CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

JANUARY 2025

Measuring Changes Caused by Generative Artificial Intelligence: Setting the Foundations

Samantha Lai, Ben Nimmo, Derek Ruths, and Alicia Wanless

*Alexandre Alaphilippe | Samantha Bradshaw | David A. Broniatowski | Graphika contributor | Josh Goldstein
Kristin O'Donoghue | Ronald Robertson | Elise Thomas | Gavin Wilde | Valerie Wirtschafter | Isabella Wright*

Measuring Changes Caused by Generative Artificial Intelligence: Setting the Foundations

Samantha Lai, Ben Nimmo, Derek Ruths, and Alicia Wanless

Alexandre Alaphilippe | Samantha Bradshaw | David A. Broniatowski | Graphika contributor | Josh Goldstein
Kristin O'Donoghue | Ronald Robertson | Elise Thomas | Gavin Wilde | Valerie Wirtschafter | Isabella Wright

Carnegie's Information Environment Project (IEP) is made possible with support from the Patrick J. McGovern Foundation and Google. Carnegie is wholly and solely responsible for the contents of its products, written or otherwise. We welcome conversations with new donors. All donations are subject to Carnegie's donor policy review. We do not allow donors prior approval of drafts, influence on selection of project participants, or any influence over the findings and recommendations of work they may support.

Carnegie does not take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Carnegie, its staff, or its trustees.

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Introduction

In 2024's so-called year of elections, fears abounded over how generative artificial intelligence (GenAI) would impact voting around the world.[1] However, as with other game-changing technologies throughout history, the sociopolitical risks of GenAI extend far beyond direct threats to democracy. As GenAI is leveraged to power "intelligent" products, made available for public use, adopted into routine business and personal activities, and used to refactor whole government and industry workflows, there are major opportunities for these disruptions to have negative consequences as well as positive ones.

These consequences will be hard to identify for two reasons. First, GenAI is being integrated into already complex processes. When the outputs of such processes change, it can be hard to trace changes back to their root causes. Second, most processes—whether in industry, government, or our personal lives—are not sufficiently well understood to allow detection of changes, especially those that are just emerging.

Informed policy that leads to beneficial change is extremely challenging to develop without being able to measure the material impacts of GenAI on governance, social services, criminal activities, health services, and myriad other aspects of social, political, and personal life. The act of measurement is necessary to help identify negative consequences that warrant prioritization and to understand whether claimed threats are over-hyped or under-recognized. Without measurement, we may fail to target policies directly towards issues that need the most attention. Worse, we may risk making changes that yield worse outcomes than the status quo.

This is the central problem we consider here. While democracies should be concerned about the potential impact of new technologies introduced into the information environment, how can those changes (good or bad) be measured? The Information Environment Project at the Carnegie Endowment for International Peace convened a workshop to explore this question in the context of common fears about the use of GenAI. The workshop ran in person with sixteen participants,[2] comprising investigators tracking GenAI abuses and researchers with experience measuring change, and was informed by a prior literature review to identify pre-existing knowledge gaps and points of debate.

Through the course of the workshop, we identified *four foundational questions* whose answers will underpin any serious scientific attempts to measure the changes wrought by GenAI in a given information ecosystem:

- What detection methods can reliably indicate that a piece of content is AI-generated?
- What is the baseline against which change in the ecosystem will be measured?
- Is the information ecosystem under observation complex and sprawling, with numerous variables and multiple sub-systems impacting each other, or more controlled, with fewer inputs and variables that a third party can more readily observe?
- Besides the new technology, what other factors influence the system, and how can they be accounted for?

Starting Point: Knowledge Gaps

A significant issue in tackling this topic is a lack of clarity around AI. The concept means many things to different people, leading to confusion on what exactly is being discussed.[3] Broadly speaking, AI is a collection of tools that allow computers to perform sophisticated functions. GenAI is a subset of these technologies, consisting of deep-learning systems that use training data to produce high-quality text, images, and other content.[4] Given the role of GenAI in generating content, known abuses of these tools tend to be tied to the use of preexisting technologies such as social media. Investigators participating in the workshop shared that GenAI had been used in influence operations to create fake accounts, avatars, and websites.[5] It has also reportedly been used to impersonate people, both to produce defamatory content[6] and to bolster the credibility of financial frauds,[7] romance scams,[8] or inaccurate voting information.[9] These use cases suggest that GenAI is often used with other communication tools. The fragmented nature of the information ecosystem, and the relative scarcity of reporting on AI-generated activity other than content distribution, mean that current knowledge of the role that AI plays, alone and in combination with other technologies, remains limited.

