Public Knowledge

August 1, 2023

Council Members,

Public Knowledge thanks the President's Council of Advisors on Science and Technology for inviting public input on this critical issue. Generative AI, and AI technology more broadly, presents both astounding opportunities and serious challenges. Ensuring that AI technology is developed and deployed safely, responsibly, equitably, and for the good of all must be one of the highest priorities of the United States.

1. ***In an era in which convincing images, audio, and text can be generated with ease on a massive scale, how can we ensure reliable access to verifiable, trustworthy information? How can we be certain that a particular piece of media is genuinely from the claimed source?***

The single most important way we can ensure access to verifiable, trustworthy information is to implement public policy solutions to address the crisis in local news. The focus of such efforts should not be to sustain or perpetuate legacy business models, many of which have not served communities well. Instead, policy approaches to the crisis in local news should be rooted in principles of the public interest; preservation of editorial independence; supporting innovators and new business models; enabling competition and choice; engaging served communities and emphasizing the needs of those communities; and considering reparative models that ensure continued movement toward community representation and social justice.

The Journalism Competition and Preservation Act (JCPA) is not the solution. The JCPA would exempt news organizations from antitrust laws, allowing them to collectively negotiate (and ultimately arbitrate) with dominant platforms for a share of the platforms' advertising revenue. This model is fatally flawed and will do more to harm our information environment than help it. Among other ills, it undermines decades of established copyright law; actively discourages content moderation by platforms; re-entrenches existing power structures in the media and technology sectors; encourages consolidation (and discourages competition) in news media; and will do little to put more reporters on the beat, especially in underserved communities. The White House should not support this proposal.

Conversely, the Community News and Small Business Support Act (H.R. 4756) empowers small businesses and news organizations themselves through two tax credits: one to small businesses that advertise in local news, and a payroll tax credit for newsrooms to hire or retain local reporters. All news outlets under a certain size qualify, provided they satisfy some consistent and foundational requirements; there are provisions to avoid news organizations supported by political organizations; and there is no "picking and choosing" by any government body. The tax credit model has proven so popular that several states have proposed one or both of the included credits through their own legislation. The White House should support this proposal.

News organizations are also among those wrestling with how to apply AI in responsible ways. Associated Press and OpenAI recently reached an agreement to train OpenAI's products on the AP's news archive. Meanwhile, Google's Genesis project seeks to use artificial intelligence to write news stories. Given their emphasis on source verification, fact-checking, and corrections, news organizations have significant incentive to ensure that citizens are seeing content that is verifiable and trustworthy, even if it is generated in part through AI. This is another reason to favor solutions that support local news as one way to ensure media is from the claimed source. Other solutions to ensure that citizens can verify the provenance of digital media are discussed below.

2. ***How can we best deal with the use of AI by malicious actors to manipulate the beliefs and understanding of citizens?***

Malicious actors present a particular challenge when it comes to disinformation—deliberately created false information. There are numerous, often countervailing, motives at play for malicious actors that engage in disinformation, and many are highly motivated to thwart countermeasures and push through systems designed to discourage simple misinformation. Disinformation can originate from sources ranging from sophisticated state-sponsored efforts at destabilization, to ideologically motivated propagandists, or even garden-variety online trolls with no purpose or motivation other than to create trouble. Accounting for this variety of motivations and interests makes it difficult to rely on the norms and incentives that compel most people and organizations to participate responsibly in our information ecosystem. To that end, it is important to consider how potential technical and policy solutions will operate when faced with a malicious actor who may behave without regard for personal benefit, economic rationality, or even common decency. Systems that rely on trust, on voluntary adoption, on norms, or simple bureaucratic or technical friction are likely to be easily exploited and subverted by malicious actors.

Therefore, it is critical to focus on robust information ecosystem scale solutions. There is no silver bullet for dealing with disinformation generated by malicious actors, but a combination of trusted and reliable fact-based journalism institutions, user-facing tools for transparency and information, AI and information literacy and education, robust content moderation standards, and real oversight and accountability for the digital platform and AI sectors, would all contribute towards developing a healthier information environment with users that are more resistant to the dangers of malicious disinformation.

3. ***What technologies, policies, and infrastructure can be developed to detect and counter AI-generated disinformation?***

A range of solutions will be required to best deal with the use of AI by malicious actors, including technical solutions, content moderation approaches by platforms, industry self-regulation, and highly targeted regulatory solutions.

