Symposium on AI, Free Speech, and Human Rights – Key Takeaways

On October 12 and 13, 2023, The Future of Free Speech (FFS) and Vanderbilt University brought together in Nashville, TN, thought leaders, researchers, activists, and industry professionals from around the world to discuss the implications of emerging Artificial Intelligence (AI) technology for freedom of expression and access to information. The event's title was Symposium on AI, Free Speech, and Human Rights.

This document, prepared by FFS, provides a summary of the key takeaways of the discussions that took place in the context of the Symposium. For a more complete and accurate account of the discussions, please watch the Symposium recordings.

Thursday, Oct. 12 – The John Seigenthaler Freedom Forum First Amendment Center, Nashville, TN

Welcome: Jacob Mchangama, Founder and Executive Director, The Future of Free Speech

Jacob Mchangama welcomed the audience and emphasized Vanderbilt University's deep commitment to free speech, as shown by initiatives like co-hosting this Symposium with FFS and Vanderbilt's Free Speech Week. FFS is dedicated to fostering a thriving and resilient global culture of free speech, conducting research and advocacy on freedom of expression.

Mr. Mchangama stressed the importance of analyzing the interaction between AI and free speech. Like the printing press in the 15th century, AI greatly facilitates the generation of all types of content. This content will likely impact the ecosystem of information and opinion as the printing press did in the past. Fearing new technologies, like AI, is not new; many voices warned about the risks of the printing press, the telegraph, the radio, and the internet. After their initial impact, most agreed that all these technologies benefited humanity. It is certainly possible that AI is different than previous technologies. Indeed, it is the first time in history that new communication depends very little on human input. A scribe needed to put a pen to papyrus, the printing press relied on the proper ordering of movable type, and social media platforms depend on user-generated content, even if algorithmic distribution has broad powers to determine who sees it. But generative AI is different; based on a few prompts, AI systems can deliver persuasive arguments, "reason," and perform what to humans appears as critical thinking.

Regardless of whether AI is different than any technology in the past, the best chance of making it one that benefits our species is to engage with these complex issues head-on. The Symposium explores the opportunities and challenges for free speech in the generative-AI context, like whether generative AI can reinvigorate free speech or whether we need a new free-speech framework to deal with AI.

Opening Remarks: Daniel Diermeier, Chancellor of Vanderbilt University

Chancellor Diermeier welcomed the audience and expressed Vanderbilt University's pride in sponsoring these essential conversations on AI, free speech, and human rights. The Chancellor stressed the university's commitment to promoting debate between people with different views – a great antidote to hatred and violence.

Discussing AI is urgent and crucial – AI is already widely used both in our personal and professional lives and its capabilities keep improving quickly. The questions – practical, ethical, and legal – surrounding AI are numerous, like how we can make sure that AI protects speech and champions

facts, and it is not used as a tool for disinformation and censorship or how can we eliminate bias and make AI as inclusive as possible. Vanderbilt University is conducting important research on issues like advancing cutting-edge innovation related to computing in AI and an AI model to predict and improve the treatment of depression.

Vanderbilt's Professor Jules White, who participated in the Symposium, developed one of the first mass online courses on prompt engineering. Vanderbilt University works under the motto of "radical collaboration" greatly encouraging cross-disciplinary work and research to bring diverse perspectives. Since its founding, the university has also committed to being a forum for all sorts of voices – for example, in the mid-60s it hosted controversial figures at the time, Martin Luther King Jr. And Stokely Carmichael. The premise is, as expressed by Vanderbilt's fifth Chancellor, that a university's obligation is not to protect students from ideas, but rather to expose them to ideas, and to help make them capable of handling and having ideas. Vanderbilt University continues to tackle difficult conversations head-on with initiatives like Dialogue Vanderbilt.

Panel No. 1: The First Amendment, Human Rights Standards and AI Governance

The first panel was moderated by Jacob Mchangama – Founder and Executive Director of FFS – and explored the intersection of generative AI, the First Amendment, and human rights standards as well as how companies deal with the tension between freedom of expression and other values.