Fears Without Measures

Threats posed by the abuse of GenAI are frequently the focus of research papers, but impact measurements remain few and far between. We conducted a literature review of eighty-six journal articles and reports published between 2019 and 2024 that discussed the harms of GenAI on the information environment.[10] Of the eighty-six articles, thirty were published in 2024 alone. Many of them featured fears of how GenAI could be abused. Just over a third of the papers feared that democratic decisionmaking would be undermined by GenAI, while a little more than one-fifth worried, respectively, that such tools would make influence operations easier to scale, be used for persuasion, or decrease trust in institutions. Just nine of those eighty-six papers claimed to measure impact, consisting of either surveys or experiments whereby respondents were asked if they could discern between AI and human-generated content, felt they trusted in their ability to do so, or tested to assess if exposure to GenAI content affected their memory. All of these studies were conducted as one-off lab experiments on limited topic areas and mediums.[11] Two further publications provided suggestions for how impact could be measured on a societal level, such as by measuring changes in levels of trust over time, the prevalence of synthetic content, or by studying the aftermath of natural experiments such as policy shocks.[12]

As developments in GenAI are relatively recent, existing literature focuses primarily on projected risks rather than observed ones and has mainly been conducted in experimental settings. However, there are limitations to what we can learn from experiments, which assume that a single piece of content can be isolated as the main cause for people formulating a belief or decision. The reality is messier than that. Many variables go into human decisionmaking. Replicating the environment in which people form beliefs is extremely difficult, as it would need to consider preexisting notions, the various means for accessing information, and the scale and repetition of exposure of different types of information to put the content created by GenAI in context. To measure real-world consequences, there is a need to supplement existing lab-based studies with observational data collected from real-life settings where GenAI is used. In other words, it requires understanding the systems in which GenAI content is experienced. This makes some forms of impact measurement more practicable than others.

What Did We Do?

It’s clear that while feared threats related to the use of GenAI are receiving attention, measurements about their impact remain a gap in the literature. Thus, the workshop was designed to foster exploration of how an observable change thought to be caused by a specific use of GenAI could potentially be measured. To achieve this, the group was led through several interactive steps.

As a first step, participants were asked to come to a consensus on common feared outcomes related to the use of GenAI, following an initial briefing from investigators on abuses of the technology they had identified in their work. Using the KJ Method, participants were asked to articulate what precisely they feared would change as a result of GenAI, then directed to group these fears into categories and prioritize ones that were the most concerning.[13] This was then followed by a discussion aimed at identifying what events would have to happen in each scenario that would be clear indicators that the feared outcome had come to pass. These scenarios might not represent all potential feared outcomes of GenAI, but can be used as exemplars to guide exploration on paths to identifying impact and measurement. Seven categories of concern emerged related to the use of GenAI:

1. Healthcare (including the use of AI for diagnosis, self-diagnosis, and medical note-taking);

2. The legal system (including the use of AI to prepare cases, as well as the possible presence of AI-generated evidence);

3. Scams;

4. The creation of AI-generated child sexual abuse material (CSAM);

5. Information flows around natural disasters (including the rapid provision of accurate and *inaccurate* information);

6. Elections (including the provision of accurate or deceptive information); and

7. The generation of nonconsensual intimate media.

Many of these examples involved criminal activity, possibly reflecting the grave social concerns many participants investigate in their work. These issues were also reflected in the literature review.

As these scenarios have considerable overlap with problems that existed before the introduction of GenAI, participants then worked together as a group to brainstorm how GenAI-specific tools would change the situation. In so doing, participants were asked to identify an observable change that might be caused by GenAI, while also considering other possible causes and what would be required to measure GenAI's role in those changes. In pairs, participants then worked on a respective scenario to propose an approach to measuring an observable change in relation to that feared outcome.

While the topics discussed and approaches to measurement varied greatly, participants coalesced around four foundational questions. Answering these will be critical for designing experiments and approaches to measure the change caused by the use of generative AI in a given field.

Four Foundational Questions

What Detection Methods Can Be Used Reliably?

Almost across the board, the first common need identified was the ability to detect that GenAI was, in fact, used to generate an output. Several methods exist that hold potential, but further study is needed to assess their reliability at scale.

