ARTIFICIAL INTELLIGENCE, EXTREME SPEECH, AND THE CHALLENGES OF ONLINE CONTENT MODERATION

POLICY BRIEF

Sahana Udupa, Elonnai Hickok, Antonis Maronikolakis, Hinrich Schuetze, Laura Csuka, Axel Wisiorek, Leah Nann

AI4Dignity is a Proof-of-Concept Project hosted at LMU Munich and funded by the European Research Council (2021–2022) under the Horizon 2020 research and innovation program (grant agreement number: 957442).

Project Website: ai4dignity.gwi.uni-muenchen.de

Research program: fordigitaldignity.com

Twitter: @4digitaldignity

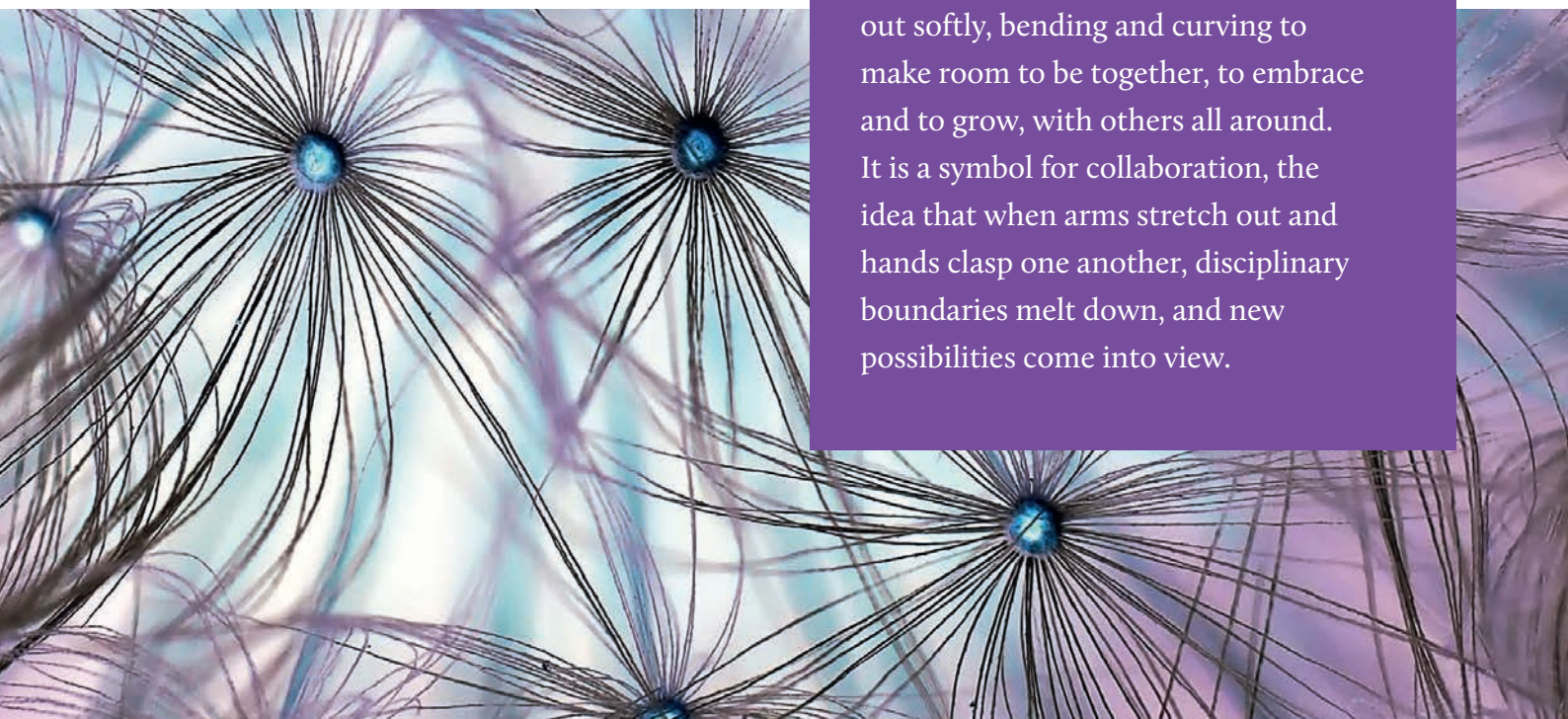
To cite: Udupa, Sahana, Elonnai Hickok, Antonis Maronikolakis, Hinrich Schuetze,Laura Csuka, Axel Wisiorek, Leah Nann. 2021. "AI, Extreme Speech and the Challenges of Online Content Moderation". AI4Dignity Project, https://doi.org/10.5282/ubm/epub.76087

European Research Council

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

AI4DIGNITY

TABLE OF CONTENTS



The delicate petals of dandelion spread out softly, bending and curving to make room to be together, to embrace and to grow, with others all around. It is a symbol for collaboration, the idea that when arms stretch out and hands clasp one another, disciplinary boundaries melt down, and new possibilities come into view.

ARTIFICIAL INTELLIGENCE, EXTREME SPEECH, AND THE CHALLENGES OF ONLINE CONTENT MODERATION

POLICY BRIEF

SUMMARY

Online extreme speech has emerged as a significant challenge for democratic societies worldwide. Governments, companies, and academic researchers have responded to this phenomenon by increasingly turning to Artificial Intelligence (AI) as a potential tool that can detect, decelerate, and remove online extreme speech.

AI deployment is expected to bring scalability, reduce costs, and decrease human discretion and emotional labor in the removal of objectionable content. However, even though digitalization is now a global phenomenon, AI tools for extreme speech detection that are globally applicable, inclusive, and yet resource-efficient and feasible are lacking.

In this policy brief, we outline the challenges facing AI-assisted content moderation efforts, and how the collaborative coding framework proposed by the European Research Council's Proof-of-Concept project 'AI4Dignity' (2021–2022) offers a way to address some of the pertinent issues concerning AI deployment for content moderation. The project's key objective is to operationalize the 'human-in-the-loop' principle by developing a community-based **human–machine process model with curated space of coding** to detect and categorize extreme speech. The methodology is based on collaborations between fact checkers, ethnographers, and AI developers.

Building on ongoing AI4Dignity project experiences, this policy brief will provide a short review of policy and corporate practices

around AI and content moderation, highlight existing challenges, discuss what lessons can be learned from ongoing efforts, and underline what new areas and questions are to be charted on priority. In the current context where the excitement around AI's capacities has run up against anxieties about the development and deployment of the technology, this policy brief will propose ways to develop context-sensitive frameworks for AI-assisted content moderation that are centered around human collaboration.

Our recommendations:

- Social media companies and governments should institutionalize people-centric frameworks by reaching out to communities and incorporating feedback to shape the future development of AI assisted content moderation.

- Social media companies should directly recruit content moderators and involve communities, such as fact checkers as well as academic researchers, on a fair and regular basis in the development and implementation of content moderation policies and practices. They should incorporate community inputs as part of the regular job mandate of in-house AI developers rather than as extraordinary and episodic arrangements during critical events like elections or under the banner of corporate social responsibility.

- Beyond company practices, collaborative models for identifying extreme speech independent of corporate and government spaces need to be fostered and supported. The intermediation we have built in AI4Dignity could be adopted for community involvement in a systematic and transparent manner in ways that AI researchers and developers remain in constant conversation with communities and academic researchers.

- Safeguards need to be put in place to avoid political or other misuse of collaborative AI models as well as to monitor their effectiveness and impact and adjust models accordingly.

