project sufficiently consider the complexity, energy and emissions of the AI system, including, but not limited to, measures to reduce these?
22. Natural resources	Does the project sufficiently consider the resource use of their AI system?
Cluster: Economic sustainability	
23. Model and dataset sharing	Are the model and the dataset publicly shared, or are there good reasons not to share them, and did the project take measures to make it easy to use them?
24. Future perspective and collaboration	Are there efforts to establish collaboration and to sustain the project and the fulfilment of its objectives in the future?

3.2 Overview of the assessment results

The visualization in Figure 2 gives an overview of the assessment results. It displays the 24 rated criteria allocated to the five assessment clusters. The map legend shows the rating levels, as well as how clusters are assigned to colours and icons[6].
One can say, the darker the shade the better the AI project was rated in regards to the criterion; hence, the fuller the circle the better the AI project was rated overall.
Nevertheless, in order to draw conclusions about the AI project, one should not rely solely on the visualisation, but should also take into account the assessment questions and, above all, the qualitative evaluation for each criterion.

[6] All icons by Boxicons on Unsplash.

Regarding the 24 criteria that were evaluated on a scale of five, we obtained the following: seven criteria were assessed with “Exemplary”, twelve criteria were assessed with “Satisfactory”, and five criteria were assessed with “Ambivalent”. None of the criteria received the label “Weak” or “Insufficient”.



3.3 Cluster 1: Impact

Meta-level criterion: Defined goal in connection to public interest and sustainability

Assessment question
Does the described goal of the project create a comprehensible connection to conceptions of public interest and/or dimensions of sustainability?

Our understanding of public interest is closely connected to democratic values. What constitutes public interest depends on deliberative processes in which the interests of the public are collectively negotiated. Public interests cannot be found universally, but still manifest in legislative expression of common interests of societies and publics. In the simplest understanding, the concept of 'public' interest is often seen as the opposite of 'individual', 'private' or 'group' interests and aims for collective well-being.[7]
In our multidimensional understanding of sustainability, the concept incorporates environmental, social, and economic dimensions, with the overarching aim of securing equal or improved living conditions for present and future generations. The 2030 Agenda for Sustainable Development and the 17 Sustainable Development Goals (SDGs) represent a holistic set of goals that integrate the three dimensions of sustainability (UNGA, 2015). Democracy is inherent to many of these SDGs, in particular goal 16, 'Peace, Justice and Strong Institutions'. This goal is clearly linked to democracy as it proposes to "[e]nsure access to information" as well as "[e]nsure responsive, inclusive, participatory and representative decision-making at all levels".[8]

We see clear connections between the goals of Full Fact and of the software Full Fact AI to objectives of public interest and of sustainability.
Full Fact as an organization describes itself as an "*education charity*" with the described goal of a "*better information environment*". This is linked to democracy as "*[s]haring accurate information helps to build a society where better decisions are made, trust is earned, and democracy thrives*" (Website Full Fact AI).
Full Fact claims that:

> *"Daily insights help fact checkers to take proactive steps to challenge false claims, limit their impact and harm, and keep their audience better informed. Without such efforts, we are in danger of reaching a place where no one believes anything anymore. That would be bad for our democracy, and for our understanding of what it means to be a functioning modern society."* (Full Fact, 2025)

Further, the goal as an education charity is to encourage good public debate: "*So we're trying to ensure that people have the tools and skills and knowledge needed to help them navigate the information environment that we're going to improve.*" (Andrew Dudfield, Head of AI, Full Fact)

[7] More on our understanding of the public interest see Züger and Asghari (2023).
[8] See Goal 16: Promote just, peaceful and inclusive societies. Accessed: 2026-07-06.

Full Fact as an organization describes their motives for their work with the importance of high quality and trustworthy information ecosystems in the public realm. Conversely, false information can develop harmful consequences for this ecosystem. Trust and truthfulness in public debates and media are crucial for democratic public debate. This is enhanced by empowered and informed citizens which is also supported by Full Fact's engagement in offering training and education. All this therefore also matters to public interest and sustainability as the efforts of Full Fact contribute to peace, justice, and strong institutions.

> *"We're talking about harmful information and usually we're defining that by, 'Is it false? Is it believable? And is there an action that somebody could take from that information?' [...] So harm doesn't necessarily mean harm to life, it can mean harm to democracy, it [can] mean harm to business, it could mean harm to society. There's lots of different definitions of that." (Andrew Dudfield, Head of AI, Full Fact)*

One step towards a better information environment, which is the described goal of Full Fact, is to provide Full Fact AI, which supports fact checking organizations and in particular fact checkers in their work.
We see, however, that Full Fact AI can only be effective and impactful when it is embedded in further efforts and an active organizational environment. A fact which Full Fact highlights and realizes within its own work.

> *"I think people need access to good information. I think fact checking is not about the idea of 'if we can just do more fact checking, everything can be better'. [...] You can't scale fact checking to just say 'there's more good information than bad'. But I think what it does do, is hold people to account. It shows that change is possible. And it shows the behaviours and techniques that are used to allow us to do things like media literacy. It allows us to go into schools and talk about these kinds of things. And it shows that we can use it as evidence to start a bunch of other conversations that are in the public interest." (Andrew Dudfield, Head of AI, Full Fact)*

Full Fact AI alone will not solve the problem of the increasing spread of misinformation. Also, it is neither supposed to replace fact checkers nor automatize fact checking itself. It does, however, contribute to democratic public debate.

Meta-level criterion: Impact for sustainability and/ or public interest

Assessment question
What is the relation between the AI project and sustainability and PI from a holistic perspective? Does the overall impact of the project justify the claim to support public interests and/or sustainability?

This criterion offers our reflection on the contribution from the AI project to public interest and sustainability. This reflection can be seen as a meta reflection of various aspects we include in the assessment of the other criteria of the report. We highlight only the selection of aspects which we see as most relevant for the system's impact in regards to its contribution to public interests and sustainability.

Overall, we see a contribution of the Full Fact AI project to the goal of the organization Full Fact, which is to create a "better information environment". Accredited fact checking organizations like Full Fact contribute to resilient democratic discourse in media and beyond, and thereby to the quality of public discourse (Amazeen, 2017; Wintersieck, 2017). This positive contribution can also be seen in the context of SDG 16 (Peace, Justice and strong institutions), as it contributes to the trustworthiness of the media. There is evidence that factchecking reduces misperceptions, agreement with misinformation and improves factual accuracy (Cobb et al., 2013; Nyhan et al., 2020; Porter et al., 2019; Walter et al., 2020). Though, whilst fact checking does lead to the acceptance of a correction by an individual, it is not sufficient in removing the effects of this misinformation on their beliefs and behaviour (Johnson & Seifert, 1994; Lewandowsky et al., 2012; Thorson, 2016). More concerningly the "directionally motivated reasoning" theory puts forward evidence that individuals are more accepting of misinformation that is consistent with their prior worldviews and are worse at accepting corrections which contradict it (Flynn et al., 2017; Walter et al., 2020).
In addition, even if a fact check is written, it is unclear how it will actually reach its audience as simply publishing in article format on a website, will not guarantee that individuals and in particular individuals at risk of accepting some misinformation will encounter the relevant correction. Full Fact is aware of this challenge and seeks to meet it: "*We're always thinking about how we get it [the fact check] out far and wide*" (Andrew Dudfield, Head of AI, Full Fact). To increase their reach, the fact checks are integrated in different systems and databases (according to the interview, these include the Google scheme or a central database for fact checks as well as tagging fact checks with metadata so it can be easier connected with other websites).
In summary, while fact checking alone is not the one solution to the combat of misinformation and work towards a better information environment (as also voiced by Full Fact themselves), there is evidence that fact checking is an important practice and could be used to defend journalism against discrediting (Pingree et al., 2018), improve misinformation discernment (Guess et al., 2020) and increase self-perceived misinformation identification (Goyanes et al., 2025). We can conclude that fact checks work towards building resilience and accountability.

The fact that 45 fact checking organizations around the world use and pay for the AI system of Full Fact is a sign that it contributes to their work in a meaningful way. While Full Fact assesses its organizational impacts in reports with meaningful transparency, the organization so far does not assess the exact impact of the system specifically (e.g. how many hours of work are saved? What is achieved with this additional time? How does the system influence the quality of work?). Therefore, no concrete isolated indicators for the system's impact could be shared with us. However, we inferred them more indirectly. Within Full Fact, and for the external fact checking organization we interviewed, the meaningful contribution of the system has been confirmed. It was described as time-saving, especially in the transcription function, as well as in supporting the quality of the fact checking process. This is because the tool has the potential to provide a quicker and more comprehensive overview of potential misinformation than the manual research of fact checkers in the same amount of time. The responsibility to realize this potential, however, lies with the individual fact checking organizations using the system, who themselves make choices over media input, topic selection, and claim selection.

In regards to the environmental sustainability of the Full Fact system, self-reported numbers suggest a relatively low energy consumption of the overall system. However, not all data to confirm this low consumption was fully transparent.[9]

From a public interest and sustainability-focussed perspective, the use of Gemini can be criticized for reasons such as the problematic resource consumption of Gemini in training and inference in general (Jegham et al., 2025) and the exploitative working conditions under which presumably all generative AI models are trained and maintained (Aapti 2026).[10] At the very least, this use presents a moral dilemma: while serving democratic publics and their discourses (in many regions of the world, including especially many global majority countries) other communities and their rights are harmed. While we understand the argument of technical capability, using Gemini instead of open source models supports the perpetuation of a power asymmetry and dependence by indirectly supporting the structures of Big Tech empires.

We see room for improvement in regards to the transparency and testing of the models. This is especially for different language scenarios where performance often varies, as well as in training with external users to raise awareness for potential harms in their specific contexts, particularly given the diversity of regions the system is operated in.

1. Objectives and Realisation

Assessment					*Exemplary*

Assessment question

Can goals, outcomes and outputs of the AI project be defined clearly and plausibly following a clear impact logic?

[9] Neither from Full Fact regarding the use of various AI models nor from Google, as the operator of Gemini, one of the models used.

[10] See also: How hungry is AI? for live data on the consumption of several large models, incl. Gemini.

For the assessment of this criterion, we outline the measurable deliverables (**outputs**), how these contribute to **outcomes,** and how all this feeds into the envisioned long-term **impact** of the intervention. This conceptualization draws on the concept of a theory of change. A theory of change is a framework that explains how a given intervention, or set of interventions, are expected to lead to a specific development change, drawing on a causal analysis based on available evidence (UNDG, 2017). In the best case scenario, concrete outputs lead to achieving envisioned outcomes and thereby contribute to a project's overall social goal. Hence, there should be a convincing impact logic between these levels of impact.

Full Fact AI is embedded in the work of Full Fact. When outlining the envisioned and real impact, it is difficult to distinguish between the impact of the work of the organisation as a whole, and the impact of Full Fact AI alone. Describing the impact levels of Full Fact AI can only be done in context of the envisioned societal goal and the work of the fact checking organization.