In six of the seven scenarios discussed, the ability to detect AI-generated content was essential to the research question of measuring the change brought on by the technology's use. This poses a challenge for measurements, as means for detecting AI-generated material are still emerging, and as such, the accuracy of existing detection methods remains contested.[14] Currently, there are three main detection methods. The first two detect GenAI content after the output is created, while the third form of detection is enabled during the content creation process.

Both perceptual and computational methods of detection aim to identify the use of GenAI after content has been created.[15] With the perceptual method, individuals rely on their perceptions to assess the authenticity of a piece of content. This can include detecting facial asymmetries or the alignment of eyes on AI-generated faces or noticing temporal inconsistencies between video frames and the accompanying audio.[16] This puts the onus on consumers and investigators to detect AI-generated content, which will become increasingly tricky as output quality continues to improve.[17] The computational method involves using technological methods to identify GenAI, through tools such as machine-learning classifiers or databases of inauthentic content. Both the perceptual and computational methods are of varying reliability and scalability.

The third common approach focuses on identifying the provenance of content, attempting to tie the output to its progenitor. A commonly proposed approach includes adding a label or watermark to verify that a piece of content is AI-generated. This incorporates a provenance-based approach into synthetic pipelines, adding digital watermarking into AI-generated content so that it can be traced back to a specific engine. Another approach involves developing provenance methods for showing that a piece of media is not AI-generated and is what it claims to be.[18] The provenance of non-AI-generated content is exemplified by the Content Authenticity Initiative, which has created an end-to-end technical specification to validate authentic content. The initiative plans to help third-party individuals and organizations adopt Content Credentials-enabled capture devices and applications, which can cryptographically hash recorded content and its optional metadata. This data will then be stored on a centralized trust list. When a user interacts with the media, a publisher that uses the protocol can validate content provenance and show that it is authentic.

Both methods have their limitations. Watermarks can easily be faked, removed, or ignored.[19] Provenance methods will be difficult to enforce with nonmajor GenAI services, such as open-source models and tools. Requiring provenance as a sign of credibility, however, may create an advantage for major providers, leading to a monoculture of models and further concentration of power. These challenges will continue to make measuring potential changes caused by GenAI difficult.

One exception might be the health scenario, which is within the more controlled system of a hospital or health provider. If such providers use GenAI for purposes like note-taking or diagnosis, this could be clearly recorded, allowing a high-confidence study. Conversely, though, if patients use GenAI to self-diagnose, there will still be a need to detect somehow that the technology was used. In this example, the situation is perhaps more challenging than in other scenarios in the sense that the diagnosis might not have been shared beyond a prompt response to the patient directly, and worse, that the patient is unable to confirm they had used GenAI to inform their health choices, along with any other resources they might have consulted. Measuring such health impacts would require the cooperation of GenAI companies and health service providers to connect usage with outcomes, raising serious privacy and ethical considerations for how such studies could be conducted.

What Is the Baseline Against Which Change Will Be Measured?

By nature, change is the act of something different occurring than what was before. For it to be measurable, something must be observably altered. This requires a baseline of the situation before, against which the altered state (or lack thereof) can be assessed. However, baselines may be more difficult to collect if that information had not been systematized in the past, or if measurements are needed on more intangible factors such as public sentiment. In some scenarios, more than one baseline may be necessary, further complicating what needs to be studied.

Some scenarios lend themselves to a degree of baseline comparison more readily than others. For example, given the nature of fraud, being both illegal and tied to financial loss, online scams were tracked and analyzed before the introduction of GenAI. It is possible to measure whether there has been growth in successful scams against currently reported levels of monetary loss. While the detection of all potential uses of GenAI, such as in phishing emails, might be challenging, it might be possible to isolate cases whereby media (images, audio, or video) were generated to appear to be someone known to the target and used. Presumably, targets would have been able to confirm after the fraud whether the media used was real if they had reported it to authorities. At a minimum, the increase in usage and the total costs of such fraud could be measured. Building on past assessments, it is possible to determine cascading costs related to the well-being of victims and costs associated with victims accessing public services they might have been able to afford if they hadn't been defrauded.[20]

For other scenarios, setting a baseline is more challenging. As mentioned, many studies (and indeed news coverage) point to the potential for GenAI to increase the scale of threats like influence operations or disinformation. Setting aside the relationship between the scale of content production and the ability to reach an audience, which is complex and nonlinear, measuring the extent to which GenAI enhances these threats would still require a baseline for what such operations produced before the introduction of GenAI. However, most work on known examples of disinformation and influence operations consists of unsystematized case studies and investigations. There is a critical requirement for meta studies and longitudinal analysis across examples to establish a baseline for the presence of such a phenomenon.