The approaches developed in the AI4Dignity project are aimed at evolving responsible practices around AI-assisted content moderation as part of a broader effort at tackling online extreme speech. Admittedly, online extreme speech is a larger social and political problem, which cannot be addressed without accounting for the structures and impacts of repressive regimes and oppressive histories.1 These larger questions should bear on AI-based content moderation systems while AI's potential for content moderation as a specific node in the larger problem should be addressed in its fullest possible scope.

EXTREME SPEECH AND CONTENT MODERATION: SCOPE

In this policy brief (and the AI4Dignity project), we define extreme speech as expressions that challenge and stretch the boundaries of legitimate speech along the twin axes of truth/falsity and civility/incivility. Extreme speech research stresses that the same expression can be repressive or subversive based on the context (speaker, target, historical factors, and technology). It also foregrounds evolving digital practices, including recent trends of hateful language that comes cloaked in 'funny' memes and wordplay. Following this emphasis, the AI4Dignity project utilizes a tripartite definition: derogatory extreme speech, exclusionary extreme speech, and dangerous speech.

Derogatory extreme speech refers to expressions that do not conform to accepted norms of civility within specific regional/local/national contexts and target people/groups based on racialized categories or protected characteristics (caste, ethnicity, gender, langauge group, national origin, religious affiliation, sexual orientation) as well as other groups holding power (state, media, politicians).[2] It includes derogatory expressions not only about people but also about abstract categories or institutions that they identify targeted groups with. It includes varieties of expressions that are considered within specific social-cultural-political contexts as "the irritating, the contentious, the eccentric, the heretical, the unwelcome, and the provocative, as long as such speech did not tend to provoke violence".[3]

Exclusionary extreme speech refers to expressions that call for or imply exclusion of historically disadvantaged and vulnerable people/groups from the "in-group" based on caste, ethnicity, gender, langauge group, national origin, religious affiliation, and/or sexual orientation. These expressions incite discrimination, abhorrence and delegitimization of targeted groups.[4] The label does not apply to abstract ideas, ideologies, or institutions, except when there are reasonable grounds to believe that attacks against ideas/ideologies/institutions amount to a call for or imply exclusion of vulnerable groups associated with these categories. For example, if attacking a particular religion in a specific context has a reasonable chance to trigger violence against people who practice this religion, such expressions would fall under 'exclusionary extreme speech'.

Dangerous speech refers to expressions that have a reasonable chance to trigger/catalyze harm and violence against target groups (including ostracism, segregation, deportation, and genocide).[5]

We further recognize that the use of AI in content moderation primarily involves the deployment of machine learning models in automated actions and decisions around (user generated and paid) content with text, image, video, and/or audio. Our reference to AI includes actions that have also been described as 'automated content moderation'. Such actions can center around detecting content (flagging and labelling content), removing content (blocking, suspending, and removing content), and curating content

(recommending content, making content more or less prominent, and ranking content). While recognizing the significant and important efforts that are in motion to bring more transparency into how companies use automated tools to curate content, the AI4Dignity project is focused on the use of AI for **delineating** and **detecting** problematic content since this constitutes the first step towards other content related actions with direct and immediate implications for public discourse and democratic possibilities.

Through technological mediation, as is widely acknowledged, social media platforms have offered several avenues for user-based communication and interaction to unfold across public and quasi-public domains. This mediation includes changes in who can create content, the scope and speed by which it can be shared, the range of content types including the weight and legitimacy that any piece of content is given and why. The generation, sharing, consumption, and monetization of content on these platforms is a key part of their business model, and companies have developed algorithms to personalize and optimize end-user engagement with content based on metrics of online user practices and the 'data traces' they leave behind.[6] In the words of tech ethicist Tristan Harris, this "extractive attention economy" has sparked a "race to the bottom of the brain stem…which is this race to reverse engineer human instincts".[7] This has created a situation where harmful content may be elevated or prioritized because of the attention it gains on a platform.[8]

Given the sheer scale of content on these platforms, the potential speed and reach of such content, and the desire that harmful content never reaches publication on a platform or is removed as soon as possible, the use of AI is increasingly being looked upon as a solution by both companies and governments.

PRIVATE SECTOR USE OF AI FOR CONTENT MODERATION

Online platforms have faced increased scrutiny for failure to address extreme speech and other problematic content on their platforms as well as the opaque functioning of algorithms on their platforms.[9] Companies have used AI in content moderation to augment decisions taken by human moderators through actions like detecting, flagging, prioritizing, and ranking content.[10] Yet, critical insights into when and how these techniques are used are difficult to gain. Social media companies have revealed their use of automated content moderation tools with variable transparency.

Twitter and Facebook have maintained regular online publications to share the details of their usage of machine learning, often also to invite ideas from the broader tech world and to proactively avoid state regulatory actions. Twitter Engineering's 'Insights' blog[11] offers glimpses into the company's recent technological advancements, including the extent to which such innovations have been relevant to its practices. Facebook Engineering has a similar forum for all AI-related topics, and a separate

website for more in-depth material.[12] In addition, it is a regular practice for major social media companies to invite participation in the programming process by hosting coding 'challenges', conducting open surveys, and seeking collaboration with external researchers.[13]

Publicly available Twitter and Facebook documents have outlined some general aims regarding AI use in their operations. Refining the tools for content moderation has been taken as a high priority, which includes the combating of misinformation,[14] trolling, and abuse,[15] as well as efforts for achieving broader language functionality.[16] Examples of AI tools for content moderation that companies have developed and publicly shared demonstrate that platforms are looking to develop models for personalization purposes and the moderation of certain types of content. Rosetta from Facebook, a machine learning model that recognizes and analyses images and texts together and is being used to improve the accuracy of photo search, enforces the company's hate speech policy, improves the classification of photos in NewsFeed, and surfaces more personalized content.[17] Facebook has also developed DeepText to undertake intent, sentiment, text, and visual analysis to enable personalization.[18] Reports have noted Facebook's use of 'whole post integrity embeddings', which analyze text, images, videos, transcribed text from recordings, interactions between users, external context, and knowledge base information.[19]

Company implementation reports submitted by signatories of the European Union's (EU) Code on Disinformation also shed light into the ways that AI is being applied. According to Google's report, the company uses automated classifiers to ensure their policies on political advertisements are adhered to.[20] Facebook uses AI tools to identify clickbait at the individual post level in addition to the domain and page level. The company utilizes AI in combination with human review processes to identify, capture, and verify cloaking for removal. AI is also used to assess whether a page should be down-ranked.[21] In Twitter's report, the company notes that automation detection methods are being increasingly used to identify misuse of the platform before users are impacted.[22]

Though optimistic and clear about the necessity of AI in content moderation, companies are being realistic about their shortcomings. Twitter has noted an increase in the use of automated tools but also recognizes that the use of automation may lack the context needed to make an accurate determination, and thus the company will not use the technology for critical actions like permanently suspending an account.[23] Though still faced with challenges such as language and access to training data, companies have claimed that the accuracy of these tools is increasing.[24] Two newly launched large-scale projects, *Responsible ML* (Twitter) and *Fairness Flow* (Facebook), build on an understanding that these systems are not perfect and require improvement.[25] Twitter's *Responsible ML* initiative that was launched in April 2021, for example, aims to provide end users with more information on how machine learning technology is used in the service. The program is built around four proposed aims: taking responsibility for the companies' algorithmic decisions; ensuring equity and fairness of outcomes; ensuring

transparency about algorithmic decisions; and how they were reached, and enabling agency and algorithmic choice. The initiative aims to research potential harms that can emerge from the use of algorithms and find ways to address these and build end user control into products and experiences shaped by algorithms.[26]