Technical Solutions

The explosion of focus on generative AI has ignited a parallel explosion in technological solutions to track "digital provenance" and ensure "content authenticity" – that is, tools to help detect what content is created with AI. These tools, some of which come from the creators of AI systems, can be applied in different places on the value chain. For example, Adobe's Firefly generative technology, which will be integrated into Google's Bard chatbot, attaches "nutrition labels" to the content it produces, including the date an image was made and the digital tools used to create it. The Coalition for Content Provenance and Authenticity, a consortium of major technology, media, and consumer products companies, has launched an interoperable verification standard for certifying the source and history (aka provenance) of media content. Various systems for so-called "digital watermarking" – modifications of generated text or media in ways that are invisible to people but can be detected by AI using cryptographic techniques – have also been proposed. Several companies, including Meta for its new Llama 2 product, encourage the use of classifiers that detect and filter outputs based on the meaning conveyed by the words chosen. Digital forensics techniques (including IP address tracking, and reverse image searches for pre-existing content) are also helpful tools that can be used downstream from dissemination to detect inauthentic content.

While each of these solutions has its own strengths and weaknesses, even in aggregate they are imperfect, and fundamentally limited by the technological arms race they are in with new generative tools. Malicious actors will not utilize opt-in standards; in fact, they may be able to defeat some AI detectors simply by copying, resaving, shrinking or cropping images. Much like the content moderation systems of dominant platforms, many AI detection systems struggle with writing that is not in English. It is unlikely these tools would be able to keep pace with motivated propagandists, particularly those backed by state actors. And some of these methods also raise concerns that they may enable platforms to detect and moderate certain forms of content too aggressively, threatening free expression.

Content Moderation Solutions

Another category of solutions has to do with how downstream companies, such as search engines and social media platforms, are expected to moderate content created by generative AI. As generative AI comes into broader use, not all content will be malicious; some may be beneficial and entertaining, making blanket removal policies by search and social media platforms as undesirable as they are legally problematic. In an effort to enforce their community guidelines and terms of service (as well as avoid accusations of partisan bias), several major platforms have already shifted their content moderation emphasis away from actual content, and toward identifying account and behavioral signals, such as networks of accounts that amplify each other's messages, clusters of simultaneously-created accounts, and hashtag flooding. All of these may be helpful if lower cost, higher volume and more difficult detection are the hallmarks of generative AI in disinformation. Many other solutions are simply extensions of platforms' existing disinformation mitigation strategies. A holistic approach would encompass four major components: labeling (which may be done in conjunction with fact checking), algorithmic down-ranking, contextual prohibition, and platform-wide transparency reports.

Labeling is, arguably, the least invasive method of moderating AI-generated content. Labeling and identification can be accomplished at several points in the content chain. Users can be required to indicate whether the work they are uploading is AI-enabled. As noted below, methods for identifying the digital provenance of AI enabled images is a work-in-progress; however, some generative AI systems already include invisible watermarks to indicate AI involvement, and there is reason for optimism that such systems will expand with time. This may allow platforms to detect, at point of upload, whether the work is AI-enabled, and label it accordingly. Platforms may also harness existing fact-checking partnerships to verify and label inauthentic content. Finally, cross-platform systems (such as those which currently exist for fingerprinting and detecting non-consensual intimate images and child sexual abuse material) may be developed to identify known AI-generated content. At a minimum, platforms should require the labeling and identification of AI-generated content in paid advertising--including political ads. (It is worth noting that advertisers themselves may eventually require it, as they are also looking for ways to both exploit and guard against the capabilities of generative AI in order to protect their brands.)

However, it is important to bear in mind that not all AI-enabled content is harmful, and what is harmful will not always be easy to identify. We have already seen numerous examples of parody being misread as genuine, creating its own kind of misinformation. Platforms will need the flexibility to make judgment calls on how to treat AI-enabled content, once it has been identified. Risk assessments may determine that, in certain contexts (such as elections or public health information), the severity of potential disinformation harms may warrant stricter moderation policies. In those circumstances, they may or may not choose to respond to AI-enabled content as they would under their existing disinformation policies, or they may take a more aggressive response. Platforms must retain the flexibility to make these decisions in ways that support their stated terms of service and community standards.

Another crucial component is transparency at a platform level about the scope and scale of AI-generated content in a given forum. Platforms could add information about AI-generated content (such as its prevalence, or the type moderated) to existing transparency reports, so that lawmakers (and the public) have more complete information.

These approaches should all be considered, but their value depends on the various, and varying, policies of the platforms and their willingness and ability to enforce them, including in languages other than English.

AI Industry Self Regulation

Until or unless there are government regulations, the field of generative AI will be governed largely by the ethical frameworks, codes, and practices of its developers and users. (There are exceptions, such as when generative AI systems are deployed within the ambit of existing regulators.)