Full Fact AI is "*designed to save time, money and effort by identifying the most critical misinformation*" (Website Full Fact AI) with the objective to scale fact checking. It focuses on the issue of the rapid spread of misinformation.

In the impact logic of Full Fact, the use of Full Fact AI is only one of the outputs to contribute to different outcomes and the overall societal goal.

The embeddedness of Full Fact AI and Full Fact (and probably any other fact checking organization using the AI software) became apparent during the interview.

This attitude is something which we see as positive, since it is most realistic: impact and change comes not through the AI tool alone, but is embedded in further action. It is a realistic estimation of what the AI tool can do, achieve, or contribute to.

Also, Full Fact's description of their goal, outcomes, and outputs reflects the clear embeddedness and entanglement of both the AI software, and Full Fact's work in general. Considering the work of Full Fact as a whole, the goal, outcomes, and outputs can be aligned with a clear theory of change and impact logic.

The **societal goal** that Full Fact envisions to achieve with Full Fact AI is a "*better information environment*" (Andrew Dudfield, Head of AI, Full Fact).

To reach the goal of a better information environment, Full Fact is aware that it is not only about publishing fact checks and providing more good qualitative information, but that there is a need for further action beyond the fact check.

> *"Fact checking is not at all about the idea of 'you publish a fact check and just assume the world's got better'. You know, that is not how it works, unfortunately. [...] So fact checking is the start for us. It's producing a piece of evidence. And then we're thinking about, how do we take that evidence and do something with it?" (Andrew Dudfield, Head of AI, Full Fact)*

Therefore, the **outcomes** which are intended by Full Fact can be described as the following:

Outcome 1: Persons who spread misinformation need to be held accountable for their action. In the case of Full Fact, this mostly refers to politicians, as Full Facts work focuses on UK politics. Therefore, one of the outcomes is to *"raise standards in public debate and create consequences for misleading the public" (Andrew Dudfield, Head of AI, Full Fact)*.
Outcome 2: Another intended outcome is to positively influence UK legislation and legislation of internet companies: *"So we are looking for impact in terms of we want to be able to directly say that legislation changed because of us and because of our fact checking" (Andrew Dudfield, Head of AI, Full Fact)*.
Outcome 3: Another step towards improving the information environment is to have an influence on the public. This is achieved through an influence on the people who consume the information published by Full Fact.

It becomes apparent that Full Fact works on reaching the set societal goals not only by their own work as a fact checking organisation, but also by making the Full Fact AI software accessible and usable for other fact checking organizations.

There are several **outputs** to fulfill the outcomes and to work towards the societal goal:

Output 1: Producing and publishing fact checks. The fact checking process is supported by the AI software.
The reason for and role of Full Fact AI in this impact logic is an important one: the main practical goal of the use of Full Fact AI is to support fact checkers and to facilitate and therefore scale their work. In the interview, the advantage of using Full Fact AI is explained in more detail:

> *"So, when we're doing this, we're trying to help out with that triage process, so identifying the most important things to fact check each day. We're also interested in finding repeats of things that we've previously fact checked, and we're interested in speeding up the process of fact checking. So those are the three things that we're trying to bring to the picture." (Andrew Dudfield, Head of AI, Full Fact)*

Although the fact check itself is a small piece in the impact logic, it is a crucial one: *"I think the fact check is the atomic unit of our work, but that is not often the thing that we get the most impact from" (Andrew Dudfield, Head of AI, Full Fact)*.
Whilst the fact checking process is being facilitated faster, it is saving (time) resources on fact checking, which enables work on other outputs and outcomes to achieve impact.
Output 2: Request corrections: This involves the contacting of people who published misinformation and requesting to correct it.
Output 3: Scaling the outreach of fact checks. This includes the publication of fact checks not only on the Full Fact website but to integrate them in databases to achieve a high media reach.
Output 4: Advocacy work: This involves campaigning work for which Full Fact engages a team

in its organisation and involves changes in legislation, law, or policy. Also, Full Fact publishes recommendations for different target groups, for example governments or internet platforms, with the goal of improving the information environment.
Output 5: This involves (public) education and training efforts on different topics, in particular related to media literacy

We see a link between the different outputs, outcomes, and the final societal goal, and a clear impact logic. Whilst Full Fact AI is only a small part of this puzzle, we recognize the overall high reflection level within the theory of change guiding Full Fact's work, which is the argument for the *exemplary* (5/5) rating.

2. Impact achievements

Assessment				*Satisfactory*	

Assessment question
Is there convincing evidence supporting the claim that the project has achieved envisioned outcomes?

Part of assessing the impact achievements of Full Fact AI is to understand how well the software actually works. No quantitative evaluations were provided to us for the topic classifier or the live tool.
With regards to the claim type classifier and according to Full Fact, a manual evaluation of a small sample was carried out every month over a period of one year, with a focus on the 'quantity' type. A typical result was: precision 0.96, recall 0.95, F1 0.95. While this is a very good result, it should be noted that 'quantity' is only one of the claim types which can be classified by the software. For other claim types, such as correlations or opinions, the results of the evaluation may vary. According to Full Fact, however, the focus was on the category 'quantity', as this is the one most frequently used by fact checkers.
With regard to the matching tool, at the start of 2025, Full Fact carried out a manual evaluation of the tool on a data sample (we do not have information about the size of the data sample). The following table 2 displays the distribution of the quality of the predicted matches, which was tested in a manual setting as explained:

> *"To produce that table for evaluation, we took a set of claims that the model predicted were repeats of previous claims and examined them manually to decide which were actually matches, i.e. which predictions were accurate. We found that about 10% were exact repeats (perhaps differing in a word or two but with the same meaning) - that's what row 5 means. Nearly 66% of the predicted matches were not exactly repeating known matches but were making related claims (row 4). I don't have the exact figures to hand, but suppose we're checking 100,000 new claims per day against 30 recently published checks. Then there are 30 * 100,000 possible pairs of claims, and we want to know which pairs are actually saying*

the same thing. Of those 3m pairs, we show to the user just the top 20 or 30 strongest predictions, and of those, about 10% are exact repeats. So the 10% figure is the precision of the model; the recall is unknown because we don't know how many of those 3 million pairs (per day) are actually matches." (David Corney, Senior Data Scientist, Full Fact)

Table 2: Results of Full Fact's evaluation of their matching tool indicating the distribution of the quality of the predicted matches.	
1. Totally unrelated claims	0
2. Some overlap in meaning but just coincidental (e.g. both mention the same country or person)	12.77%
3. Claims on same broad topic but largely unrelated	10.64%
4. Claims are different but related to each other	65.96%
5. Repeats of the same (or very nearly the same) claim	10.64%

While no match was indicated for two completely independent claims, only around 10% of the matches indicated were actually repetitions of the same claim, shown by the manual check. This means that for the task of finding claims that have been made before the model is producing a relatively high number of false positives in regards to being a complete match. Many of the claims however do relate to each other, which the model wrongly interprets as full matches of claims. While this performance might seem unsatisfying, in practice it seems to offer value. In an interview with Africa Check, we were told that in their experience on average 50% of the matches flagged by the alert function are correct in indicating a previously checked claim. Africa Check described this performance of the tool as helpful to assist in the workflow of detecting repeated misinformation, as the claims are manually checked anyway and 50% out of millions of potential matches is still a helpful filter.

Further insights into the effect of the use of Full Fact AI on the work of the fact checkers were provided in the interview with Africa Check. Predominantly, the fact checker highlighted that in their perception the use of the AI software saves work hours. It was not specified in more detail, but the benefit of the saved time was described to increase impact since it allows them to focus on other tasks.
According to Africa Check, in their own work processes not all functionalities of Full Fact AI are used to the same extent. On average, each fact checker spends two hours a week using the different tools. The alert and live function were described as the most used tools for the fact checking organization. It was estimated that the latter had saved a fact checker four to six hours of work per speech that might be fact checked. Also the search function was described as useful, in particular because it allows one to filter for specific claim types. While being described as useful, the function is used less often due to the overall fact checking strategy of the organisation, not because of a lack in practical functionality as such.

For assessing the impact achievements, we must once again note the difficulty in distinguishing between the impacts of the organisation Full Fact, and the impacts of the Full Fact AI alone. Full Fact has established monitoring procedures for the impact of their work. There is an existing strategy to monitor changes and progress towards internally defined goals. However, again, it must be noted that these account for all activities and all of the work that Full Fact is doing, not only for the use of Full Fact AI.
We positively rate their high awareness that the Full Fact AI software alone cannot reach the envisioned societal goal of a better information environment, but rather just plays a part in this process. Nevertheless, there is no holistic description of the effects of the AI tool on the users, in this case the fact checkers. On their website, Full Fact states that the use of AI has increased the number of identified checkable claims by 1,000, leading to 100,000 detected claims a day (Dudfield, 2021). However, the direct effects of the AI software on the fact checker, their daily work routine, and their organizations remains unreported. We see it as an implicit expression of added value, that 45 fact checking organisations in 30 different countries are using their AI software and in most of these cases are paying a license fee for its use.

To answer the assessment question for this criterion: The objective of Full Fact AI is to facilitate and scale the work of fact checkers. Africa Check and fact checkers within the Full Fact AI staff reported on the benefits of Full Fact AI for their fact checking activities and also the many users of Full Fact AI are indicators that the software supports the work of fact checkers. But besides numbers on the increase in checked claims, clear evidence on the more detailed effects on workflows for fact checkers are missing. The direct impacts of Full Fact AI and its different functionalities should be monitored and reported more holistically. Examples of relevant questions are: How many claims suggested in the search functionality were checked? How accurate are the matches found with the respective tool? Or, more generally, to what extent does the use of the AI tool lead to a reduction in workload (work hours saved) but also unforeseen changes in workflow? Based on this, we rate this criterion with *satisfactory* (4/5).

3. Relevance of impact achievements

Assessment				*Satisfactory*	

Assessment question
Are these outcomes relevant achievements in the given targeted problem area?

Fact checking is one important practice to contribute to a better information environment and to serve the democratic public. However, as elaborated under “Meta-level criterion: Impact for sustainability and/ or public interest” fact checking by no means can be seen as the only solution against misinformation but is one piece in a multitude of measures. We therefore can agree that Full Fact AI is improving and supporting the work of fact checkers and is thereby also relevant for the overall goal of a better information environment.

The relevance of the Full Fact AI software for fact checkers lies in its ability to scale the work of fact checkers, particularly with regard to monitoring the media landscape and deciding what to fact check to support Full Facts overall mission.
According to an interview with Full Fact, the idea to develop Full Fact AI was based on the challenges and needs for fact checking, highlighted by fact checkers at Full Fact (see also under 22. Inclusion, participation and co-design). The task of finding the most relevant claims is described as the one to which fact checkers usually dedicate the most of their time. It becomes even more challenging in times when misinformation is easily generated and spreads quickly.