However, studies have also shown how the classification algorithms are limited by the homogenous workforce of technology companies that employ disproportionately fewer women, minorities, and people of color.[27] Some initiatives have tried to address these limitations by involving users' experiences and opinion.[28] Google's Perspective API and Twitter's Birdwatch have experimented with crowd sourcing models for determinations around content. Launched in 2021 as a pilot, Birdwatch allows users to label information in tweets as misleading and provide additional context. Google's Perspective API offers "toxicity scores" to passages based on user inputs feeding the machine learning models. Although efforts that leverage 'crowd intelligence' are promising, studies have exposed that they can result in false positives as well as racial bias.[29] Some studies have argued that crowd sourced models have the problem of unevenness. Whereas racist and homophobic tweets are more likely to be identified as hate speech, gender-related comments are often brushed aside as merely offensive speech.[30] Lack of public oversight can also give companies unchecked discretionary power to label and remove online content. To address this issue, Twitter's Birdwatch has proposed some important measures. When consensus has been reached by a diverse set of contributors, the notes will be made publicly visible. This means the data contributed to Birdwatch will be made publicly available and downloadable alongside the code for reputation and consensus systems and the ranking system for Birdwatch.[31]

While such proposals are at different stages of implementation, there is still the looming problem of transparency. Notwithstanding the various assurances of social media companies, it can still be unclear to the end user when a decision or action was taken with AI and how it shaped their experience. In the 2020 index of the Ranking Digital Rights (RDR) Project, companies were assessed on specific questions: whether they publish policies that "clearly describe the terms for how they use algorithmic systems across their services and platforms, and if they publish a clear and accessible policy stating the nature and functions of these systems".[32] The RDR methodology recommends that companies "publish clear policies describing their use of algorithmic curation, recommendation, and ranking systems, including the variables that influence such systems". The 2020 assessment found that of the companies assessed, no company provided full disclosure on how users' online content is curated, ranked, or recommended. And only some provided explanations that were easy to find and understand.[33]

AI IN CONTENT MODERATION: OVERVIEW OF CHALLENGES

Such limitations signal a broader challenge facing AI use in content moderation, and the stakes of addressing problematic content online. Though the use of AI in content moderation can potentially be effective in removing certain categories of content and augmenting human decisions, relying on AI for content moderation also comes with risks that can impact users' rights and affect democratic systems of inclusion and justice. How AI systems for content moderation are designed, developed, trained, used, communicated, and governed can have far-reaching impacts upon what content is allowed, magnified, or deleted, and consequently, upon diverse contestations expressed by such content and the multiple publics who drive them.

Challenges in using AI for content moderation can include:

- **Inability to fully account for evolving context and practice:** AI does not necessarily have the ability to understand the context, intent, linguistic nuances, and cultural variation, and the changes that occur around these required to evaluate content. Some formats of content can also be more difficult for AI to identify and correctly moderate, including audiovisual content.[34] It can also be difficult for AI to account for aspects like misspelled words, changing syntax, and the use of images, GIFS, and memes to convey offense.[35] This can result in the removal of legal content, including reportage of extremist events and political or dissenting speech, and raises concerns about accuracy, potential censorship, and implications for freedom of expression.[36]

- **Biased/inaccurate/limited training data**: AI can be trained on limited, skewed, or biased data sets, resulting in decisions that are inaccurate and/or reproduce these limitations.[37] It can also result in decisions that are inconsistent across contexts. This raises concerns of potentially discriminatory content moderation practices and the removal of legitimate content—particularly with respect to speech from minoritized and historically disadvantaged groups.[38]

- **Linguistic diversity**: While companies are continuing to invest in natural language processing (NLP) models that cover a diversity of languages[39] and large global languages including English, Spanish, and Mandarin are often covered by existing AI models, smaller languages and those spoken in poorer countries have been left behind except where international outcry has placed pressure on social media companies to step up efforts to respond to humanitarian crises (as evidenced in Facebook's efforts in 2021 to ramp up regional language capacities for Myanmar).[40] The lack of linguistic diversity can result in extreme speech being unidentified or misidentified.

- **Function creep**: Relying predominantly on AI to monitor and moderate content online can result in function creep: "the spilling over of technologies devised for certain purposes into other areas, uses, purposes, with

impacts on safety, privacy and bias."[41] The broader monitoring of users online has raised concerns about the impact on user privacy.[42]

- **Opaque decisions and lack of transparency**: A lack of transparency around how AI is developed and trained for content moderation purposes can result in opaque decisions escaping public scrutiny.[43]

- **Lack of notice**: A lack of notice on when and how AI is being used in decisions pertaining to content limits the ability for users to appeal the decision.[44]

- **Proactive moderation**: The use of AI to moderate content prior to publication can result in censorship and a lack of due process for the users as they may not be aware that their content was filtered/restricted or have the ability to appeal the decision.[45]

- **Shifting the burden of determining unlawful content**: Legal requirements for companies to use AI in moderating content shifts the burden of determining legality to companies and removes important safeguards such as judicial review.[46]

- **Authoritarian use**: Governments can use AI to facilitate or mandate authoritarian or unlawful censorship practices that are in violation with international human rights law.[47]

- **Amplification of harmful content**: When AI is used in systems that recommend and prioritize content, this can amplify problematic content depending on the characteristics that the AI has been trained to prioritize.[48]

REVIEW OF POLICY

There are a number of policy proposals emerging that both recognize the challenges and potential harms that can emerge when AI is put to use for content moderation. These proposals have evolved in conjunction with shifting national and regional regulations around online misinformation, disinformation and hate speech. For instance, in Kenya, the Computer Misuse and Cybercrimes Act (2018) penalizes, among other things, content that amounts to false publication of communications and communications understood by the recipient as indecent or grossly offensive. In Brazil, the Internet Freedom, Responsibility and Transparency Bill (2020), developed in response to disinformation, requires companies to take steps such as traceability of end users and the use of "technical means'" to monitor their platforms for misinformation and unauthorized/fake accounts.[49]

Broadly, policies have recommended and encouraged the investment in and use of the technology for different purposes:

- *Demoting* the ranking of content that exposes users to problematic content.[50]
- *Prioritizing* relevant, authentic, and accurate and authoritative information where appropriate in search, feeds, or other automatically ranked distribution channels.[51]
- *Identifying, reviewing,* and *reducing* illegal content.[52]
- *Redirecting* users from problematic content.[53]
- *Promoting* counter narratives.[54]

In particular, the proposed EU Digital Services Act,[55] the United Kingdom's (UK) Online Harms White Paper (Online Harms White Paper),[56] and the Indian Guidelines for Intermediaries and Digital Media Ethics Code Rules 2021 (Indian Intermediary Guidelines)[57] are examples of policy commitments, self-regulatory codes, and (draft) legislations that have created (draft) frameworks for the use of automated tools in content moderation. These frameworks focus on developing oversight, accountability, and transparency mechanisms as well as defining configurations for the use of automated tools in content moderation.

These proposals have highlighted several important regulatory elements:

- **Mandatory nature and authorization:** Different configurations have emerged on whether the use of automated tools for content moderation can be mandated or if this determination will remain with companies. For example, para 30 in the UK Online Harms White Paper proposes that the Authority will have the power to require that companies use automated tools whereas section 4(4) of the Indian Intermediary Guidelines encourage companies to endeavor to deploy these technologies but does not mandate the use.

- **Principles to guide use:** Different principles are being defined to guide the use of automated tools. For example, under para 2.62 Interim Codes of Practice, the UK Online Harms White Paper recommends that determinations for use of automated tools should be guided by the persistence of the illegal content and based on a determination of whether there are no less intrusive means for identifying and removing the content. Section 4(4) of the Indian Intermediary Guidelines requires that the use of such tools and subsequent actions need to be proportionate with respect to free speech and privacy.