Virtually every major AI developer has begun to articulate their own principles for responsible AI development, develop accountability structures, and disseminate use policies that ostensibly govern the tools' use by others. Ideally, these encompass each stage of the product development process, from pretraining and training of data sets to setting boundaries for outputs, and incorporate principles like transparency, privacy and security, and equity and inclusion.

However, as in every other industry, voluntary standards and self-regulation are subject to daily trade-offs with growth and profit motives. This will be the case even when voluntary standards are agreed to collectively (as is a new industry-led body to develop safety standards) or secured by the White House (as was announced last week). For the most part, we're talking about the same companies – even some of the same people – whose voluntary standards have proven insufficient to safeguard our privacy, moderate content that threatens democracy, ensure equitable outcomes, and prohibit harassment and hate speech.

Regulatory Solutions

Any discussion of how to regulate disinformation in the United States – no matter how virulent, and no matter how it's created – is bounded by the simple fact that most of it is constitutionally-protected speech. Policy makers should exercise caution and restraint in considering solutions. We favor an approach of highly targeted and incremental regulation; that is, regulation that recognizes and accounts for a breadth of use cases and potential benefits as well as harms. It encourages us to focus on applications of the technology, not bans or restrictions on the technology itself.

Every sector and use case comes with its own set of ethical dilemmas, technical complexities, stakeholders and policy challenges, and potential transformational benefits from AI. For example, lawmakers should develop solutions that address the harms associated with disinformation whether they originate with generative AI, Photoshop, or foreign troll farms. The resulting policy solutions would encompass things like requirements for risk assessment frameworks and mitigation strategies (as called for in the European Union's Digital Services Act), transparency on algorithmic decision-making and its outcomes, access to data for qualified researchers, ensuring due process in content moderation, impact assessments that show how algorithmic systems perform against tests for bias, and enforcing accountability for the platform's business model (e.g., paid advertising).

We also need to account for the rapidity of innovation in this sector. One solution that Public Knowledge has favored is an expert and dedicated administrative agency for digital platforms—and perhaps now AI as well. Such an agency should have broad authority to enhance competition, protect consumers, and promote civic discourse

and democracy. An expert agency can set standards and even lead in AI algorithmic auditing and other oversight such as product safety validation. Data privacy protections are also relevant here, as they would limit the customization and targeting of content that can make disinformation narratives so potent and so polarizing. But it would be most beneficial to implement protections that cover *all* the data collection, exploitation, and surveillance uses rather than focus exclusively on AI-related use cases.

4. ***How can we ensure that the engagement of the public with elected representatives—a cornerstone of democracy—is not drowned out by AI-generated noise?***

First, we need to recognize how many ways generative AI can be used to *deter* authentic engagement of the public with elected representatives. These systems can:

- Mimic human interaction in lobbying campaigns;
- Overwhelm the volume of authentic content in public comments and create "snow blindness" in agencies;
- Distort elected representatives' perceptions of public opinion by flooding communication channels (such as by executing public relations campaigns and writing opinion commentaries, letters to the editor and community comments customized to different channels);
- Bolster the voices of the already-powerful, as tools take resources to develop and train and lobbying is most effective when backed with money as well as message; and
- Accelerate the development – or illusion – of expertise that would normally require years of experience. Bad actors using AI can be the first and produce the most on any issue that bursts into the scene; what is normally the product of highly specific industry or sector expertise can be generalized across sectors and legislators.

But the solutions we propose should not negate the positive ways that generative AI could increase the engagement of the public with elected officials. Greater awareness and access to these tools could:

- Lend eloquence and efficiency that empowers the authentic citizen-lobbyist to participate in public processes more easily;
- Make the public comment process more manageable and meaningful by allowing government agencies to collate and analyze responses;
- increase civility and persuasion in communication channels, through tools such as pro-social chatbots designed to promote positive social interactions;
- Quickly synthesize research in a given field to substantiate policy preferences or positions.

5. ***How can we help everyone, including our scientific, political, industrial, and educational leaders, develop the skills needed to identify AI-generated misinformation, impersonation, and manipulation?***

As the Surgeon General recently pointed out, it will take a whole-of-society approach to restore integrity and trust in our information environment, and we need to accelerate solutions that have already been proposed. Solutions may include equipping Americans with better tools to identify misinformation and disinformation and make informed choices about what information they share; expanding research into how disinformation is seeded and spread and how to counteract it; creating incentives for the technology platforms to change their policies and product design; fostering more competition and choice among media outlets; and convening stakeholders, including from the communities most impacted by misinformation, to research and design solutions – all while protecting privacy and freedom of expression. There should also be public education about AI, including through direct engagement with it. And as noted above, we should pursue policy interventions that ensure the availability of local civic information, since the crisis in local news has opened information voids that disinformation rushes in to fill.