> *"[...] [I]dentifying what to fact check is probably the place where we spend the most time in what we're doing, because that has been about media monitoring. So, when we're talking about podcast and YouTube and short video and newspaper websites and social media and parliament and TV, [...] that is where technology can come in, because it can help scale our ability to listen." (Andrew Dudfield, Head of AI, Full Fact)*

Africa Check confirmed the relevance of the Full Fact AI software for their work through the work hours saved. Overall, they indicated that the software's functions are helpful and that they would recommend it to other fact checkers as well. As Africa Check was involved in the initial development of Full Fact AI, they have free access to the tool. However, it was said that they would also be willing to pay for the use of Full Fact AI, confirming the relevance of the use of the AI tool for their organisation's work. We recognize that the alternative of fact checking without the use of such an AI tool comes with greater disadvantages than the mindful application of the software, in particular when considering the growing challenges regarding the increasing flood of (mis-)information, also due to generative AI (OECD, 2024).

In general, we consider the impact of Full Fact AI to be relevant for fact checkers and their work. When assessing Full Fact AI as a standalone AI software, we therefore assess the relevance of the tool to achieving the envisioned goal with *satisfactory* (4/5) . We acknowledge, however, that the overall impact of the work of Full Fact as an organisation might achieve an even higher rating.

4. Potential harm and mitigation

Assessment			*Ambivalent*		

Assessment question
Does the AI system in any way potentially cause or exacerbate harm, whether individual or collective, and including harm to social, cultural, economic, natural, and political environments and how is this harm prevented or mitigated?

Full Fact showed awareness about risks and limitations of Full Fact AI. Two specific risks for harm were mentioned proactively: The first one is related to ill-intended actors who might use

the Full Fact AI tools with harmful and non-democratic intentions, "*if our tool was used by an oppressive government, it could be used to find troublemakers*" *(David Corney, Senior Data Scientist, Full Fact)*. To prevent this scenario, Full Fact makes its AI software available only to registered fact checking organizations. The second risk proactively mentioned, relates to language and variations in the performance and quality of Full Fact's AI results depending on the language. Whilst Full Fact is aware of this and assesses the likelihood of this risk as high, it sees only low severity as users would simply just not rely on the tool for their editorial choices if the outputs prove to be faulty, which would show in the fact checking editorial process.
In regards to technical prevention of biases and discriminatory outcomes of the model, particularly concerning aforementioned language bias and a bias relating to the different input media (the AI system operating better for some media than others), Full Fact did not share any results of an evaluation with respect to these biases or information that such an evaluation exists. Also, no further mitigation strategies are deployed. For more details, see also "13. Non-Discrimination".

Overall, there is awareness for existing bias, and beyond diverse perspectives in the team, transparent debate internally seems to be an ongoing practice.

> *"It's a question that we ask ourselves constantly. We want to say that fact checkers operate without bias. We want [to] say that AI operates without bias, but those things are not true. I mean, everybody has their own bias, their own things that they bring to information. And we have to be mindful of that." (Andrew Dudfield, Head of AI, Full Fact)*

This is also a reason why transparency plays an important role for Full Fact and its AI tool:

> *"[W]e're making sure that the tools we're giving to people are transparent in how they're produced, how they are working, how they commit to decisions so that people in different environments can choose whether they place their trust in that or not." (Andrew Dudfield, Head of AI, Full Fact)*

Besides these issues mentioned by Full Fact, we found additional potential scenarios for fundamental rights risks, with different degrees of severity and likeliness for Full Fact AI. These scenarios were developed in collaboration with experts from the "European Center for Not-For-Profit Law", and are part of a Fundamental Rights Impact Assessment (FRIA)-inspired triage that we conducted. We communicated these to the project staff of Full Fact and discussed their likelihood, severity and existing mitigation measures.

For many scenarios, the most important strategy for prevention or risk mitigation is to rely on the professionalism and journalistic skills of the human fact checkers, whilst generally recognizing the importance of human oversight. Risks we identified and discussed, are for instance the risk of fact checkers becoming overly reliant on Full Fact AI, which, in the case of media monitoring, can lead to not checking any other sources than the ones displayed by Full Fact AI. Or, in the case of the matching or the checkworthiness tool, to blindly rely on the

outputs of the tool without properly verifying them. All this could result in potentially missing out more important and harmful information which is not checked, in certain media sources being monitored more closely than others, or in certain actors being checked (and accused) for misinformation disproportionately more often.

> *"So, the tool has potential to cause harm because we're suggesting that these things need to be fact checked. [...] That could be incorrect. We could be saying this thing is really important and there is something that's more important that we haven't spotted. People could potentially outsource responsibility [to the AI system] or consider this to be guiding them to something." (Andrew Dudfield, Head of AI, Full Fact)*

This means that mitigation of this potential harm is a widespread awareness amongst fact checkers, that they should never overrely on the outputs of an assistive system, which we recognized in this but also other statements.
Another harm that needs to be considered is the possibility of hallucination when using Gemini for transcribing the audio input for the live tool. As a first measure to counteract such possible hallucinations, inside of Full Fact AI the transcribed text is linked back to the original medium so that a user of the software can consult it when looking at the transcription. In the interview with Africa Check it was explained that the software makes it easy to verify matches between that which is generated by the live function, and the original source, since the transcriptions are directly linked to the original source. This facilitates double checking which was articulated to be a common practice.

Another scenario under consideration is the misuse or hacking of Full Fact AI. This may result in entities who could work out how to appear informative, despite actually being misinformative. To prevent this potential harm, Full Fact is closed source and does not publicly share their code (for more details see also under 23. Model and dataset sharing). In addition to that, IT security checks are periodically performed (for more details see also under 10. IT security).

Related to the risks of sensitive data protection infringement, scenarios are neither very likely nor severe, since Full Fact AI only processes already published information (for more details see also under 14. Data protection). According to their statements Full Fact also did a documentation of an internal data protection assessment.

Full Fact's own conclusion is that Full Fact AI as a tool itself does at the current moment not cause any significant harm or infringe upon fundamental rights, as potential harms are sufficiently mitigated. We agree to this overall assessment based on our current information. However, we did see that the process of a FRIA was insightful and could be communicated and repeated internally. Also, as Full Fact AI is delivered to external organizations for use, it is key to raise and share awareness of potential fundamental rights risk, e.g. in bad actor scenarios, especially since human oversight and the role of fact checker is the key mitigation to prevent harms from materializing. As we see room for improvement in this regard, we overall, assess this criterion with *ambivalent* (3/5).

5. Harm management

Assessment			*Ambivalent*		

Assessment question
Are there appropriate procedures to assess and monitor potential harms?

Full Fact stated to include the consideration of risks within their overall governance approach of the organization:

> *"We try to disperse the responsibility for those across the organisation. And then we have a portfolio governance approach as well. So we'll be looking across the organisation at where the risks exist, which we'll formally record as part of a process there." (Andrew Dudfield, Head of AI, Full Fact)*

If this process includes reflection of specific risks related to the AI system remains underspecified. At the moment, Full Fact does not carry out systematic risk assessments in relation to potential harm of Full Fact AI. Also, so far, no risk assessment findings from third party providers of the AI system (in this case Google's Gemini) were requested or evaluated. Risk assessments from foundational models can be an important point of reflection for relevant risks also for specialized applications, as in the case of Full Fact AI.

In one of the organizational interviews it was reported that the project conducts benchmarking tests with their model to assure high performance. While we received some information about the performance of the ML models trained by Full Fact, this information is very limited, both in terms of making a claim about the correctness of the system and whether sufficient testing of the models was done.

According to the interview, there are existing internal documentations on data protection processes and IT security.
Other than that, no monitoring of risks or potential harm exists and at the moment no monitoring of the effectiveness of mitigation strategies is performed in a structured way.
In conclusion, core risks and harms seem to be considered and managed internally. However, these processes could be improved with inclusion of more aspects and a more structured approach, which includes the mitigation strategies (also in respect to measures including external users) and the assessment of the effectiveness of mitigation measures. Therefore, we rate this criterion with *ambivalent* (3/5).

3.4 Cluster 2: Governance

6. Necessity and proportionality

Assessment				*Satisfactory*	

Assessment questions
Is the use of an AI system, including the specific AI method(s) deployed, justified, necessary and proportionate in the given context and does it not exceed what is necessary and proportionate to achieve envisioned outcomes?

Overall, the integration of different AI models into Full Fact AI appears necessary and proportionate to the task for several reasons.

From an application perspective, Full Fact AI is an appropriate solution for fact checkers since the idea to use AI was born out of the actual experiences and challenges of fact checkers and the design and functionalities were developed with them.

First of all, we find it convincing that automated claim classification as well as the other functions of the system are relevant improvements for fact checking, which speaks to the necessity of such a system within today's information environment. Full Fact made it very clear that the fact checking itself is not automated, even though on their website Full Fact AI is described as "*AI powered fact checking tools for everyone*" (Website Full Fact AI) which might be misunderstood as a more holistic functionality at first sight. The webpage however does explain the functions that are offered (monitoring, alerting for existing misinformation and transcription of media content for monitoring) which are not automating fact checks themselves. Full Fact AI is there to support fact checkers in their work who make the decisions in the end.

Full Fact convincingly explained why Natural Language Processing (NLP) technology is more capable than other technologies for identifying misleading or potentially harmful claims. Furthermore, Full Fact demonstrated awareness to use LLMs only where necessary (live tool) and to otherwise rely on the use of task-specific models and resources (BERT and SentenceBERT). Generally, Full Fact highlighted that, whenever possible, they try to use already available, task-specific models to build Full Fact AI rather than more general models that can also be used for other tasks. This generally minimizes resource use because models do not need to be built again and "oversized/too general" models are avoided.

We positively rate that for Full Fact the decision to use AI technology was not the starting point, but rather that the decision to use AI was generally made on a problem-oriented basis.

> *"[...] Our default is to think about: how do we do it without technology, and then work out where we do need to use technology. So I would say our approach is to use tech very lightly,*

because everything that we do, our whole thing, is about humans writing fact checks, and for that to be trusted." (Andrew Dudfield, Head of AI, Full Fact)

We view the decision-making and design of their infrastructure as a necessary application of NLP technology in this context.

We gained the impression that Full Fact is well aware of the capabilities but also limitations of the use of AI. Further, we view it positively that Full Fact recognises the specific potential of the AI software in the fact checking process and the points where fact checkers can use the tool as support, whilst at the same time acknowledging the necessity of human contributions for a meaningful fact check.
While fact checkers can use Full Fact AI to identify checkworthy claims which are potentially misleading or harmful, the rest of the process relies on human responsibility and journalistic quality. The decision-making about misinformation is never left to the AI software.

From a technical point of view, there is a lack of information about model performance and the datasets which is necessary information to assess the appropriateness of the technical implementation of the AI system and is needed to assess the overall necessity and proportionality of the AI system (Is the dataset appropriate to the task? Is the model appropriate to the task?). Limited information about the performance of the ML models that Full Fact trained (the two classifiers and matching tool, see also Section 2 "Description of Full Fact AI") was provided. Given this, from an outside perspective, it is not possible to rate the overall performance of their AI system or whether this performance is sufficiently high. Full Fact also did not share any numbers about uncertainty and variability in their results. There is a "confidence score" that the model produces, but it is not shown to the users of Full Fact AI, because Full Fact does not think it is reliable enough.
Another aspect that they consider is how often fact checks are published based on the software they provide. However, this is done not in a systematic way, "*for the media monitoring part of it, showing claims we think are worth checking, we may show 200 claims per day. And one that leads to a story, that's a great success. In percentage terms, it's only like half or 1%*" *(David Corney, Senior Data Scientist, Full Fact)*.
Similarly, since we don't know the details of the internal datasets, it is not possible to say from the outside whether these are appropriate to the task.