- **Instances for use:** Proposed instances for use have focused on specific types of clearly illegal content but have not always been clear if the technology should be used only on public spaces on platforms or if it should also be used on private channels. Under para 2.62 Interim Codes of Practice, the UK Online Harms White Paper proposes the use of the technology for identifying illegal content such as child exploitation and terrorist content, and para 30 notes that this includes on private channels if necessary. Section 4(4) of the Indian Intermediary Guidelines recommend the use of the technology to identify content depicting rape, child sexual abuse or conduct, and any information that is identical to content previously removed. The Guidelines do not clarify if the technology should be used on public and private channels.

- **Human–AI collaboration:** Different frameworks for human-AI collaboration are emerging with respect to whether automated tools should augment human decisions or take decisions, and the extent of human oversight needed for the same. Para 2.60 under Interim Codes of Practice in the UK Online Harms White Paper recommends that automated tools should be used to identify, flag, block, or remove illegal content for human review as

opposed to taking an independent decision. Section 4(4) of the Indian Guidelines for Intermediaries require the establishment of human oversight measures including a review with regard to the accuracy and fairness of such tools, the propensity of bias and discrimination in such tools and the impact on privacy and security of such tools.

- **Risk assessment and review:** Different structures for risk assessment and review are emerging for both automated content moderation and the algorithmic amplification of content. Article 26 of the Digital Services Act will require very large online platforms to undertake a risk assessment that includes a review of how their content moderation systems, recommender systems, and systems for selecting and displaying advertisements influence any of the systemic risks identified. Section 4(4) of the Indian Guidelines for Intermediaries requires a review of automated tools with regard to the accuracy and fairness of such tools, the propensity of bias and discrimination in such tools, and the impact on privacy and security of such tools.

- **Notice:** Different configurations for communicating the use of automated tools are beginning to emerge including who must provide the notice, what the notice needs to contain, who receives the notice and the timeframe for giving notice. Para 2.62 of the Interim Codes of Practice in the UK Online Harms White Paper recommends that the Regulator provides notice to the public when a decision is made to require a company to use automated tools. Article 14 of the EU Digital Services Act requires that when responding to a notice of illegal content submitted by a user, intermediaries must, among other things, communicate if and where automated tools were used as part of the decision-making process. Article 15 of the EU Digital Services Act requires that this information is also communicated to the user if content is removed and must include a distinction to reveal whether automated tools flagged the content or were used to take the final decision to remove the content.

- **Transparency and reporting requirements:** Structures of transparency and reporting requirements that are emerging have proposed different frameworks for communicating when automated tools are used, how they are used, and what rights users have with respect to decisions augmented or taken by automated tools. Article 12 of the EU Digital Services Act requires that intermediaries communicate in their Terms of Service information on any policies, procedures, measures, and tools used for the purpose of content moderation, including algorithmic decision-making and human review. Article 13 requires intermediaries to publish at least once a year information about content moderation engaged in at the providers' own initiative. This contains the obligation to present the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorized by the type of reason and basis for taking those measures.[58] Article 23 requires that

providers of online platforms report every six months, among other things, any use made of automatic means for the purpose of content moderation, including a specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes, and any safeguards applied.

- **Trusted flaggers:** Through the category of "trusted flaggers", Article 19 of the EU Digital Services Act proposes to expedite the process of notices, complaints, and removal of illegal content. Online platforms are obligated to process and decide on the notices submitted by trusted flaggers "on priority and without delay". Trusted flagger status is "awarded to entities and not individuals that have demonstrated…that they have particular expertise and competence in tackling illegal content, that they represent collective interest, independent from any online platform…and that they work in a diligent and objective manner." This measure is geared towards bringing community perspectives to platforms' decisions around content moderation.

COLLABORATIVE MODELS: TOWARDS PEOPLE-CENTRIC APPROACHES

As the above discussion demonstrates, the challenge of AI use for content moderation is multidimensional, involving different regulatory principles, foremost of which are transparency, timeliness, accuracy, due diligence, and accountability. While available frameworks are making advancements in addressing the key challenges, ongoing efforts also reveal fragmented responses at the national, regional, and international levels, lacking coordinated efforts to achieve global standards. In addition, by focusing on social media companies as the key agents of action, several regulatory frameworks have tended to take the questions around AI deployment further away from community-based interventions and people-centric systems of accountability. This is not to say that social media companies are any less responsible for ensuring accountable AI practices for content moderation. On the contrary, the key challenge is to ensure corporate actions are embedded within a rigorous system of accountability where communities have a direct stake in how AI systems are designed and deployed for moderating online speech. On this aspect, there are wide disparities between, for instance, the EU's proposals to develop a systematic structure to involve communities as "trusted flaggers" and national legislations such as the Indian Guidelines for Intermediaries that have remained vague on processes of human oversight.

As high-level principles, the latest EU regulations have laid out the agenda for human oversight

for AI use. These can provide guidance on the key characteristics that should be maintained when structuring different gradients of human–AI collaboration. For example, as per Article 14 of the 'EU Proposal for a Regulation laying down harmonized rules on artificial intelligence' (released in April 2021), individuals responsible for human oversight of high-risk AI systems should be cognizant of the possibility of automation bias and must be able to:

- understand the capacities and limitations and monitor the operation of the high-risk AI system
- correctly interpret the high-risk AI systems output
- decide when to and not to use high risk AI systems
- stop or interrupt the functioning of an AI system.[59]

Operationalizing such high-level principles, however, remains a huge challenge. In several regulatory frameworks, the emphasis on 'human oversight measures' as an appeal to abstract ethical principle is not accompanied with the procedural clarity and specific steps needed for implementation. The very abstract construction of 'human' or 'crowd' begs the question on who represents and constitutes these categories and what measures are needed in ensuring 'human oversight'.

The AI4Dignity Project seeks to unpack the ethical principle of 'human oversight' by embedding it within actual conditions of human–machine collaboration. It aims to anchor the abstract construction of 'human' to ethical and practical ways of involving communities beyond the purview of governments and corporates. The project further recognizes that imagining and creating such possibilities are invariably political in nature. It therefore requires a keen understanding of the specific social-political contexts within which different communities can be identified as reliable and legitimate partners in anti-hate initiatives.

AI4DIGNITY: RATIONALE, IMPLEMENTATION, AND INTERVENTIONS

Taking up one specific challenge in the broader set of issues outlined above, and building on the recognition that human oversight is invariably a political process, AI4Dignity focuses on operationalizing the "human-in-the-loop" principle by developing a **hybrid human–machine process model** with **curated space of coding** as a key approach to detect and categorize extreme speech.

The project centers the role of independent fact checkers as critical human interlocutors who can bring cultural contextualization to AI-assisted extreme speech moderation in a meaningful and feasible way. This approach follows directly from the two key challenges in AI deployment for content moderation. First, as discussed above, there is no catch-all algorithm that can work for different contexts. Lack of cultural contextualization has resulted in false positives and overapplication. Second,

hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings, satire, and coded language. The dynamic nature of online hate speech—where hateful expressions keep changing—adds to the complexity.

If human supervision is critical, it is then important to devise ways to **connect, support, and mobilize existing communities** who have gained reasonable access to meaning and context of speech because of their involvement in online speech moderation of some kind. Building spaces of direct dialogue and collaboration between AI developers and (relatively) independent fact checkers who are not employees of large media corporations, political parties or social media companies is a key component of AI4Dignity.