Another commonly expressed fear in the literature was that GenAI would lead to change in generalized attitudes, such as trust in knowledge or democracy. This fear, while strongly held, is a difficult outcome to measure at that level of abstraction. For something to be measurable, there needs to be some observable change specific enough that it could be measured, and a causal link between the input (creation of AI content) and output (lowered trust) that can reliably exclude or hold constant other possible influencing factors. Baseline measurements here would need to establish both the prevalence of an attitude ("I do not trust") and the current panoply of sources that underpin it, so that changes in sources and attitudes could both be identified over time. To understand the complex series of questions that establishing a baseline might involve, we focused on one subset of this larger concern that emerged from our literature review and which is of particular concern in academic articles: fears over an increase in fabricated academic articles leading to an erosion of knowledge.

Measuring the change in the reliability of academic articles would first require an assessment of the historical prevalence of (non-AI-generated) unreliable articles. For example, in the pre-AI era, how many were rejected for publication or were accepted and then retracted? Retractions could, in theory, be measured, as there are databases that track this activity;[21] setting a baseline for rejected papers would require some form of insight from the existing mitigants in the academic field: the peer reviewers and editors who rejected certain papers. If those baseline figures could be established, a longitudinal study could then assess whether the rate of rejections and retractions is rising or falling. Given adequate detection capabilities (as discussed above), this could, in turn, provide a measurement of the *prevalence* (or lack thereof) of AI-generated scientific articles. This approach would need to be paired with a longitudinal survey or poll of opinions about the reliability of scientific articles or trust in scientific knowledge. Therefore, this question would require not one baseline but two.

Is the System Under Consideration Complex or More Controlled?

In exploring possible approaches to measuring change, it is also essential to consider whether the information system under question is more controlled, with fewer inputs and variables, such as a court or hospital system, or more complex and sprawling, with numerous variables and multiple subsystems impacting each other, such as a society, democracy, or the information environment.

More controlled systems lend themselves more readily to measurement, especially under circumstances where the use of GenAI is routinely recorded. For example, a hospital using GenAI for recordkeeping or diagnosis may record that fact.[22] The closed nature of the medical system would then make it easier for researchers to evaluate whether overall health outcomes improve or deteriorate, as all relevant changes can be tracked in patients' electronic health records and hospitals' integrated AI models. Assuming a high level of compliance from hospitals, that can decrease the number of external influences or inaccessible sources of information that researchers will have to worry about. If researchers want to measure how GenAI affects the accuracy of doctors' diagnoses, they can work with hospitals and doctors to access patient data, conversations with patients, and doctors' LLM activity and use. Then, they can calculate how often doctors implement the output they receive from GenAI models, and see if instances of GenAI use correlate to an increase or decrease in patients' adverse outcomes.

A court of law is also a more controlled system that maintains records over time. This makes it possible to establish baselines for comparison. For example, if the question is whether evidence created or manipulated by GenAI might increase wrongful convictions, then a starting point would be to monitor and assess cases that used media as evidence after the introduction of GenAI tools. If the use of GenAI can reliably be identified in the future, then it becomes possible to measure changes in case outcomes before and after the introduction of faulty GenAI "evidence." In a similar vein, another approach might look at cases that used media as evidence before the introduction of GenAI to baseline acquittal rates using arguments that the evidence was fabricated versus similar arguments made in cases after the technology's introduction.[23] It might also be possible to assess changes in case outcomes where GenAI was used to complete paperwork or draft submissions to the court. The detection issue could be overcome by requiring courts to include a question asking those involved in a case to indicate if and how GenAI was used and then comparing case outcomes from periods before such tools existed.

Policing CSAM is also a relatively controlled system, given the illegality of storing such content. A concern identified in the literature review and by workshop participants was that GenAI would lead to an increase in synthetic CSAM, which would strain investigative resources working to track down and save abused children. The detection issue remains a problem, but assuming there are baselines on investigative operations, it would be possible to measure changes in the volume of content reported to the United States' National Center for Missing & Exploited Children and efficacy rates in finding and saving real children both before and after the introduction of GenAI.

Conversely, complex and sprawling systems feature a far wider range of potential variables, a patchwork of technologies and systems potentially impacting the greater whole, and a far more fragmented pattern of data ownership and access to study it. Rather than working with one healthcare provider or court in a closed system, an analysis of change in a complex system such as a society or democracy would require the ability to assess and weigh against each other many concurrent, perhaps contradictory sources. Research aimed at understanding the change due to GenAI in a complex system, therefore, presents significantly greater challenges that should be built into the design of any experiment or study.