Despite these limitations and gaps in documentation, from what we can infer (based on the considerations above) we overall consider Full Fact AI as an AI software to be necessary and proportionate in a *satisfactory* (4/5) way.

7. Accountability

Assessment					*Exemplary*

Assessment question
Are there clear roles of responsibilities to ensure accountability for the AI system, its impacts and liabilities?

Full Fact named clear roles and responsibilities within the organisation for the decision-making process for the AI system, for liability cases, as well as for ethical and social matters of the AI project. Regarding cases of liability a clear structure of responsibility was stated:

> *"[...] [W]e're ultimately beholden to our Board of Trustees. So we have a cross-party, from across the political spectrum, board of trustees who have formal legal responsibility for the charity. So if we were to hold somebody liable in a fact check in the way that UK sort of charity law works, our trustees are ultimately the people who are responsible for that." (Andrew Dudfield, Head of AI, Full Fact)*

Full Fact has not reported any previous liability claims brought against them. In its contracts with its users of Full Fact AI, Full Fact excludes all liability for errors in the information provided by the AI software. According to the interview, Full Fact's contracts do not contain any non-disclosure agreements.
In regards to their internal governance, Full Fact describes an internal "*formal due diligence process*" *(Andrew Dudfield, Head of AI, Full Fact)* in regards to deciding about new collaborations as new partners, suppliers or funders related to any risks these collaborations might bring. Full Fact refers here in particular to reputational, technical, and strategic risks.

Full Fact has also undergone official accreditation processes to be recognized as fact checkers under European and international standards and is part of the European Fact-Checking Standards Network (EFCSN) and the International Fact-Checking Network (IFCN). This included transparent reviews of their work as fact checkers, including an independent assessment of their work, political neutrality, quality of correcting processes and many other criteria.
According to the interview, Full Fact only provides its tool to fact checking organizations which are also part of the EFCSN or IFCN.
For all of these explicit reasons, we rate this criterion *exemplary* (5/5).

8. Decision processes and design

Assessment				*Satisfactory*	

Assessment question
Is the design and decision-making process for the AI system reflecting principles of public interest and sustainability?

Overall we gained the impression that, within their workflow of decision-making for the AI system, processes are constructed with caution and include many loops to several people to ensure the highest journalistic standards. Positive examples here are a very problem-oriented and focussed AI usage, participative design-process including the needs and feedback of fact checkers (the core user group of the system), and the high awareness for potential harm within the organization. However, certain risks of harm remain, which warrant continuous caution, especially for the use of the Full Fact AI in other organizations and regional contexts. From a sustainability perspective, it can be criticised that Full Fact AI is based on LLMs (in this case Gemini) which are problematic due to their resource intensiveness as well as their developments by data workers in precarious working conditions. Full Fact AI communicated awareness of and a high degree of moral reflection on both of these realities. While, in this particular case we see legitimate arguments for the decision to use this technology, as no feasible better alternatives exist at the moment, the environmental and human rights issues which come with the use of proprietary generative models such as Gemini cannot be aligned with public interest and sustainability.
Another relevant design decision for Full Fact AI is the languages in which the interface of Full Fact AI is currently available. While the AI software itself works in any language which is supported by Gemini, the interface can be used in three languages: English, French and Arabic. This design choice was made according to new partnerships and the priority of their language needs. In general, there is the possibility to adapt the interface to also be available in other languages. This, however, would require work effort and would therefore cause additional costs.
Overall, we found that the design and decision-making process for the AI system reflect principles of public interest and sustainability as inclusive, deliberative and participatory processes. However, generally, the use of proprietary LLMs cannot be aligned with public interest and sustainability principles.
We therefore assess this criterion with *satisfactory* (4/5).

9. Inclusion, participation and co-design

Assessment					*Exemplary*

Assessment question
When designing, deploying and using AI systems, does the project take an inclusive, interdisciplinary and participatory approach?

From our interviews, we gained the insight that an inclusive and participatory approach was adopted during the development and deployment of Full Fact AI.

Full Fact clearly defines their target user group, is itself an active fact checking organization and Full Fact's fact checkers are using their own tools. We see that this leads to them being attentive to the needs of fact checkers and tailoring the functionalities of the Full Fact AI to these.

> *"We developed this for fact checkers. And so they are the people that we've worked with most. We have worked with other user groups as well: [...] with regulators, with charities, other people who are interested in what bad information looks like at scale online. And that has been helpful and beneficial to us, but we're really focused on that small group of users because that is the people we're engaging with." (Andrew Dudfield, Head of AI, Full Fact)*

Additionally, Full Fact was in exchange with potential users of the system to adjust its design to their needs. A good example of an inclusive design approach is that Full Fact adapted its AI tools to enable the usage on mobile phone devices after learning from fact checkers in other working environments that this would be needed. Also, the checkworthiness component was retrained and adjusted after fact checkers reported to Full Fact that the component would classify anything related to health as checkworthy.
A responsive and collaborative approach was also confirmed by Africa Check who described that special requests by them were taken into account by Full Fact and improvements were made. Overall, they expressed satisfaction about the support that they received from Full Fact and described the communication as "*very hands-on*" and "*proactive*" *(Fact checker Africa Check)*, especially in regards to information about new updates or functionalities.
What we are unsure of is the regularity and depth of feedback from all 45 organizations using the tool. There is no formal structure in place through which external users are regularly asked for feedback on Full Fact AI. However, there is a person who is responsible for external liaison with the users. Users forward their feedback to this person by email or raise it during regular monthly meetings that this person holds with users. Overall, Full Fact stated that it does not proactively ask for suggestions for improving the AI tools, "*we don't specifically say, 'how do we change the model to make it easier for you to use?' [...] So it's a rather other question of saying, [...] 'are there any problems with the system?'" (David Corney, Senior Data Scientist, Full Fact)*. This is justified by constraints in resources and time required to improve the system on the basis of

this feedback.
During the development of Full Fact AI, Full Fact also experimented with participatory approaches, including the collaborative annotation of data. However, Full Fact explained that this approach did not prove to be helpful, despite motivated participants.

> *"We have previously done work with volunteers to help on our technology work. And we found that that has not been so successful for us. We did a lot of work with annotations. And whilst it got a lot of people enthusiastic and we got quite a lot of data, we found that the difference between a fact checker and an enthusiastic volunteer was so stark that it wasn't actually that helpful to work with large amounts of volunteers. [...] We just didn't find the tasks that could benefit from the desire for people to give their time and energy to us." (Andrew Dudfield, Head of AI, Full Fact)*

We consider such experiments exemplary, as it shows the openness of Full Fact AI to include interested citizens in the process of training the model. This can be an educational experience and to a certain extent, a meaningful form of digital civic engagement. However, if there is a large drift in quality, it is understandable to choose other strategies and be honest about the downsides of participatory processes.

Full Fact has a dedicated contact address on their website and according to their information receives several inputs from citizens a day. Also they set up a dedicated LinkedIn account for Full Fact AI to answer questions and be open to discussing technical issues. In addition, Full Fact has a process for feedback to emails they receive, stating a very positive attitude towards interaction with citizens.

Finally, Full Fact AI displayed a high degree of organizational awareness for the importance of diverse perspectives within the team as a factor that influences inclusive decision making also regarding Full Fact AI

> *"[...] Within the team, we're a UK-based team working with fact checkers around the world, and so we are very mindful of that [...]. The person who's working with the fact checkers on a day-to-day basis is based in South Africa. And part of that is, again, to make sure that we're bringing in a slightly different perspective on things to make sure there's somebody who was actually originally a fact checker, but also a fact checker not based in the UK. It's been important for us to create a bit more diversity of the different ways of viewing information, different ways of working and also to ensure that it's not a small group of people in a small geographic area who are making choices. And then, we have to check our own privilege and choices as we're going through this. [...] There's no magic way through this, but considering the diversity and the team in a number of different ways about people's social background, their ethnicity, their religious beliefs, their neurodiversity. Like, we consider this stuff all of the time in thinking about: How do we ensure that we're not defaulting to any kind of worldview? We're challenging ourselves on these things". (Andrew Dudfield, Head of AI, Full Fact)*

Despite the fact that improvements for inclusive and participatory design are always possible but require resources and feasibility, we find sufficient arguments in the multitude of described efforts to give an *exemplary* (5/5) rating for this criterion, as in these different aspects we see important steps other organizations could learn from.

3.5 Cluster 3: Social sustainability

10. IT security

Assessment				*Satisfactory*	

Assessment question

Are there organizational responsibilities and measures to identify, address and mitigate IT security risks throughout the AI system lifecycle to prevent where possible, and/or limit any potential or actual harm?

Full Fact shared some thoughts on with whom they collaborate and why they did not share their trained models and internal datasets with us. We discuss them below ("25. Model and dataset sharing). Users of the software need a password (and username) to access it. However, from the user's perspective, two-factor authentication is missing, which would increase and improve security.

Apart from these aspects, Full Fact shared in the interview that security checks are periodically performed (e.g. pen testing was done), involving third parties and consultants. According to their statements, no "*major security issues*" *(David Corney, Senior Data Scientist, Full Fact)* were found during these tests. Additionally, for several years Full Fact got certified with Cyber Essentials[11]. However, the decision was taken not to renew the certification since participating in the process was not worth the effort and the overall process was described as a "*piece of bureaucracy that didn't really reflect the security of our system*" *(David Corney, Senior Data Scientist, Full Fact)*. Still, despite deciding not to do the Cyber Essentials process anymore, Full Fact reported to have internal documentation covering assessments also in regards to IT security, which was, however, not shared with us.

Other than that, they rely on the security systems of the software and platforms they use (e.g. google, github). This was justified with their small size as an organisation. Overall, we assess this criterion with *satisfactory* (4/5), as it shows to meet the necessary requirements to secure their system, but small improvements could be made.

[11] Cyber Essentials is a cyber security certification scheme which was launched by the UK government and developed by the National Cyber Security Centre. More information can be found on the Cyber Essentials website.

11. Sovereignty and power relations

Assessment				*Satisfactory*	

Assessment question
Does the project contribute to shifting power in a democratic way (e.g. to civil society or marginalized groups) and support inclusive use and equal distribution of benefits and harms of the AI system?