STEPS

The first step in the implementation of AI4Dignity has involved discussions among ethnographers, NLP researchers, and fact checkers to identify different types of problematic content and finalize the definitions of labels for manually annotating social media content. After agreeing upon the definitions of the three types of problematic speech as 'derogatory extreme speech', 'exclusionary extreme speech' and 'dangerous speech',[60] fact checkers were requested to label the passages under the three categories.

Each passage ranged from a minimum sequence of words that comprises a meaningful unit in a particular language to about six to seven full sentences. For the second step, fact checkers uploaded the passages via a dedicated WordPress site on to a database connected in the backend to the NLP model building. They also marked the target groups for each instance of labeled speech. On the annotation form, they identified the target groups from a dropdown list that includes "ethnic minorities, immigrants, religious minorities, sexual minorities, women, racialized groups, historically oppressed castes, indigenous groups and any other". Only under "derogatory extreme speech", annotators were also able to tick "politicians, legacy media, the state and civil society advocates for inclusive societies" as target groups. Fifty per cent of the annotated passages were cross-annotated by another fact checker from the same country to achieve inter annotator agreement score.

The third step (ongoing) is to create a collaborative coding space where AI developers and partnering fact checkers enter into an assisted dialogue to assess classification algorithms and the training datasets involved in creating them.

This dialogue will be facilitated by academic researchers specialized in the regions. Currently, fact checkers from four different countries—Brazil, Germany, India, and Kenya, are participating in the project.

AI4Dignity's process model aims to stabilize a more encompassing collaborative structure in which 'hybrid' models of human–machine filters are able to incorporate dynamic reciprocity between AI developers, academic researchers and critical community intermediaries such as independent fact checkers. The classification tool that is under development will reflect this collaborative annotation and iteration process. Figure 1 illustrates the basic architecture and components of this classification tool.

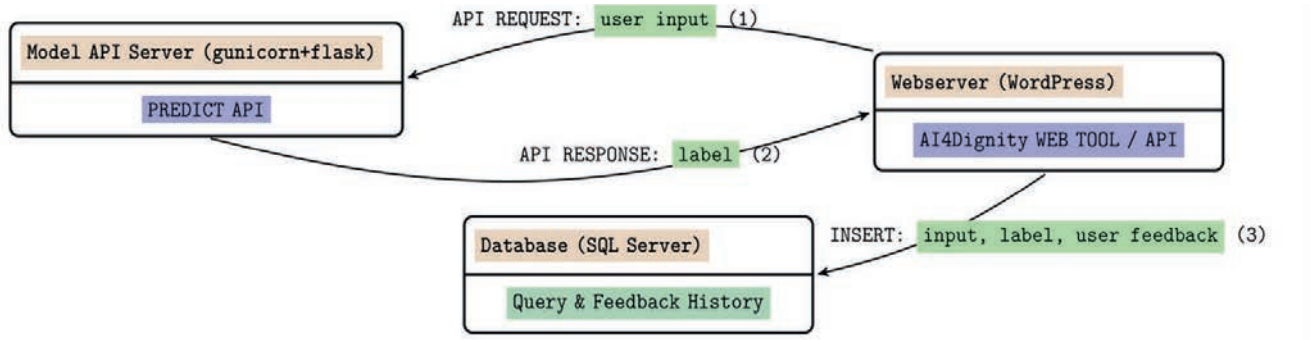


Figure 1: Architecture and components of the classification tool
(A full technical documentation will be provided on ai4dignity.gwi.uni-muenchen.de).

INTERVENTIONS

AI4Dignity's interventions relate to three interconnected areas:

1. The project addresses the vast problem of bringing cultural contextualization to big data sets by **rescaling** the challenge. It does so by identifying **fact checkers as a key stakeholder community** that can provide a **meaningful and feasible gateway into cultural variation** in online extreme speech. Fact checkers are distinct from other anti-hate groups because of their professional proximity to journalism. Exposed to volumes of disinformation data containing hateful expressions, they use—directly or indirectly—journalistic practices associated with checking and categorizing content. Therefore, they constitute a significant professional community in discussions around content moderation. The plan to involve fact checkers in this exercise comes with the risk of conflating two seemingly discordant objectives of extreme speech detection and anti-disinformation tasks. However, while they appear discordant, these speech forms come closely packed in actual practice. Without doubt, fact checkers are already overburdened with verification-related tasks, but they might still benefit from flagging extreme speech as a critical subsidiary to their core activities. Moreover, for factcheckers, this collaboration also offers the means to foreground their own grievances as a target community of extreme speech. By involving fact checkers, AI4Dignity therefore aims to draw upon the professional competence of a relatively independent group of experts who are confronted with extreme speech both as part of the data they sieve for disinformation and as targets of extreme speech. This way, AI4Dignity is creating a mechanism where the 'close cousin' of disinformation, namely extreme speech and dangerous speech, are spotted during the course of fact checkers' daily routines, without significantly interrupting their everyday verification activities.

2. AI4Dignity takes the 'hybrid' model of human–machine filters a step further. At present, initiatives implemented by entities such as the Media Authority of North Rhine-Westphalia have hybrid models that combine first stage filtering by algorithms with subsequent human verification.[61] The serialization model is limiting because of a lack of reciprocal reinforcements that can make machines more intelligible and human actors more tech ready. In place of serialization, AI4Dignity will bring **dynamic reciprocity** to the hybrid model.

3. The model does not use a classical keyword-based approach but allows for the creation of training artefacts through a **collaborative community-based classification approach**. By involving fact checkers, this model will advance Saleem and colleagues' pioneering effort to "leverage a community-based classification of hateful language".[62]

WHAT AI4DIGNITY IS DOING

With the curated space of coding as the central component of the Proof-of-Concept design, the project is generating the following:

1. Development of a **replicable process model** for collaborative AI involving different stakeholders by operationalizing the coding initiative described as **'Counterathon'**: a marathon of coding to counter online extreme speech. It brings two groups of key stakeholders in direct and systematic dialogue—AI/NLP developers and independent fact checkers—who will work in small national teams overseen by academic researchers.

2. An **AI4Dignity toolkit** published on a dedicated WordPress site to enable groups to organize Counterathon events in their respective locations and periodically publish updated extreme speech databases based on the outcomes of the event. The toolkit will help communities (in this case fact checkers) by providing: i) feature-rich technological tools for assisted dialogue with AI/NLP developers and academic intermediaries; ii) tools to classify content across the three categories (derogatory extreme speech, exclusionary extreme speech, and dangerous speech); iii) a classification API.

3. Extreme speech databases generated during the pilot Counterathon event and further contributions will be stored internally as relational SQL database for **research analysis,** including discerning patterns of extreme speech rhetoric and testing whether extreme speech data identified through Counterathon continues to persist on social media platforms (indicating inadequacies in corporate practices of content moderation).

AI4Dignity's collaborative effort is a departure from company-led initiatives. The project seeks to also overcome the limitations of unevenness in crowdsourcing through facilitated space of coding and annotations involving fact checkers as a key community. In addition, as described below, it opens up distinct pathways for the technical field of NLP and holds the possibility for overarching systemic effect in relation to AI-assisted content moderation.

BENEFITS FOR NLP AND SYSTEMIC EFFECTS

While NLP practitioiners and researchers have taken a keen interest in addressing online extreme speech, social NLP is still a fledging subfield of applied computational linguistics and consequently the proper tools and methodologies to deal with this problem have not yet been developed. Therefore, hate speech detection is currently being viewed as another variation of the staple NLP applications such as sentiment analysis or language inference. For these tasks, higher scores of the machine learning models are usually accepted as meeting the academic and professional standards. While for most tasks evaluation in such a lab setting is acceptable, for problems such as extreme speech with severe societal consequences, there is a need for a more nuanced approach—a recognition that NLP researchers have begun to increasingly take up as a challenge to tackle.