Joan Barata – Senior Fellow at FFS, and Fellow of the Program on Platform Regulation at the Stanford Cyber Policy Center – discussed the applicability of Article 19 of the International Covenant on Civil and Political Rights (ICCPR) to generative AI. Article 19 is the key global rule in international human rights law concerning freedom of expression. Mr. Barata explained that the application of Article 19 to generative AI is still unexplored territory, but that this Article most likely protects AI-generated content. Article 19 encompasses the freedom to seek, receive, and impart information and ideas, regardless of how they are created – hence, even if they are generated by an algorithm. Mr. Barata pointed out that the implications of internet shutdowns for freedom of expression can provide a useful framework to assess the interaction between generative AI and this same freedom. In this regard, both limiting access to the internet and to AI content restrict citizens' access to information and can violate freedom of expression.

Eugene Volokh – Professor at UCLA Law School and, starting in 2024, Senior Fellow at the Hoover Institution at Stanford University – pointed out that the United States First Amendment protects AI-generated content. The First Amendment, in general, enables companies to share content – this means that AI providers have the right to generate information. In addition, like Article 19 of the ICCPR, the First Amendment protects users' right to access the information generated by AI and, should they wish so, to gather such information and distribute it (this right to gather and distribute AI content is similar to the right to record images and share them). Prof. Volokh also indicated that Section 230 does not protect AI providers for the content their systems generate,

unless the systems merely reproduce content from another source, like a third-party webpage. Hence, AI system providers could be held liable for content that is not protected by the First Amendment, such as libel.

Alexandria Walden – Google Policy Lead for Global Human Rights and Free Expression – talked about how Google balances freedom of expression with other values. Ms. Walden pointed out that Google determines and enforces its policies based on the purpose and benefits of each specific product. Google has committed to respect the ICCPR and, hence, human rights also guide its policies. Users' expectations regarding products are also crucial, particularly concerning trust and safety and products' guardrails. Given that many of its employees are from the U.S., Google is influenced by the First Amendment but also pays significant attention to international standards, especially Article 19 of the ICCPR, when considering use restrictions affecting their products. When dealing with overly restrictive government requests to take down content, Google reminds the authorities of the commitments both those governments and Google have signed up for under international human rights law.

Panel No 2: Improving Linguistic Inclusion in Large Language Models

The second panel was moderated by Jesse Spencer-Smith – Interim Director and Chief Data Scientist at the Vanderbilt University Data Science Institute – and explored the importance of inclusivity in AI models, specifically focusing on incorporating smaller languages outside the Western world to promote linguistic diversity and cultural representation.

Gabriel Nicholas – Research Fellow at the Center for Democracy and Technology – discussed the implications of polyglotism in generative AI. Mr. Nicholas stressed that the quality and quantity of text available to train Large Language Models (LLMs) significantly decreases from high-resource languages (e.g., English) to medium-resource languages (e.g., Hebrew) to low-resource languages (e.g., Basque). According to Mr. Nicholas, between 80% and 90% of the data used to train LLMs is in English. LLMs can generate and analyze content in low-resource languages, but their output is not as usable. He also pointed out that while multilingual LLMs increase inclusivity, polyglotism implies risks. Training a model with text in several languages can imply that it performs worse in these languages than if it had been trained in just specific languages or groups of languages. In addition, content moderation is context-specific and can be more difficult in multilingual models – for example, the same word may be an insult or not, depending on the language. Mr. Nicholas warned that companies are making trade-offs regarding polyglotism, performance, and safety, and the public is not aware of them.

Irene Mwendwa – Executive Director of Pollicy – stressed that limited language inclusivity has been and continues to be an issue in the digital sector at large – it is not limited to generative AI. This means that many citizens cannot use digital tools for their needs, for example, as an education tool or for their businesses. Pollicy has undertaken actions to increase technology inclusivity, but companies must act too. She also mentioned that, in Africa, women are particularly

disenfranchised – when they are verbally attacked, the abusive content is often not tackled due to the lack of resources and local digital expertise. Ms. Mwendwa also encouraged companies and stakeholders to work more closely with disenfranchised communities – Big Tech companies should help in democratizing access to technology, including by providing their products in more languages.

Julie Owono – Executive Director of Internet Without Borders – explained that traditionally there has been a lack of interest in providing internet access and internet-based tools to disadvantaged communities and regions. This interest has increased in recent times. Ms. Owono talked about how hate speech and disinformation are large and complex issues and referred to the limited resources devoted to tackling them, particularly in the Global South. Both good and bad faith actors, but especially the latter, test their capabilities in smaller subsets of populations that are not as well connected, for example, in matters of electoral interference. Ms. Owono also stressed the importance of companies devoting more resources to disadvantaged regions and communities, empowering local communities and talked about the need that African institutions increase their expertise and get involved in technology and AI.