Generally, we see that Full Fact's mission is in support of a strong civil society. It is a non-profit organization with the goal to contribute to a better information environment, and is a watchdog for misinformation spread by politicians or others holding power. Full Fact describes themselves as:

> "[...] *[P]oliceman wandering around the neighbourhood. At times they're not there because there's a specific crime. They're just there so people will realise that there is somebody walking around. And so, we are that police person sometimes. We will fact check politicians to make sure they remember that we exist and so they know that we're there." (Andrew Dudfield, Head of AI, Full Fact)*

From an organizational side, Full Fact is committed to avoid dependency on a few funders. Their funding is diverse enough that the organization is not reliant on one specific funding party: *"[W]e're always trying to ensure that we diversify our income. Our goal is to have never more than 15% of our funding coming from any one source" (Andrew Dudfield, Head of AI, Full Fact)*. In addition to that, just recently Full Fact has changed its funding policy and now accepts also public funding.

A weakness in regards to digital sovereignty is the current use of Gemini for some functions of Full Fact AI. This dependency could potentially impact the functionality of Full Fact AI[12] and also indirectly support the powerful position of Big Tech Companies (in this case Google). While they reflect this dependency, at the moment it remains. However, switching from the Google platform to other open source models was described as "*work in progress*" *(David Corney, Senior Data Scientist, Full Fact)*.

> *"And so the idea of US sovereignty winning over everybody else is a real consideration for us from a legal point of view, philosophically, morally, about how we want to approach these things and to what extent do we want to use proprietary models versus open models?" (Andrew Dudfield, Head of AI, Full Fact)*

[12] This is relevant as proprietary systems may be taken offline, or users may, over time, be required to agree to more extensive use of a provider's additional features in order to maintain the operation of the systems.

Full Fact is committed to empower users of their AI software. As Full Fact sets out, its AI tools are there to support fact checkers, to "*scale [...] [the fact checkers'] own thinking*" *(Andrew Dudfield, Head of AI, Full Fact)* and to enhance the "*ability to listen*" *(Andrew Dudfield, Head of AI, Full Fact)*. The aim is not to replace fact checkers, but to expand their scope of action. In addition to that, users are given responsibility and autonomy over the tools and the way fact checkers work with them, as they can configure and customise the Full Fact AI to suit their needs:

> *"[...] [I]t creates the onus on them to curate the content. And we found that that is the most helpful way of empowering people. So, they understand that it's their choices that have been reflected, rather than some distant algorithm telling them what they should fact check." (Andrew Dudfield, Head of AI, Full Fact)*

Of course, this also might warrant some problems, as the selection of sources to the AI pipeline might create biases. Under the condition that potential biases and harms are well-communicated and sufficient training was offered to these external fact checkers (which Full Fact AI describes), we agree to their assumption that this ability to customize and adapt the tool to local contexts is important.

Further, the AI software is provided only to registered fact checkers or fact checking organizations which is supposed to prevent misuse but also to empower fact checkers who are committed to neutrality and democratic principles.

> *"[...] We only make it [the AI software] available to registered fact checkers. [...] You can imagine [...] some authoritarian government using it. If they define misleading as meaning 'attacking the government', then they [can] use that to find things journalists are saying against the government or people on social media are saying. So it can then be used as a tool to enforce censorship, and the software basically wouldn't know which way it's being used of course. So we limit it by saying it's only available to people whose ethics we basically agree with, which basically means not any government body." (David Corney, Senior Data Scientist, Full Fact)*

Full Fact AI is currently available in English, French and Arabic. When looking at where the fact checking organisations using Full Fact AI are located (as highlighted in this map), these languages represent the official languages in most of the countries. However, even if it is the official language it may not be the one most widely spoken by the population (e.g. Guinea, Namibia, Senegal). Also, Full Fact AI is used in some countries which no longer have one of these three as an official language (Burkina Faso and Israel), which is spoken, if at all, by a minority only.
From an empowerment perspective, this raises the question about who can make use of the AI software and which media can be monitored.
For the use of Full Fact AI, external organizations need to pay fees. We interpret the amount as relatively fair and not as a barrier to the inclusive use of Full Fact AI:

> *"The license fee ranges, it depends on use basically. So a tiny organization, not looking at much, [for them] it's very small amounts. But if DPA [German Press Agency] turned up, that would be probably a few thousand dollars a month to run at the scale that they would need to. So it is not a one size fits all thing. It's dependent on the literal AI throughput effectively." (Andrew Dudfield, Head of AI, Full Fact)*

While the use of Full Fact AI creates a dependency of fact checking organizations on the AI system, this seems reasonable as many especially small organizations will not have the capacities to build and maintain a similar tool themselves in equal quality.
The goal of Full Fact reflects the idea of empowering the public by access to high-quality information and through this to hold people in power accountable. We acknowledge the efforts of Full Fact to stay independent as an organisation, to enable an inclusive use of their AI tool and a commitment to empowerment. However, due to their dependency on Big Tech companies, we give an overall rating of *satisfactory* (4/5).

12. Accessibility

Assessment				*Satisfactory*	

Assessment question
Does the project convincingly describe efforts or create measures to create accessibility?

Accessibility refers to people with special user needs as well as in precarious economic situations or with limited access to the internet.

Full Fact describes efforts with the intention to increase the accessibility of their tools focusing on their specific user target group.
Firstly, since Full Fact realizes that many users of their AI tools are in a situation of poor internet connectivity, they worked on making the tools accessible via mobiles phones as well.

> *"We're often working with people who are working in difficult information environments, which can often mean very poor internet connectivity. So you spend a lot of time thinking about how to optimize for that. To start with, we didn't optimize at all for mobile usage because [...] in our worldview, a fact checker was somebody sitting at a desk in an office. And that proved to be very wrong. Sometimes, a fact checker is a person with a mobile phone. And so we wanted to make sure that we could offer that experience that was useful to people working in environments that look quite different to what we did." (Andrew Dudfield, Head of AI, Full Fact)*

Secondly, Full Fact describes the effort to design Full Fact AI in a way that it can be used by people with different input and access needs. The AI tools are "*input device agnostic*" *(Andrew Dudfield, Head of AI, Full Fact)* and can be used with various input devices. Also, customization of

color, contrast and size is possible. Despite their efforts, Full Fact themselves acknowledge that there is room for improvement.
We see the efforts to increase accessibility as appropriate since the main target group of users of the AI software is not the general public but fact checkers. In our view the prioritization of their accessibility, such as adapting it for smart phone use and low internet connectivity, is adequate. Nevertheless, the accessibility of Full Fact AI for users with special needs could be even more increased if the software would adhere to any accessibility standard or guidelines.
Overall, we rate this criterion with *satisfactory* (4/5).

13. Non-discrimination

Assessment			*Ambivalent*		

Assessment question
Does the project take appropriate measures to prevent bias, discrimination and stigmatization of any kind, in alignment with fundamental rights?

There are different ways in which Full Fact AI can be biased. There can be **differences in performance between the languages** it is used for, with a bias towards better text processing in English. Also, quality can depend on the regional dialect of a specific language (e.g. French spoken in West Africa versus France). In the worst case, the software may not be able to process a specific language (e.g. the language Wolof in the transcription of the audio).
Also, there is a **bias with respect to input medium**. Transcribing an audio to a text and processing this text is harder than directly processing a written article, because audio (speech) tends to be more informal and fragmented. Such a bias that depends on the input medium may be mixed with a language bias, because the article structure or way of speaking can be also linked to the language.
Potentially, the tools may also be subject to representation bias, if datasets for classification using BERT models are not well balanced. This can lead to a classifier that misclassifies certain sentences more often than others.
Similarly, biases that may occur as part of the "checkworthiness" component of Full Fact AI are possible, but were not raised by Full Fact specifically. Based on information gathered through the interviews, there have been no evaluations of the models with respect to performance factors (e.g. language) or evaluations of the trained models using fairness metrics. Also, while datasets do not contain sensitive features, we do not have any information about evaluation of the datasets with regards to possible bias (representation bias) used for training the BERT models. From the interviews, we understood that no mitigation measures to counteract any type of bias were conducted.
There is some limited information about model performance, but any inference about bias from that data is not possible. However, some information about the possible types of bias (language, input medium) was shared with us. Regarding bias in the model with regards to language, it was mentioned that Full Fact is aware of potential differences in performance. However, they do not

have the resources for retraining models. In addition, Full Fact assumes that errors based on languages are random and not systematic.

In summary, only limited information was provided to us about possible bias in the Full Fact AI system and no evaluation nor mitigation measures of biases exist.
Overall, we assign an *ambivalent* (3/5) rating. We thereby take two aspects into account: first, the Full Fact AI system is embedded in editorial decision processes overseen by humans and no final decisions are delegated to the system which reduces the severity of possible bias. Second, significant resource requirements are needed for counteracting language bias through retraining. Overall, this rating is given with lower confidence due to the limited information provided.

14. Data protection

Assessment				*Satisfactory*

Assessment question
Are the rights of data subjects respected, protected and promoted throughout the lifecycle of the AI system according to applicable laws?

From the interview, we did not see evidence that Full Fact technically evaluates the privacy of their ML models or the data that they collect and use for training. In general, the datasets that Full Fact creates for training their own model consists of already published media content. This published data may already contain personal details (such as names of individuals, details about their health or other sensitive information).
However, Full Fact made us aware that they refrain from publicly sharing their datasets for other reasons: sometimes, internally, Full Fact asks their fact checkers to make annotations to it, e.g. about its checkworthiness. This information, potentially together with the names of the fact checkers, can therefore be an additional part of the datasets and can be critical to share.
According to the interview, when using Full Fact AI no additional personal data is collected. The only data of users which is stored when using Full Fact AI is the email address. These email addresses are usually tied to the organization the people work for and are not private ones, and can often be found on the internet or social media platforms. It was reported to us that there is existing internal documentation on data protection processes and proper management of this data according to existing requirements.
Overall, we rate this criterion as *satisfactory* (4/5) at this moment but recommend Full Fact to exchange with a data protection expert, especially in regard to potentially providing Full Fact AI in the EU. In case Full Fact AI would be applied in the EU, the performance and documentation of a Data Protection Impact Assessment (DPIA) may be required.

15. Human oversight

Assessment					*Exemplary*

Assessment question
Does the design of the system ensure human oversight and human capability to oversee the overall activity of the AI system?

Full Fact made it clear that humans are essential for the use of the Full Fact AI software, as human labour is required from setting up the software and deciding about input media to checking and validating the outputs of the AI software. Full Fact highlighted the importance of the "*human in the loop*" *(Andrew Dudfield, Head of AI, Full Fact)*. However, we consider the role of humans as more essential than simply "being in the loop'. In this regard, Full Fact highlights that the fact check itself is not automatized. Fact checkers do not simply oversee AI decisions but remain in charge of the full fact checking processes. We see it as a positive that Full Fact showed a high sense of responsibility and awareness of the role and limitations of their AI tool. This is reflected in their awareness of the limitations on what the AI tool can provide and the scope of its capabilities.
The fact checkers play an important role in various stages of using Full Fact AI, since the use of the AI tool itself bears the risk of overreliance or of introducing bias. Firstly, there is a danger that fact checkers "*could potentially outsource responsibility*" *(Andrew Dudfield, Head of AI, Full Fact)* to the AI tool. This could lead to fact checkers relying on the AI software for media monitoring and missing other important potential misinformation because it is not flagged by Full Fact AI. Secondly, Full Fact expressed high awareness about the*"need to be mindful of what we are suggesting this person has done" (Andrew Dudfield, Head of AI, Full Fact)*. There is an awareness of the potential harm if people blindly rely on the output of the AI tool without checking it, potentially leading to wrongful accusations of misinformation.