The common practice in the NLP field is that researchers themselves collect the data, usually by querying social networks for offensive language. This does not ensure a faithful depiction of the distribution of extreme speech in the real world. A related issue is the cultural, material, and social gap between NLP annotators and people who peddle and are affected by extreme speech. Even well-meaning NLP efforts might fall short if they lack cultural knowledge or transfer implicit bias, preconceived ideologies, and goals to machine learning models, beginning with the very definitions of different classifications employed in detecting and mapping extreme speech. Problems of contextualization are more pronounced when researchers work in the language fields of underrepresented and disenfranchised communities. Lack of understanding of the cultural and socio-political fields has thus impeded the progress of NLP researchers in the field of extreme speech mapping and analysis as well as in evolving a concrete framework of defining the labels.

Responding to these challenges, AI4Dignity has placed annotators from the fact-checking community in the four different countries at the center of the NLP methodology. As described in the preceding section, discussions were held between the annotators (fact checkers), ethnographers, and NLP researchers to identify the needs, goals, and course of action for the project. The definitions around what constitutes extreme speech and the fine-grained labels that capture possible targets of such speech were formulated by ethnographers with input from the annotators.

Iterations in annotations and definitions have sought to address the quality of data collection, which is the main problem in current NLP research on extreme speech. In terms of the machine learning models, the project is using the base BERT model and DistilBERT (a smaller model with fewer parameters that has been trained to mimic BERT's predictions via a process called distillation) that has been shown to perform adequately close to BERT. From the traditional machine learning models, the project is examining Support Vector Machines, Logistic Regression, and neural models based on LSTM (Long Short-Term Memory) units.

In addition to addressing specific challenges facing machine learning model development, AI4Dignity's pairing of fact checkers, NLP practitioners, and ethnographers—each with their own expertise and limitations—will contribute towards generating a systemic effect in the classification process. Triangulating these communities for collaborative coding beyond corporate and governmental systems advances the goal to bring more inclusive and culturally sensitive data sets for machine learning models. Such a triangulation process holds the potential to address systemic issues such as bias and lack of transparency in AI-assisted content moderation.

FUTURE DIRECTIONS AND FURTHER DEVELOPMENT OF COLLABORATIVE AI MODELS

Half-way into the project, AI4Dignity has identified key benefits of people-centric approaches to AI-assisted content moderation, as outlined in the preceding sections. However, there are significant areas for further reflections and development. The limitations, challenges, and recommendations for future development we highlight below will help to fine tune and strengthen the efforts for greater accountability and cultural contextualization in AI-assisted systems for content moderation.

- **Sparsity of data and interpretative labor:** One key challenge in people-centric models is the sparsity of data. Involving communities for annotations is resource intensive, although identifying critical intermediaries such as fact checkers can provide significant levels of scalability relative to the actual scope of cultural variation and dynamism in online extreme speech. Whereas NLP researchers have often used automated collection methods and employed easy-to-access and free/underpaid labor (e.g. Amazon Mechanical Turks) to gather large volumes of data, AI4Dignity stresses the importance of quality as well as fairness in human interpretative labor. Our data collection and annotation rely on contributions from extremely busy fact checkers with limited resources and time. Therefore, even though we are expecting the data to be more representative of actual vitriolic comments directed at different target groups, we are also expecting our data set to be significantly smaller.

- **Computational costs and internet access:** In involving diverse communities from regions with vastly different levels of internet access and technological resources, it is important to not only take into account the performance of machine learning models but also the size of the model and prediction time. For example, the tool should be able to work on low-end devices. Safeguards should also be put in place to ensure data is not lost in case of internet disconnection.

- **Autonomy and effectiveness of fact checking:** The role of fact checkers as independent community intermediaries is constrained and threatened under repressive and authoritarian conditions. While fact checkers can provide meaningful intermediation, their autonomy cannot be taken for granted. International standards for fact checking should constantly keep track of repressive attacks against fact checking as well as possible collusions between political interest groups and diverse players in the field of fact checking.

- **Open models for data annotation:** Another broad challenge is to build on existing efforts to develop open models for annotation that can incorporate human input at scale. Many models and projects that have been developed in the past are proprietary.[63] Since training models in NLP are often beyond what academic projects can afford, the challenge is to make use of publicly available, open-source models as a starting point for further training. AI4Dignity is using open-source models but more efforts in this direction require funding for and coordination among academic researchers. More important, social media companies should take an active and responsible role in funding infrastructures for open models.

- **Community collaboration for defining the policies that will guide the AI:** Further steps in people-centric efforts include ways to involve communities in the process of training AI systems *and* defining and designing the platform rules that those AI systems must enforce to reflect local and cultural realities and context.

- **Diverse human feedback loops:** Further focus should be on developing rigorous and transparent processes to streamline and consistently incorporate diverse and potentially conflicting human feedback into the development of AI for content moderation.

- **Collaborative graded approach:** Further emphasis should be on developing principles that could guide a graded approach to human–AI collaboration, such as type of content, nature of content, and potential of harm, to determine the extent to which the AI augments a decision vs. takes a decision. These should include standards that could guide the configuration of AI use—for example, determining when humans move from training an AI system to supervising an AI system.

- **Standards for human review:** As more community-based annotations are activated, standards for a human review process should be developed that considers the impact on human rights and potential discrimination and bias.

- **Measuring effectiveness:** Collaborative models built outside of corporate and governmental spaces—such as AI4Dignity—can advance existing academic and civil society organizations' efforts to measure the effectiveness of company policy and enforcement mechanisms, especially by comparing research data sets curated within these projects and actual instances

of takedowns/actions that social media companies have implemented on such content.

- **Privacy, security, and misuse:** Any effort at AI deployment should consider risks of privacy, security and dual use. AI4Dignity has sought to address this challenge by following a set of basic selection criteria to determine the (relative) independence of partnering fact checkers. Fact checkers have been selected and identified based on their active involvement in their respective regions in defending independent community efforts to address hate speech and disinformation and advance social justice goals. The International Fact-Checking Network (IFCN) certification has been an important criterion for selection and so are fact checkers' association with United Nations supported peace initiatives. Fact checkers who are part of political party campaign teams or full-time employees of social media companies are not involved in project partnership. The selection of independent fact checkers will reduce the risk of misuse and is likely to positively contribute towards upholding autonomous collaborative efforts and responsible technology to safeguard the interests of marginalized and vulnerable groups. Further efforts involve developing clear guidelines for the selection of partnering communities. The procedural benchmarks set out by the EU Digital Services Act in terms of vetting "trusted flaggers" and third-party operators provide useful benchmarks in developing selection standards.

- **Linguistic and cultural diversity:** Future development in people centric collaborative models will depend on involving communities (such as fact checkers) from diverse linguistic and cultural backgrounds. This will be a significant societal goal since current NLP models are heavily tilted towards large, resource rich languages and linguistic communities. Equally, the subtle intricacies of just a single language (use of irony, leetspeak, dog whistles etc.) and dynamically evolving hateful expressions require continuous engagement with communities through well delineated and tested processes of collaboration.

- **Multimodal content:** Online extreme speech that combines moving images, text and audio—as exemplified by internet memes and GIFs—poses specific challenges to machine learning models. The process model envisaged in AI4Dignity requires further expansion to incorporate multimodal elements of hateful expressions.

- **Impact assessment reports/audits:** Processes of community-centric AI development should develop standards for impact assessment and audits. Such standards should complement broader regulatory directions around impact assessment and audits; for instance, those stipulated by the Digital Services Act for large social media platforms.