Fireside Chat with Google

David Graff – Vice President of Google's Global Policy & Standards team, within Trust & Safety – discussed Google's approach to technology and generative AI. Mr. Graff stressed the benefits that can come from generative AI, for example, in the medical field, while referring to the need to be attentive to the risks that this technology generates. He also emphasized the value that generative AI can have to enhance learning, amplifying the opportunities that products like Google Search and YouTube already provide; he considers that the interactivity that generative AI provides can be particularly helpful. Regarding risks, Mr. Graff referred to the challenge of adopting policies that affect a very large and diverse group of people, especially given that policies need to be scalable. To define its policies, Google considers the purpose the product aims to achieve. Mr. Graff also stressed the importance of transparency, so the public is aware of Google's policies and can ask questions and challenge them.

He also stressed that generative AI presents new challenges since, contrary to previous products, it does not reproduce content but creates it. He considers that Google's previous experience with other products can be useful to address the opportunities and challenges generative AI brings. In order to limit potential harm by generative AI, Mr. Graff said that Google uses a safety by design approach, as it does with its other products, which implies extensive pre-launch testing. Regarding the balance between freedom of expression and safety, Mr. Graff said that Google focuses on the manner of expression and not the underlying ideas. The policies are not designed to protect users from being offended or ideas that they disagree with, they are designed to generate constructive discussions. Google also engages with stakeholders, including from underrepresented communities, and aims to have a diverse workforce. Generative AI may potentially allow for more region-specific policies, enabling the adoption and enforcement of several sets of rules that adjust

to local values. Mr. Graff also defended the need to have "smart" regulation for AI – he argued that excessively restrictive regulation can imply high compliance costs, which affects start-ups more than large corporations.

Panel No. 3: The Future of AI: Open-source or Centralization?

The third panel was moderated by Ole Molvig – Assistant Professor of History at Vanderbilt University – and delved into the merits and drawbacks of open-source AI models, advocating for collaboration, transparency, and democratization of AI innovations versus centralized approaches that prioritize control, security, and efficiency.

Peter Stern – Director of Content Policy Stakeholder Engagement at Meta – talked about Meta's approach to generative AI. Mr. Stern pointed out that Meta has released two versions of its LLM – one pre-trained model (the least finished version of the model which includes some basic safety modifications) and one fine-tuned one (trained in additional data sets that bring the model closer to conversational capacity). They are available on an open approach, so parties can take these models and then further train and adapt them to their needs. Mr. Stern stressed that the categories open source and proprietary should be thought of as a spectrum rather than a dichotomy and he pointed out that Meta has not open-sourced all its products; for instance, Meta chose not to release an application that allows users to edit videos and generate voices. Meta believes that an open-source approach is better for the security and stability of models. Mr. Stern explained that currently the social media playbook is generally being used for generative AI – events like this Symposium enable actors to discuss whether and how this playbook should continue to be applied.

Kim Malfacini – OpenAI Product and Policy Analyst – explained that OpenAI's approach to open development of AI tools has evolved. Initially, it supported an open-source model but over the years, and in view of the risks AI can create, OpenAI concluded that there are significant risks in open-sourcing. OpenAI has now opted for a gated API approach, which, according to Ms. Malfacini, makes it hard to break the safeguards that the company spends months building. She said that in open-source systems safeguards are significantly easier to undo. Like Mr. Stern, Ms. Malfacini pointed out that the discussion between open source and proprietary need not be binary – this distinction may become more relevant in the future with frontier AI models, but currently there may be benefits to having a middle ground, for example, to have AI products with adjustments for specific communities. Ms. Malfacini considers that we do not yet have a good collective model of how free speech and other values should apply to generative AI. There are significant differences between generative AI and social media – notably, AI systems generate content, they do not merely reproduce content from others nor distribute content to a wider audience beyond the specific user generating it.