> *"If you use our tool, all it's doing is showing you, 'here is some stuff to look at'. And it's not saying 'it's right' or 'it's wrong'. [...] And we try to bake that in into the way that we describe it, in the way that we train people to ensure that they're not just gonna say, 'okay, well it came top when I logged on, so therefore it's the most important thing, therefore it must be wrong, therefore I'm gonna say it wrong.' That is not what it's there to do." (Andrew Dudfield, Head of AI, Full Fact)*

According to the interview, this understanding of the need to verify the outcomes of the Full Fact AI is a prerequisite for all users in fact checking organizations and embedded in the communication for its usage.

Secondly, human oversight also plays an important role when setting up the tool. When it arrives, it is empty, meaning that users decide which media should be monitored. The responsibility for and prevention of bias in regards to monitored media input lies with the fact

checking organization since the decision about the media which is monitored by the AI software is an editorial one.

In our view Full Fact displays a high degree of reflection on the capabilities and limitations of Full Fact AI. Moreover it places an importance on the "human in the loop" and an even higher one on human oversight which leads us to give an *exemplary* (5/5) rating for this criterion.

16. Self-determined use

Assessment					*Exemplary*

Assessment question
Do users make a self-determined choice when and how to use the system?

This criterion assesses whether the users of the AI system are able to decide on whether or also not to use the AI system or which degree of optionality users have to use the system.

Currently, at Full Fact it is not mandatory to use the Full Fact AI in acknowledgement of the different working styles of individual fact checkers. In our view this lack of mandate is positive, as it acknowledges experts' workflows and preferences and mitigates the risk of overreliance on such an assistive tool. Full Fact describes that it is open to fact checkers using whichever functionalities they please, and to whatever degree. They highlight that Full Fact AI should be seen as a tool to support and by no means to replace: *"[...] It is not mandatory in our organization and we don't suggest it becomes mandatory for everybody" (Andrew Dudfield, Head of AI, Full Fact).* Even if Full Fact itself does not advise to mandate the use of the software, we must note that it is possible that other organizations using Full Fact AI might handle this differently. A mandate to use Full Fact AI could lead to a greater dependency and overreliance on it. However, Africa Check confirmed that it is not mandatory to use the software in their organization either.

Overall, from the information on this aspect we could gain, we rate this criterion with *exemplary* (5/5).

17. AI literacy

Assessment				*Satisfactory*	

Assessment question
Does the project describe efforts or concrete measures to improve AI literacy of people directly or indirectly impacted by the AI system?

For the assessment of this criterion we want to understand if and how AI literacy has increased

alongside the use of the tool. In general, Full Fact described that users do not need any advanced skills to use the AI tools. We see it as positive that external fact checkers are provided with an onboarding process that is adapted to the needs of the users and seems sufficient to allow for independent use.

> *"The process we use is different depending on the users, but it's normally a structured onboarding. So we'll run training sessions for people to make sure that they're comfortable. There is documentation provided, too. We allow people to understand the tools and troubleshoot themselves. And then we'll probably have drop-in sessions, so with some of our users, we run a once a month thing, where people just come along and ask us questions." (Andrew Dudfield, Head of AI, Full Fact)*

In addition to that, Full Fact offers a training program on "*our own AI tools as well to highlight what is possible or how we as fact checkers use this kind of thing" (Andrew Dudfield, Head of AI, Full Fact)*. This training program is an introduction to the AI tools and their use in practice.
The website of Full Fact AI contains some more details on how certain AI features work. For instance, one blog article describes how a claim matching tool works. Whilst this can be interpreted as contributing to AI literacy, it is not a systematic or structured contribution.

In addition, Full Fact supports media literacy more generally and conducts training on various topics for different target groups in the UK. Examples are workshops, an e-learning course on fact checking as well as workshops on misinformation or critical thinking.[13]

Besides all these important efforts to increase the user literacy and media literacy in general, there is no dedicated program for AI literacy which explains how their underlying models work and where the risks of the system lie. It could be helpful however, to also offer a more general AI basics workshop to potential external users (or at least check for the level of pre-existing knowledge) to make sure they have a sufficient understanding of the basic technology behind LLMs and AI in general. Therefore, we opt for a *satisfactory* (4/5) rating for this criterion.

18. Open for validation and contestation

Assessment					*Exemplary*

Assessment question
Does the project provide sufficient transparency and openness for evaluation?

We see the agreement of Full Fact to join our research project as an effort towards transparency and an indication of Full Fact's openness to be validated, contested and evaluated. Prior to our assessment, Full Fact did not undergo any other voluntary assessment or audit. While we did not get access to the AI model and dataset (this is elaborated under 25.

[13] More information on what Full Fact offers can be found on their website.

Model and dataset sharing), Full Fact was very open and cooperative. They shared requested documents and documentation and provided access to a demo version of the Full Fact AI software to us. Overall, Full Fact was very open and cooperative in facilitating interviews with us and providing the necessary time this process required. Also, during the interviews they answered questions extensively and were available to answer follow-up questions subsequent to our interview. They showed openness to discuss the different potential harm scenarios stemming from the FRIA-triage with us and a general interest in the feedback and recommendations provided by us after the assessment.
Therefore, we assess this criterion with an *exemplary* rating (5/5).

19. Technical transparency

Assessment			*Ambivalent*		

Assessment question
Is technical explainability ensured, meaning that outputs by an AI system can be understood and traced by human beings?

Overall, technical transparency of Full Fact AI is limited. There may be good reasons not to disclose model, data, and documentation publicly. However, internally, there could be a higher degree of transparency.
Full Fact AI is using BERT models as well as Google Gemini, which are not interpretable. Furthermore, no method to increase the transparency of such models (XAI methods) is currently used. However, as stated in the interview, future updates of Full Fact AI may allow for increased transparency. So far, however, we are not aware of any such tool.

While audit trails are kept for internal documentation, Full Fact AI does not have publicly available technical documentation. This is also connected to the fact that they do not share their model and data. However, some broader articles about how Full Fact AI works are available on their webpage. Internally, the code is documented inside the github repository that they use for development. From the interview, it was not clear whether there is comprehensive written documentation of the system. Also, for our assessment, no code or data (excluding the data that was already public) was available to us.
Overall, we rate this criterion with *ambivalent* (3/5) as we believe a more comprehensive documentation would be possible and helpful.

20. User transparency

Assessment				*Satisfactory*	

Assessment questions
Is there public information about the AI system that is understandable to the user of this system? Are individuals meaningfully informed about the AI system that they use, when a decision is informed by or made based on AI algorithms?

We positively rate that Full Fact decided to include 'AI' in the name of their AI software, so it is clear for users that there is an AI system behind the software. Further, Full Fact created, next to the website of their organisation (Full Fact), a separate webspace for Full Fact AI (Full Fact AI). This website provides information on the mission and goal of Full Fact AI and describes the main functionalities of the software. While it is clear for what goal NLP is used, details on the AI system behind the software are only provided in some cases. For example, they mention that the claim-type classifier is based on their own fine-tuned version of BERT. Such information is given in their general description or through different blog articles.
We do believe it is relevant to communicate to lay users which AI models and data are in use, especially when generative models are in use, so they are aware which infrastructures are utilized. Although this is a dynamic process that is subject to regular changes, a structured overview about which models and data are in use would be helpful.

Also, on the website of the Full Fact organization, information about Full Fact AI can be found. Similar to the Full Fact AI website, the section on Full Fact AI contains goals, impacts and the basic functionalities of the AI software. Also here, no detailed information about the AI model and dataset can be found. However, a small paragraph on the limitations of AI usage and the importance of the role of fact checkers is included:

> *"There are a lot of people who say that artificial intelligence is a panacea but we have been on the front line of fact checking since 2010 and we know first hand how difficult it is. Human experts aren't going anywhere anytime soon — and nor would we want them to be."* (Website Full Fact)

We regard the disclosure of this limitation as an important aspect of transparency towards users. However, the transparency for users and the general public would be increased if it were clear which models and data the AI software is based on. This led us to give a *satisfactory* (4/5) rating.

3.6 Cluster 4: Environmental sustainability

Hardware and emissions of Full Fact AI

The following information about the hardware of Full Fact AI was reported to us: Full Fact uses Apple MacBooks in combination with the Google cloud platform. Their primary compute engine is located in the Google region "europe-west2-a", which is in London, UK (see google cloud website). It amounts to 112 virtual CPUs and 448GB of total memory, as well as an associated cloud storage in the order of 20TB. Overall, Full Fact pays £4000-£5000 per month for the Google services they use.
According to Google, the "europe-west2-a" region is considered "low carbon". This means that the Google CFE% (percentage of carbon free energy, which is based on the average percentage of CFE consumed on an hourly basis, taking into account the investments of Google into CFE at that location) is at least 75% or if not available, the grid carbon intensity is of maximum 200 gCE2eq/kWh. For "europe-west2", these values are CFE% = 79% and 106 gCE2eq/kWh.
Based on the internal dashboard of Google, emissions of Full Fact AI equate to very approx. 0.1-0.2 tCO2e per month.

21. Complexity, energy and emissions

Assessment				*Satisfactory*	

Assessment question
Does the project sufficiently consider the complexity, energy and emissions of the AI system, including, but not limited to, measures to reduce these?

We have limited insight about the direct environmental impact of Full Fact AI on energy and emissions. Apart from the general numbers provided in the Google dashboard (see general information above, 0.1-0.2 tCO2e per month), no other information about energy consumption or emissions of the Full Fact AI system are available or were provided upon request. Taking the upper limit of 0.2 tCO2e per month, this equals approximately 1000 km driven by a car (1000 km = 194,3 kg CO2e, using this tool). However, it is not transparent which emissions of the overall Full Fact AI system are part of the number reported to us. For example, does it cover only the use of Gemini[14] for transcription or also inference of the two classification models (claim type and topic), checkworthiness model and other technical steps in the Full Fact AI pipeline?[15] Lastly, there is also the open question about the emissions tied to model training, specifically regarding the models that Full Fact trained themselves, and whether these are also

[14] An overview of environmental impacts of large models, including Gemini, can be retrieved via this interactive dashboard.
[15] According to this article, it is possible to see emissions separated by project.

part of the reported number or not.
When questioning Full Fact about the impact of their models on the environment in different AI life-cycle stages, Full Fact estimated that their overall impact during deployment is higher than during training. Full Fact does not re-train the Gemini model that they use for transcribing the audio data, but the smaller BERT models (and the checkworthiness component). This is in contrast to the very frequent use of Full Fact AI. Specifically for Gemini, they stated that "*that's going to be over 400,000 sentences per day.*" *(David Corney, Senior Data Scientist, Full Fact)*.