- **Beyond category definitions:** The interface between AI/NLP, ethnographers and communities should extend beyond category definitions or achieving greater rigor in coding the categories. Other areas of cooperation could involve facilitating direct

interactions between critical communities and online posters for responsible actions around extreme speech (such as tools that can enable critical communities to directly provide positive narratives to posters) and automated tools for fact checkers and anti-hate groups to report extreme speech simultaneously on multiple social media platforms.

- **Industry practice:** Current arrangements of third-party moderation based on precarious contractual work should end. Social media companies should directly recruit content moderators and involve communities such as fact checkers on a fair and regular basis. The intermediation we have built in AI4Dignity should be adopted for community involvement in a systematic manner in ways that in-house AI researchers and developers remain in constant conversation with communities and academics to incorporate their inputs as part of their regular job mandate rather than as extraordinary and episodic arrangements during crisis situations or under the banner of corporate social responsibility. Corporates should design training processes so that AI researchers become acquainted with anthropological and sociological insights into online extreme speech and its implications. Through such measures, they should institutionalize the practice of reaching out to communities and bringing feedback to bear on the future development of AI-assisted content moderation.

- **Independent collaborations:** Beyond company practices, collaborative models independent of corporate and government spaces should be fostered and supported.

OTHER SELECTED RESOURCES FROM THE *FOR DIGITAL DIGNITY* RESEARCH PROGRAM

- Schick, Timo, Sahana Udupa, and Hinrich Schuetze. 2021. "Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP." ArXiv E-Prints. arXiv:2103.00453
- Udupa, Sahana. 2019. "Nationalism in the Digital Age: Fun as a Metapractice of Extreme Speech." International Journal of Communication 13: 3143–63.
- ———. 2020. "Artificial Intelligence and the Cultural Problem of Online Extreme Speech." Items, Social Science Research Council. 2020.
- Udupa, Sahana, Iginio Gagliardone, and Peter Hervik. 2021. Digital Hate: The Global Conjuncture of Extreme Speech. Bloomington: Indiana University Press.
- Udupa, Sahana, and Matti Pohjonen. 2019. "Extreme Speech and Global Digital Cultures." International Journal of Communication 13: 3049–67.
- Podcast: Decoding Hate: Episode 5/Moderating Global Voices (Katie Pentney and Sahana Udupa)

ENDNOTES

1 Udupa, Sahana, Iginio Gagliardone & Peter Hervik. 2021. Digital Hate: The Global Conjuncture of Extreme Speech. Bloomington: Indiana University Press.

2 Udupa, Sahana. 2021. "Digital Technology and Extreme Speech." United Nations Strategy Paper commissioned by the United Nations Department for Peace Operations. Also, Udupa, Sahana. 2017. "Gaali cultures: The politics of abusive exchange on social media." *New Media and Society* 20(4): 1506–1522.

3 Redmond Bate vs Director of Public Prosecutions before the Lord Justice Sedley and Justice Collins on July 23, 1999; *The Times,* July 28, 1999.

4 This follows from the definition of hate speech offered in the "United Nations Strategy and Plan of Action on Hate Speech: A Detailed Guidance on Implementation for United Nations Field Presences." https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf

5 Benesch, Susanne. 2012. *Dangerous speech: A proposal to prevent group violence.* New York: World Policy Institute.

6 Ranking Digital Rights. It's the Business Model: How Big Tech's Profit Machine is Distorting the Public Sphere and Threatening Democracy. New America. May 2020. Available at: https://rankingdigitalrights.org/its-the-business-model/

7 Harris, Tristan. Humane: A New Agenda for Tech with Tristan Harris - Presentation and Transcript.Ethical.net. May 5 2019. Available at: https://ethical.net/ethical/humane-new-agenda-tech-tristan-harris/

8 Munn, Luke. Angry by design: toxic communication and technical architectures. Nature - Humanities and Social Sciences Communications. July 30 2020. Available at: https://www.nature.com/articles/s41599-020-00550-7

9 Wiggers, Kyle. Stop the spread of harmful content. VentureBeat. November 13 2020. Available at: https://venturebeat.com/2020/11/13/facebooks-redoubled-ai-efforts-wont-stop-the-spread-of-harmful-content/

10 Vincent, James. Facebook is now using AI to sort content for quicker moderation. The Verge. November 13 2020. Available at: https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation

11 For more information see: https://blog.twitter.com/engineering/en_us/topics/insights.html

12 For more information see: https://engineering.fb.com/category/ml-applications/ and https://ai.facebook.com/

13 For more information see: https://blog.twitter.com/engineering/en_us/topics/insights/2020/what_twitter_learned_from_recsys2020.html https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/ https://ai.facebook.com/blog/deepfake-detection-challenge/ https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html

14 For more information see: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#moderation https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

15 For more information see: https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration.html https://ai.facebook.com/research/publications/abusive-language-detection-with-graph-convolutional-networks/

16 For more information see: https://engineering.fb.com/2018/08/31/ai-research/unsupervised-machine-translation-a-novel-approach-to-provide-fast-accurate-translations-for-more-languages/ https://engineering.fb.com/2018/01/24/ml-applications/under-the-hood-multilingual-embeddings/ https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html

17 For more information see: https://engineering.fb.com/2018/09/11/ai-research/rosetta-understanding-text-in-images-and-videos-with-machine-learning/

18 For more information see: https://engineering.fb.com/2016/06/01/core-data/introducing-deeptext-facebook-s-text-understanding-engine/

19 Wiggers, Kyle. Facebook's redoubled AI efforts to stop the spread of harmful content. VentureBeat. November 13 2020. Available at: https://venturebeat.com/2020/11/13/facebooks-redoubled-ai-efforts-wont-stop-the-spread-of-harmful-content/

20 Google. EU Code of Practice on Disinformation: Google Report. EU Commission. Available at: https://ec.europa.eu/information_society/newsroom/image/document/2019-5/google_-_ec_action_plan_reporting_CF162236-E8FB-725E-C0A3D2D6CCFE678A_56994.pdf

21 Facebook. Facebook Baseline Report on Implementation of the Code of Practice on Disinformation. EU Commission. Available at: https://ec.europa.eu/information_society/newsroom/image/document/2019-5/

facebook_baseline_report_on_implementation_of_the_code_of_practice_on_disinformation_CF161D11-9A54-3E27-65D58168CAC40050_56991.pdf

22 Twitter. Twitter Progress Report: Code of Practice against disinformation. Available at: https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019

23 @vigata and Derella, Matt. An update on our continuity strategy during COVID-19. Twitter. March 16 2020 Available at: https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html

24 Matsakis, Louise. Facebook's AI Can Analyze Memes, but can it understand them? Wired. September 14 2018. Available at: https://www.wired.com/story/facebook-rosetta-ai-memes/

25 Knight, Will. Facebook's Head of AI Says the Field Will Soon 'Hit a Wall'. Wired. December 4 2019. Available at: https://www.wired.com/story/facebooks-ai-says-field-hit-wall/ ; O'Brien, Chris. Twitter CTO on machine learning challenges: I'm not proud that we miss a lot of misinformation. Venturebeat. July 16 2020. Available at: https://venturebeat.com/2020/07/16/twitter-cto-on-machine-learning-challenges-im-not-proud-that-we-miss-a-lot-of-misinformation/; Vincent, James. YouTube brings back more human moderators after AI systems over-censor. The Verge. September 21 2020. Available at: https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns
For more information see: https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.html
https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/

26 For more information see: https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.html

27 Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

28 See Online Hate Index developed by Berkeley Institute for Data Science https://www.adl.org

29 Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. "The Risk of Racial Bias in Hate Speech Detection." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–78. Florence, Italy.