Allison Stanger – Professor at Middlebury College – pointed out that the "open" versus "closed" distinction is useful in the sense that each of these options can have different impacts in different

countries. She considers that centralized models more readily work well in authoritarian systems; open-source models are usually associated with freedom. Prof. Stanger defended that platforms like Google or Facebook do not merely provide access to content generated by third parties – they also heavily moderate content, including with algorithms that aim to generate engagement and increase profit margins. She believes these platforms should not benefit from the liability exemption established in Section 230, given the amount of content curation they do. Prof. Stanger defended that we should consider the unintended consequences of Section 230 and hold companies liable for the content they host. While recognizing the usefulness of generative AI, Prof. Stanger also considers that the challenges faced during the United States 2020 elections may be exacerbated by the use of generative AI to produce disinformation.

Presentation of an AI-based Application to Counter Hate Speech

Jesse Spencer-Smith – Interim Director and Chief Data Scientist at the Vanderbilt University Data Science Institute – presented an AI-based application that will allow users to promptly counter hate speech using best practices drawn from the Toolkit for Using Counter Speech developed by FFS and the Dangerous Speech Project. This toolkit aims at empowering internet users, online activists, and civil society organizations. The AI-based application presented by Mr. Spencer-Smith was developed by the Vanderbilt Data Science Institute (DSI) and FFS and relies on information provided by the user and the toolkit to generate responses to hate speech that reflect the user's personally held beliefs and voice. More information on the toolkit and the application can be found at https://futurefreespeech.com/a-toolkit-on-using-counterspeech-to-tackle-online-hate-speech/.

Friday, Oct. 13 – The John Seigenthaler Freedom Forum First Amendment Center, Nashville, TN

Keynote Address: Max Tegmark, Professor of Physics at MIT and President of the Future of Life Institute

Max Tegmark discussed freedom of expression, AI, and disinformation. Prof. Tegmark first emphasized the progress made by AI in recent years and pointed out that we should be both excited and careful about this technology. He identified deepfakes as one of the most obvious risks stemming from AI and advocated for the adoption of "bot-or-not" and digital watermarking rules. Prof. Tegmark also talked about how difficult it is to deal with disinformation while not over-censoring, unduly limiting freedom of expression. He also referred to the challenges of distinguishing censorship, propaganda, and disinformation and emphasized the importance of humility when dealing with content moderation. Prof. Tegmark also warned about the risk of giving powerful entities – such as the government or companies – influence over fact-checking. Prof. Tegmark also referred to the importance of paying attention to conflicts of interest in content

moderation. He also warned about the "cancel culture" in campuses and the "invisible censorship," this is, when speech is restricted not because it is illegal but because of the social cost that it implies, like not obtaining a promotion for having said something controversial. Prof. Tegmark also talked about how news coverage greatly changes depending on where the media outlet stands politically. He studied this phenomenon in a paper considering over 100,000 articles from, approximately, 100 newspapers – the paper identifies, among others, the differences in the words used to refer to the same phenomenon (e.g., rally vs. riot, oil producers vs. Big Oil) depending on whether the outlet is right- or left-leaning or whether it is mainstream or critical of the establishment. In addition, Prof. Tegmark argued that merely blocking false information is not an effective way of addressing the disinformation phenomenon as it only addresses the symptoms and not the underlying causes of the situation we currently face. As a potential solution, Mr. Tegmark proposed applying the scientific method to freedom of expression and disinformation – the truth should be determined through inquiry in a democratic way not established by authority.

Panel No. 4: Trust and Safety in Generative AI

The fourth panel, moderated by John Samples – Vice President at the Cato Institute and member of the Oversight Board – discussed the crucial issue of trust and safety in generative AI, exploring measures to mitigate potential risks and build responsible AI systems.

Sam Gregory – Executive Director at WITNESS – expressed his concerns that generative AI may result in decreased trust in videos shared by regular citizens. Generative AI facilitates the creation of synthetic media and makes it hard to say, even for forensic experts, whether videos are authentic or not – this lack of trust in the authenticity of videos limits the ability of people to share their realities and denounce human rights violations. Mr. Gregory also pointed out that the field of generative AI should learn lessons from the previous wave of technology, social media. In social media, some communities felt excluded and that services were biased; they also considered that content moderation did not work well, was not well resourced and limited their freedom of expression. Mr. Gregory stressed the need to fix these issues and also to help people be better consumers of AI; he also emphasized the need to provide a certain degree of autonomy for users on the content they see.