Full Fact shared that their priority is usually to directly reduce their financial but not environmental costs. However, such costs are in their opinion often tied to environmental costs. For example, using only one location of a datacenter is cheaper compared to using several locations. Using only one location also implies that no data needs to be transferred which in turn reduces energy consumption.
However, the organization also reported measures taken to directly minimize their environmental impact. Firstly, incoming media for which monitoring has been requested by various organisations (users of the software) are analysed only once in total, but the results of the analysis are shown to all organisations with an interest in them. Also, they stated that they try to use smaller models if possible. Full Fact stated that they also try to reduce emissions by running training in times which are "*out of cycle [...] on downtime*" *(Andrew Dudfield, Head of AI, Full Fact)*.
Lastly, while it appears that the Google server is located in a "low carbon" region (see also information box at the top), according to the interview, this was not a deliberate choice, *"[...] we don't have a choice of which Google data centres to use" (David Corney, Senior Data Scientist, Full Fact)*.

When talking about the environmental impact of Full Fact AI, Full Fact describes a dilemma. They acknowledged the environmental costs of AI systems but also emphasized their usefulness: "*[...] we do use the models, even though we know they are environmentally damaging, because they're very, very useful. So it's a trade-off that we are conscious of*" *(David Corney, Senior Data Scientist, Full Fact)*. At the same time, they showed particular interest in the topic of environmental impact and ways to improve their actions in this regard.
Given these considerations, the direct or indirect measures to reduce their footprint, and the open questions about reported information discussed above, we rate this criterion with *satisfactory* (4/5).

22. Natural resources

Assessment			*Ambivalent*		

Assessment question
Does the project sufficiently consider the resource use of their AI system?

Using an AI system comes at the cost of natural resources that are used e.g. in the production of hardware. Here we assess the AI project's considerations about the resource use of their AI system, which includes decommission plans for the hardware and certification of hardware in use.

Similar to what is already described under "complexity, energy use and emissions", Full Fact stated in the interviews that their focus is primarily on financial costs but that they think these costs are tied to environmental costs, including natural resources. Here, an example is their "*very broad policy*" *(David Corney, Senior Data Scientist, Full Fact)* that once laptops are replaced, they could be used by other people or get recycled, but there are no clear guidelines. This "*very broad policy*" is also the only information that was provided to us about decommissioning plans. Additionally, except the "low emission" label that is given to the Google servers they use–which is not an official certificate–, no certification of their hardware was shared with us.
In addition to this, Full Fact described the trade-off between buying new hardware infrastructure that requires new material but can improve, for example, computational efficiency, versus keeping the same hardware and not profiting from such improvements.

Overall, considerations and actions about natural resources behind the Full Fact AI system can be improved. There is no concrete plan on what happens to their hardware once decommissioned. Also no information about certificates of their hardware or other considerations on how their hardware was selected was provided to us. Taking these aspects into account, but also considering that the organization is small and the number of hardware devices limited, we assess this criterion with *ambivalent* (3/5).

3.7 Cluster 5: Economic sustainability

23. Model and dataset sharing

Assessment				*Satisfactory*	

Assessment question
Are the model and the dataset publicly shared, or are there good reasons not to share them, and did the project take measures to make it easy to use them?

We see and appreciate Full Fact's high engagement in efforts to increase transparency in different domains (e.g. related to funding sources, processes in model design). However, Full Fact shares neither their trained models nor the datasets that they use for training of the AI tool. A fact that Full Fact does not hide, but rather gives reasons for. Only the dataset used for topic classification is available, but this dataset was not created by Full Fact. At the point of conducting the interviews, Full Fact mentioned that they are preparing a conference paper for which they will also publish parts of their internally created datasets.

Full Fact gives two reasons for not making the code for Full Fact AI publicly available. The first is related to the business model that they are trying to establish. Their business plan is based on selling licenses for the use of the Full Fact AI software. Currently, Full Fact is very dependent on external funding (see under 13). A goal for the future is to be more self-sustaining by generating revenue through selling commercial licenses for Full Fact AI. The second reason is risk related, as it concerns the possibility to find work-arounds if the code, or too much information about the system, is available. As a consequence, entities could work out how to merely appear as informative, despite actually being misinformative.

> *"We are working in a space that, if we tell people how we work out if something is potentially harmful, it's easier for people to then work out how to not seem harmful. And so there's a sort of a challenge for us to just make sure that we are transparent, but not so much that people can understand how to work around us. And that's an exercise that we have to work with." (Andrew Dudfield, Head of AI, Full Fact)*

Full Fact provides a number of justifications for not publishing the dataset. Firstly, Full Fact mentions the issue of copyright. The dataset consists of annotations of newspaper articles and other published text which would raise copyright questions if published. Secondly, the effort to prepare the data for publication (e.g. clean it, fully anonymize it) as well as to provide explanations (or clarifications on the data after publication), outweigh the benefits of doing so.

> *"[...] [W]e don't have millions of records. Having a few thousand records, a few thousand sentences, someone saying, 'yeah these two are the same thing' or 'this is a prediction', there's no massive value there and anybody can recreate the same data sets themselves with*

[...] a few hours of effort or a few thousand dollars." (David Corney, Senior Data Scientist, Full Fact)

As a third reason not to publish the data, Full Fact cites its commercial value.

> *"So, we use the BERT model, we use Gemini, and we are using annotations from our own fact checkers. And, whilst we are wanting to be transparent in the way that we work, those annotations have commercial value. And so, we are very sensitive to the fact that, whilst we're giving away our fact checks, we give away a lot of meta data: the actual annotation. So how a fact checker's brain works and the codifying of that is information that we don't necessarily want to openly publish because it can be commercially used. And we're not sitting here making any money from that ourselves, but also we're sensitive to the fact that it does have value to it." (Andrew Dudfield, Head of AI, Full Fact)*

We believe that open source and publicly available models and datasets alone are not the only viable approach and do not, by themselves, justify a positive rating in this category. For some use-cases it may be beneficial or needed to act open-source for others less. For the evaluation of this criteria, there is a deeper and more situated understanding needed for why a project decides not to make a model and dataset publicly available.
Overall, we find the reasons provided for choices made in regards to the models and data related to Full Fact AI understandable and consider them as proper justification.
Additionally, Full Fact is not opposed to open-source in principle. In other projects and collaborations (e.g. with Maldita), the code is open source.
However, we see it as a downside that Full Fact makes no exceptions in not sharing models and datasets. For instance, access could be granted to researchers for research purposes. This is why we opt for a *satisfactory* (4/5) rating of this criterion.

24. Future perspectives and collaboration

Assessment					*Exemplary*

Assessment question
Are there efforts to establish collaboration and to sustain the project and the fulfilment of its objectives in the future?

To guarantee the availability of the project and sustain it in the future, two aspects are of high importance: secure funding, and active collaboration with other organizations and initiatives. For both aspects, we could identify high efforts by Full Fact for sustaining Full Fact AI.
To secure funding for Full Fact, and in particular Full Fact AI, is described as a constant challenge by Full Fact. For instance, in its initial phase, for the development of Full Fact AI, Full Fact received money from the philanthropic arm of Google, called google.org. After six years, this funding ended. The software has been developed with funding from multiple projects and

sources over the last eight years. Details of Full Fact's funding are published on their website. At the same time, Full Fact prioritizes their independence, including with regard to their funding:

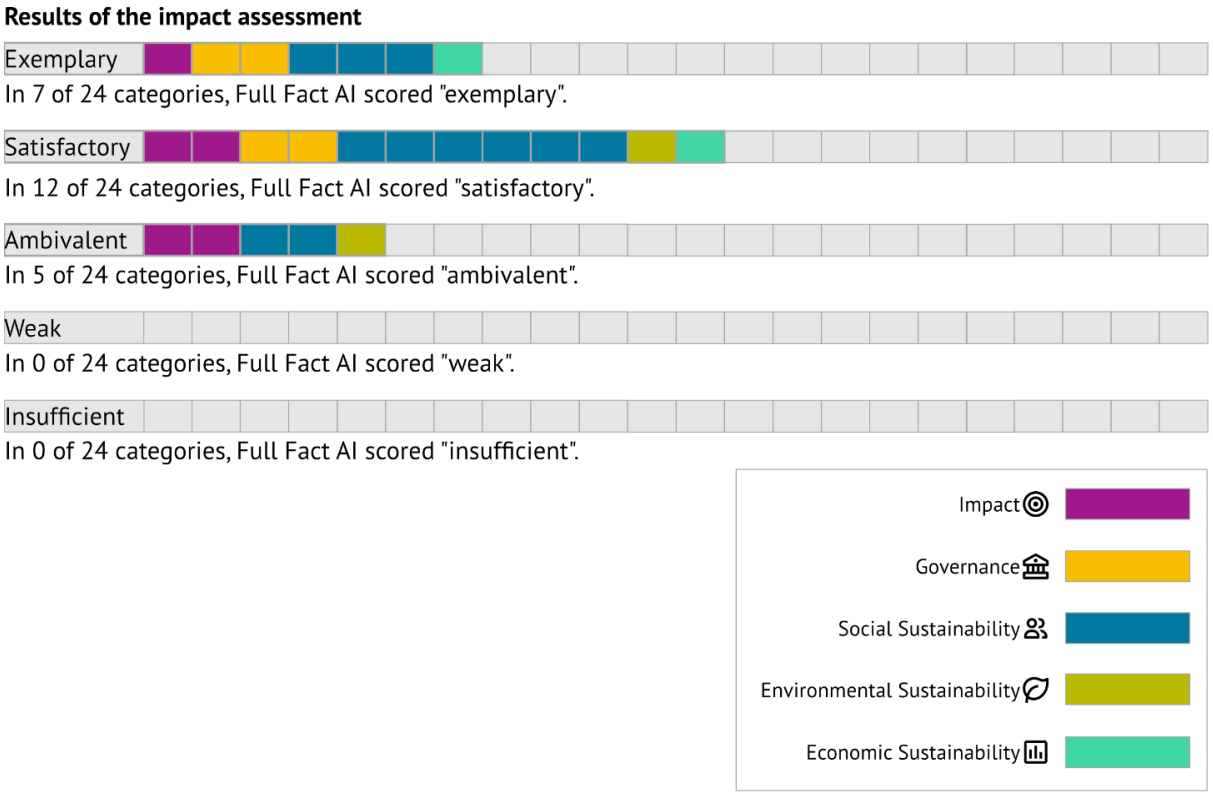
> *"Our goal is to never have more than 15% of our funding coming from any one source, and that means we constantly need to think about how to make sure that we're not reliant on anybody […]." (Andrew Dudfield, Head of AI, Full Fact)*

As funding is a constant challenge, Full Fact works on the diversification of their income sources and implements a business model to generate revenue through Full Fact AI. This is based on paid licenses, which organizations wanting to use Full Fact AI must obtain. Full Fact works on increasing revenues generated through these means, aiming for 10-15% of their organizational revenue to be based on providing their AI software to other organizations.
In regards to collaboration, Full Fact shows high efforts. The Spanish fact checking organisation Maldita is a long-standing partner of Full Fact, leading to collaboration in different projects. Full Fact also actively collaborated with two fact checking organisations, Chequeado and Africa Check, within a joint research programme (Full Fact's joint research programme with Africa Check and Chequeado). Further, the participation in the European Fact-Checking Standards Network (EFCSN) and the International Fact-Checking Network (IFCN) enables collaboration with various organisations. Full Fact sees collaboration and widespread access as more important than a business advantage, since it prevents developers from creating fact checking products which compete with each other. It is more important to create one tool to ensure "*that fact checkers have access to the best technology*" *(Andrew Dudfield, Head of AI, Full Fact)*.