30 Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Langauge." ArXiv:1703.04009v1

31 For more information see: https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html

32 For more information see: https://rankingdigitalrights.org/index2020/indicators/F12

33 For more information see: https://rankingdigitalrights.org/index2020/indicators/F1d

34 https://firstdraftnews.org/latest/ai-moderating-media/

35 https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

36 https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/case-study-facebook

37 For example, a seminar on "AI, Hate Speech and Online Content Moderation in the Global South" held by Freie Univerität Berlin noted that many algorithms have been trained on content from the global north—resulting in the misinterpretation and identification of hate speech and dangerous speech. For more information see: https://www.polsoz.fu-berlin.de/en/kommwiss/arbeitsstellen/mediennutzung/news/2021/Seminar-Series-on-_AI_-Hate-Speech-and-Online-Content-Moderation_-_March-18---June-10_-2021_.html

38 https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf; https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/artificial-intelligence-and-the-cultural-problem-of-online-extreme-speech/

39 https://about.fb.com/news/2020/10/first-multilingual-machine-translation-model/ and https://www.infoq.com/news/2020/11/multilingual-ai-models/

40 https://about.fb.com/news/2021/02/an-update-on-myanmar/

41 Ponzanesi, Sandra. 2020. "Digital cosmopolitanism: Notes from the underground". *Global Perspectives*, 1(1): 12548

42 Llanso, Emma; Hoboken, Joris, Harambam, Jaron. "Artificial intelligence, content moderation, and freedom of expression". Transatlantic Working Group. February 26 2020. Available at: https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

43 Donovan, Joan. "Why social media can't keep moderating content in the shadows". November 6 2020. Available at: https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/

44 New America. The limitations of automated tools in content moderation. Available at: https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation/

45 Llanso, Emma J. 2020. No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society* 7(1): 1-6. https://journals.sagepub.com/doi/full/10.1177/2053951720920686

46 For example, the 2018 draft Intermediary Liability Rules in India proposed a provision (section 9) that would have required intermediaries to use automated tools to remove unlawful content or information. For more information see: https://www.meity.gov.in/writereaddata/files/Draft_Intermediary_Amendment_24122018.pdf

47 Freedom Online Coalition. FOC Joint Statement on Artificial Intelligence and Human Rights. Available at: https://freedomonlinecoalition.com/wp-content/uploads/2020/11/FOC-Joint-Statement-on-Artificial-Intelligence-and-Human-Rights.pdf

48 Llanso, Emma; Hoboken, Joris, Harambam, Jaron; Artificial Intelligence, Content Moderation, and Freedom of Expression. Transatlantic Working Group. February 26 2020. Available at: https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

49 Instituto de Technologia & Sociedade du Rio. Technical Note on Congressional Bills No. 2927/2020 and Bill No. 2630/2020. Available at: https://itsrio.org/wp-content/uploads/2020/06/ITS-Technical-Note.pdf These regulations have drawn criticism from digital rights groups for regulatory excess and threats to freedom of expression and privacy. See: Tech Against Terrorism. The Online Regulation Series: Brazil. Available at: https://www.techagainstterrorism.org/2020/11/11/the-online-regulation-series-brazil/; Freedom House. Brazil: Disinformation Bill Threatens Freedom of Expression and Privacy Online. June 29 2020. Available at: https://freedomhouse.org/article/brazil-disinformation-bill-threatens-freedom-expression-and-privacy-online

50 Australian Code of Practice on Disinformation and Misinformation. Digi. February 22 2021. Available at: https://digi.org.au/wp-content/uploads/2021/02/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-PDF-Feb-22-2021.pdf

51 European Commission. Code of Practice on Disinformation.2018. Available at: https://ec.europa.eu/digital-single-market/en/code-practice-disinformation

52 For example see: Australian Code of Practice on Disinformation and Misinformation. Digi. February 22 2021. Available at: https://digi.org.au/wp-content/uploads/2021/02/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-PDF-Feb-22-2021.pdf

53 The Christchurch Call. Available at: https://www.christchurchcall.com/call.html

54 The Christchurch Call. Available at: https://www.christchurchcall.com/call.html

55 European Commission. The Digital Services Act package. Available at: https://eur-lex.europa.eu/legal-content/en/T/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

56 Gov.UK. Consultation Outcome - Online Harms White Paper. April 8 2019. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/944310/Online_Harms_White_Paper_Full_Government_Response_to_the_consultation_CP_354_CCS001_CCS1220695430-001__V2.pdf

57 The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Available at: https://www.meity.gov.in/writereaddata/files/Intermediary_Guidelines_and_Digital_Media_Ethics_Code_Rules-2021.pdf

58 https://ec.europa.eu/info/sites/info/files/proposal_for_a_regulation_on_a_single_market_for_digital_services.pdf

59 European Commission. Proposal for a Regulation laying down harmonised rules on artificial intelligence. 2021. Available at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

60 See the section "Content Moderation and Scope" for the definitions.

61 Finck, Michèle. 2019. "Artificial Intelligence and Online Hate Speech." https://www.cerre.eu/sites/cerre/files/CERRE_Hate Speech and AI_IssuePaper_0.pdf

62 Saleem, Haji Mohammed, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. "A Web of Hate: Tackling Hate Speech in Online Social Spaces." ArXiv Preprint ArXiv:1709.10159

63 Laaksonen, Salla-Maaria et al. The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring.Big Data. February 5 2020. Available at: https://www.frontiersin.org/articles/10.3389/fdata.2020.00003/full

AUTHORS' BIOGRAPHIES

- **Sahana Udupa** is professor of media anthropology at LMU Munich, and the principal investigator of the For Digital Dignity research program. Her publications include *Making News in Global India* (Cambridge University Press); *Digital Hate: The Global Conjuncture of Extreme Speech* (Indiana University Press); Extreme Speech and Global Digital Cultures (*International Journal of Communication*), among others.

- **Elonnai Hickok** is a non-resident scholar at the Carnegie Endowment for International Peace and an independent expert. She has guided research with international organizations and has presented worldwide on issues of digital rights and emerging technology and the counterbalancing of governmental and individual interests and rights.

- **Antonis Maronikolakis** is an NLP research associate in Project AI4Dignity. He is a doctoral candidate at the Center for Information and Language Processing, LMU Munich, where he is working at the intersection of Natural Language Processing and Deep Learning.

- **Hinrich Schuetze** is professor and chair of computational linguistics and director of the Center for Information and Language Processing at LMU Munich. He is the principal NLP collaborator for Project AI4Dignity. He is the author of Introduction to *Information Retrieval* (Cambridge University Press). His publications are available at: https://www.cis.uni-muenchen.de/publications/.

- **Laura Csuka** is a doctoral candidate in computer science at the University of Oxford and has contributed as a research assistant for projects ONLINERPOL and AI4Dignity. She received her B.A. and M.A. in social and cultural anthropology from LMU Munich. Her current research is on natural language processing, machine learning, and misogynist online communities.

- **Axel Wisiorek** is a data science research associate in Project AI4Dignity and a doctoral candidate in general and computational linguistics at LMU Munich. He is a research assistant at the Center for Information and Language Processing (CIS) and LMU Center for Digital Humanities (ITG), where he is involved in the development of various web-based research projects in the field of digital humanities.

- **Leah Nann** is a scientific research assistant at Project AI4Dignity. She has received her B.A. in social and cultural anthropology from LMU Munich, where she is currently pursuing her M.A. in the same field with a focus on artificial intelligence and data capitalism.

Design: Florian Geierstanger, Miriam Homer

AI4DIGNITY