Jules White – Associate Professor of Computer Science at Vanderbilt University – warned about the risk of focusing excessively on generative AI risks, such as the generation of disinformation, and not paying enough attention to the positive use cases and its potential, like using AI as a tutor for students or to generate code. Prof. White argued that excessively restricting what people can do with generative AI will limit the benefits all users can extract from this technology. He also pointed out that one of the most interesting uses of generative AI is challenging one's own perspective, our 'tunnel vision,' and getting new ideas – this exercise needs to be done purposefully and, to do this, users need to be educated on how to effectively use AI. In addition, it is important that regulation does not excessively limit the views and perspectives AI can share. Prof. White agreed with Mr. Gregory that users should have a certain degree of autonomy on what

they see and mentioned that generative AI is particularly well suited to provide diverse content in line with users' requests.

Abby Fanlo – Policy and Strategy Lead at the Chief Digital and Artificial Intelligence Office of the Department of Defense (DoD) – explained that the DoD Artificial Intelligence Office is focused on building safe and trustworthy AI systems. The Office enables personnel in the DoD to design, develop, and deploy AI in accordance with the Department's legal and ethical obligations. The ethical obligations are the same regardless of the technology, including AI. Ms. Fanlo emphasized the importance of educating all those who are going to work on AI governance as well as end users. To adequately govern AI, she also stressed the need to identify the use cases and conduct risk and performance assessments for each use case. Ms. Fanlo pointed out that it is important that end users trust the technology, and this requires that they believe that the technology is going to work as advertised. For this to happen, there needs to be explainable and measurable criteria for how AI systems perform.

Panel No. 5: The Challenge of Disinformation in Generative AI

The fifth panel, moderated by Charreau Bell – Senior Data Scientist at the Vanderbilt University Data Science Institute – examined the growing concern of disinformation generated by AI and explored strategies to combat this challenge, taking into account its impact on society and democratic processes.

Doug Fisher – Associate Professor of Computer Science and Associate Professor of Computer Engineering at Vanderbilt University – started discussing the role of AI as an intermediary between humans. The application of AI as an intermediary can be positive, like FFS' and DSI's AI-based application to counter hate speech or the use of AI for "productive talk", e.g., finding common ground between people with different views to send a letter to a Congressperson; but it can also be negative, like when AI is used to generate disinformation or hidden algorithms that guide indirect asynchronous communication in social media. Prof. Fisher argued that confirmation biases affect humans regardless of generative AI and pointed out that he is not sure that generative AI will have an additive effect on the impact of this bias. Regarding disinformation, Prof. Fisher expects an arms race between AI perpetrators – those generating and distributing disinformation – and AI guardians – those fighting disinformation. Prof. Fisher also pointed out that we should pay attention not only to disinformation concerning current events, but also to that affecting historical accounts.

Yi-Ling Chung – Research Associate at The Alan Turing Institute – begun by discussing positive generative-AI use cases, such as drug discovery or assisting users to prepare documents; she then referred to AI harms, like the generation of disinformation. Ms. Chung argued that to analyze AI's future impact it is crucial to think about how people may use generative AI. She also referred to research suggesting that while changing what people consider true or false is difficult, it is possible. For fact-checking to be effective, it is important that it is highly accurate; otherwise, it

can rapidly lose credibility. Ms. Chung also stressed the importance of ensuring that enough information on the abilities and limitations of AI is available, so users can know the products better and adjust their expectations accordingly.

Ari Cohn – Attorney and Free Speech Counsel at TechFreedom – defended that generative AI is a democratizing force for speech. Mr. Cohn argued that this technology puts tools that before were only accessible to wealthy people in the hands of everyone; it facilitates communication. Limiting generative AI may lead to cutting off development. Mr. Cohn considers that while generative AI facilitates the creation of bad content, this does necessarily imply an increase in the quantity of bad content in circulation. Mr. Cohn agreed with Prof. Fisher that it is not clear that generative AI will make the impact of confirmation bias worse; preliminary research suggests that the emotional manipulative effects of a deepfake video is not higher than that of other non-AI media. Mr. Cohn emphasized the importance of data literacy and general political knowledge as a way of reducing the impact of disinformation – users should be responsible for the information they consume and choose to trust.