As Full Fact demonstrated strong commitment to sustain Full Fact AI in the future by securing funding, establishing a business model while at the same time balancing this revenue model with providing wide spread access and building active collaborations with others, we give an *exemplary* (5/5) rating.

4 Conclusion and recommendations

The following visualization shows an overview of the results of our impact assessment and indicates which and how many categories achieved a specific rating.



We have passed on the following concrete recommendations to Full Fact during a feedback conversation, which was received in a constructive manner:

"*It's been very useful. [...] So, it's much easier for me to go to my new line manager and say, 'Hey, here are things to think about when we buy new laptops or when we look at non-Google platforms'. [...] So, I think, providing that sort of clarity, of ideas, is useful already.*" *(David Corney, Senior Data Scientist, Full Fact)*

Following the impact assessment, our feedback and ideas for improvement are as follows, organised according to our evaluation criteria:

Objectives and Realisation

- (Nice to have): Think about formalizing the impact logic between your envisioned goal, the outcomes and output, e.g. by formulating a theory of change

Impact achievements

- Establish (systematic) monitoring and reporting of direct impacts on fact checkers to

generate clear evidence and numbers on impact for the AI system specifically. This could include questions like:

- How many of the suggested claims were checked?
- Accuracy of alerts?
- Reduction in workload (work hours saved)?
- Do fact checkers see a change in quality of their work?
- How do work processes change?
- Which other impacts on their work do fact checkers notice?

Potential harm and mitigation

- Request risk assessment findings from third party provider (Google in case of using Gemini)
- Full FRIA assessment in collaboration with European Center for Not-for-Profit Law
- Awareness and training around FRIA and potential harms of Full Fact AI, in particular for external users of Full Fact AI (e.g. in form of a training or elaborated onboarding)
- Conduct a yearly survey on any biases or other FRIA related harms that external users might have noticed in their contexts

Necessity and proportionality

- Systematically assess data on model and data appropriateness (e.g. model performance, uncertainty and variability)

Sovereignty and power relations

- Move towards non-proprietary LLMs (connected to big tech), explore (open) more sustainable options if possible

Accessibility

- Adherence of Full Fact AI to any accessibility standard or guideline

Data protection

- Exchange with a data protection expert about the possibility to be required to do a Data Protection Impact Assessment (DPIA), in particular if Full Fact AI will be applied from users in the EU

Self-determined use

- When introducing users to Full Fact AI, be clear about a non-mandatory use of Full Fact AI

AI Literacy

- Offer a general AI basics workshop to potential external users to make sure they have a

sufficient understanding of the basic technology behind LLMs and AI in general

User transparency

- On your website, include a clearer communication for lay users which AI models and data are in use. E.g. a structured overview about model and data at one place (not fragmented in different blog articles)

Non-discrimination

- Evaluate models with fairness metrics / compute performance measures of model disaggregated by factors (classification models)
- Evaluate training datasets with regards to their representation
- Use of bias mitigation strategies if results of evaluation require it

Technical explainability

- Use explanation methods to increase transparency of models
- Create an internal technical documentation as a written document

Complexity, energy and emissions

- Have a look at what is behind the emission number from the google dashboard that you reported to us. Are there other possible emissions (larger in scale)?
- Track how often the different functionalities of Full Fact AI are used (deployment): this is a scaling factor for emissions and will help you to estimate your environmental impact in more detail

Natural resources

- Create a decommissioning plan for hardware
- Think about the hardware you use: are there environmentally friendlier alternatives (e.g. other laptop brands that are more environmentally friendly, use of refurbished laptops)?

Model and dataset sharing

- Share models and datasets in exceptional cases, e.g. for research purposes without full publication

The assessment overview we provide is designed to serve several purposes: it is a project development asset, as it gives a comprehensive analysis and feedback for many layers and aspects of a specific AI use case. We are confident that it serves as a meaningful and actionable outcome to improve concrete projects in their impact for public interests and sustainability. Secondly, it can provide guidance for other AI use cases as they might help to reflect their own

strengths, but also options for improvements or even potential harms they have not considered. Finally, every case delivers insights for further research and contributes to a more nuanced understanding of AI technology use in practice. We thank our interview partners and Full Fact for their openness to contribute to all of these purposes.

References

Aapti Institute. 2026. *Invisible Workers, Visible Harms: Perils and Precarities of AI Labour*. On behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ). https://www.bmz-digital.global/wp-content/uploads/2026/02/GIZ_2026_InvisibleWorkersVisibleHarms-1.pdf.

Amazeen, M. A. 2017. Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, *21*(1): 95–111. https://doi.org/10.1177/1464884917730217.

Cobb, M. D., Nyhan, B., & Reifler, J. 2013. Beliefs Don't Always Persevere: How Political Figures Are Punished When Positive Information About Them Is Discredited. *Political Psychology*, 34(3): 307–26. https://doi.org/10.1111/j.1467-9221.2012.00935.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805.

Dudfield, A. 2021. How AI helps us detect 100,000 potential claims a day. Full Fact. https://fullfact.ai/blog/ai-google-100000-claims-day/.

Flynn, D.J., Nyhan, B., & Reifler, J. 2017. The Nature and Origins of Misperceptions:
Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, 38(2): 127–50. https://doi.org/10.1111/pops.12394.

Full Fact. 2025. Full Fact Report 2025. https://fullfact.org/media/uploads/full-fact-report-2025.pdf.

Goyanes, M., Lee, S., Salgado, S., & Gil de Zúñiga, H. 2025. Enhancing Self-Perceived Disinformation Identification in Democracy: The Impact of Fact-Checking Integration into Daily News Consumption Practices. *Journal of Broadcasting & Electronic Media*, 69(3): 331–352. https://doi.org/10.1080/08838151.2025.2508767.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. 2020. A digital media literacy intervention increases discernment between mainstream and false news. *Proceedings of the National Academy of Sciences*, 117(27): 15536–15545. https://doi.org/10.1073/pnas.1920498117.

Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., & Hendawi, A. 2025. *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference* (arXiv:2505.09598). arXiv. https://doi.org/10.48550/arXiv.2505.09598.

Johnson, H. M., & Seifert, C. M. 1994. Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6): 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3): 106–131. https://doi.org/10.1177/1529100612451018.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. 2020. Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. *Political Behavior*, 42(3): 939–960.

OECD. 2024. Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity. OECD Publishing, Paris. https://doi.org/10.1787/d909ff7a-en.

Pingree, R. J., Watson, B., Sui, M., Searles, K., Kalmoe, N. P., Darr, J. P., Santia, M., & Bryanov, K. 2018. Checking facts and fighting back: How fact-checking impacts elites and the mass public. *New Media & Society*, 20(3): 990–1010. https://doi.org/10.1371/journal.pone.0208600.

Porter, E., Wood, T. J., & Babak, B. 2019. Can Presidential Misinformation on Climate Change Be Corrected? Evidence From Internet and Phone Experiments. *Research and politics*, 6(3). https://doi.org/10.1177/2053168019864784.

Thorson, E. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3): 460–480. https://doi.org/10.1080/10584609.2015.1102187.

United Nations Development Guide (UNDG). 2017. UNDAF Companion Guidance: Theory of Change. https://unsdg.un.org/resources/theory-change-undaf-companion-guidance.

United Nations General Assembly (UNGA). 2015. Transforming Our World: The 2030 Agenda for Sustainable Development A/RES/70/1. https://docs.un.org/en/A/RES/70/1.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3): 350–375. https://doi.org/10.1080/10584609.2019.1668894.

Wintersieck, A. L. 2017. Debating the Truth: The Impact of Fact-Checking During Electoral Debates. *American Politics Research*, *45*(2): 304–331. https://doi.org/10.1177/1532673X16686555.

Züger, T., & Asghari, H. 2023. AI for the public. How public interest theory shifts the discourse on AI. *AI & SOCIETY, 38*(2): 815–828. https://doi.org/10.1007/s00146-022-01480-5.

Appendix

Table 3: Overview of the topics used by the topic classifier, based on the NewsCodes Media Topic taxonomy (source here, additional info here).

Name	Description
Arts, culture, entertainment and media	All forms of arts, entertainment, cultural heritage and media
Conflict, war and peace	Acts of socially or politically motivated protest or violence, military activities, geopolitical conflicts, as well as resolution efforts
Crime, law and justice	The establishment and/or statement of the rules of behaviour in society, the enforcement of these rules, breaches of the rules, the punishment of offenders and the organisations and bodies involved in these activities
Disaster, accident and emergency incident	Man made or natural event resulting in loss of life or injury to living creatures and/or damage to inanimate objects or property
Economy, business and finance	All matters concerning the planning, production and exchange of wealth
Education	All aspects of furthering knowledge, formally or informally
Environment	The protection, damage, and condition of the ecosystem of the planet Earth and its surroundings
Health	All aspects of physical and mental well-being
Human interest	Item that discusses individuals, groups, animals, plants or other objects in an emotional way
Labour	Social aspects, organisations, rules and conditions affecting the employment of human effort for the generation of wealth or provision of services and the economic support of the unemployed
Lifestyle and leisure	Activities undertaken for pleasure, relaxation or recreation outside paid employment, including eating and travel

Table 3: Overview of the topics used by the topic classifier, based on the NewsCodes Media Topic taxonomy (source here, additional info here).	
Politics and government	Local, regional, national and international exercise of power, the day-to-day running of government, and the relationships between governing bodies and states
Religion	Belief systems, institutions and people who provide moral guidance to followers
Science and technology	All aspects pertaining to human understanding of, as well as methodical study and research of natural, formal and social sciences, such as astronomy, linguistics or economics
Society	The concerns, issues, affairs and institutions relevant to human social interactions, problems and welfare, such as poverty, human rights and family planning
Sport	Competitive activity or skill that involves physical and/or mental effort and organisations and bodies involved in these activities
Weather	The study, prediction and reporting of meteorological phenomena

Table 4: Overview of the claim types used by the claim type classifier, including short description for each claim type.	
Claim type	**Description as in Full Fact AI documentation**
Personal	"Statements about people's personal experiences can be facts, but usually aren't capable of being checked"
Quantity	"These are often statistical or numerical in nature"
Correlation	"One event causes or correlates with another"
Rules	"These refer to sentences about the laws and rules of a country (or region) and procedures of public institutions"
Predictions	"These are hypothetical claims about the future"
Voting	"Claims about official votes, including parliamentary votes, elections, referendums"
Opinion	"Claims about public opinion, whether a formal opinion poll, survey, or otherwise"
Support	"Claims that someone supports a policy or position or has suggested support"

Table 4: Overview of the claim types used by the claim type classifier, including short description for each claim type.	
Other	-
Not claim	-