How Complex Is the System, and How Can Its Complexity Be Accounted For?

A related but distinct challenge pertains to the complexity of the system to be studied. The complexity of a system is shaped by the number of interlocking factors that may influence outcomes within it.

The greatest measurement challenge identified throughout our workshop was that complex systems are exceptionally difficult to study. They comprise many interrelated processes and variables that all might affect particular outcomes. In poorly understood complex systems, like the information environment, conducting measurements to assess change or impact is even more challenging. Designing valid studies in these complex environments will most likely require the ability to account for a very large number of confounding factors over time.

This complexity applies in both technological and human aspects of measurement development: technological, in the interplay between GenAI and other technologies, such as distribution platforms, that makes it difficult to separate the impact of GenAI from that of other technologies; human, in the interplay between factors that make it unclear whether GenAI or other sociopolitical factors contribute to human decisionmaking.

Many of the abuse examples raised by investigators at the workshop involved means of distribution like social media, mass media, and influencers. It is not merely the use of GenAI that could cause a change, but its use in conjunction with several other technologies, many of which remain poorly understood in the context of changes within the information environment. For example, feared outcomes related to the degradation of reliable search returns—such as how fabricated or hallucinated results might make it difficult for people to evacuate during disasters or find reliable medical information—might require understanding several interconnected processes, working backward from how results are identified and curated by search engines. It isn't merely the presence of GenAI-created outputs, like fake academic papers or news articles, but how they interrelate to the wider system. This would require a greater degree of understanding as to how the search engines work. Are they using GenAI themselves? What training material is being used to develop them? What factors do search engines use to select and sort results? The answers to these questions can help shed light on whether GenAI or other technological influences are the cause of change. In the past, website traffic, hyperlinks, citations, and social media trends were factors for sorting online search returns. If that still holds, assessing the legitimacy of those activities introduces the need for detection across multiple processes.

Beyond the role of technology, many of the most prevalent fears about the use of GenAI are related to human decisionmaking, whether in crisis situations or political events.

One scenario discussed during the workshop was the impact of GenAI information on people's responses to disasters. This was considered a potentially tractable area for measurements because some outcomes in real-world incidents (such as the number of evacuations versus the number of people who refused evacuation) can be observed and measured. However, the critical gap is the ability to isolate the causation of any particular human decision (stay or go) and attribute it to one single piece of information. Due to the complexity of the system, each piece of content is filtered through many potential avenues for accessing information, some of which are poorly understood. In past disasters, people failed to follow evacuation orders because they didn't receive them,[24] didn't understand evacuation instructions, or couldn't afford to leave.[25] Logistics for evacuation are complex (for example, figuring out which evacuation route to use or shelter to go to).[26] If people have trouble navigating information online, this could delay evacuation plans. This is further complicated when communication services are lost.[27] The projected scale of the disaster can also shape evacuation decisions, the demographics of people affected (whether they are low-income, have pets, or have disabled family members),[28] and other geographical factors that might make evacuation difficult. These are variables that researchers have identified in studying past evacuations, which provide insight and some baseline to compare how the situation might change after the introduction of GenAI. These causes have primarily been identified through surveys conducted after natural disasters. All of these variables would need to be controlled to measure the role of GenAI in evacuations as it is added to this information ecosystem.

Attempts to understand the threats posed by nonconsensual intimate media targeting women are another example of how complex systems are difficult to evaluate. A rise in such material is a legitimate fear. In one 2019 study, of the 14,678 synthetic videos where a woman's face was superimposed on another's body found online, 96 percent were pornographic.[29] Barring significant changes, the situation will likely worsen with better tools for creating nonconsensual intimate media. It's also likely that the use of such material to target women in politics would further discourage women from participating in politics.[30] However, measuring GenAI as the specific cause will still be challenging. The most direct rate, surveying women over time who left politics to identify their reasons for doing so, risks retraumatizing those who departed because of nonconsensual intimate media. There might also be concerns about listing this as the reason when the person wants to avoid drawing more attention to the existence of such material. The sad reality is that women in politics are likely to face a wide variety of gendered harassment, making it difficult to pinpoint the exact instance that led someone to leave politics.

The most common fear across the literature reviewed in this project, as well as in the abuses mentioned by those articles and workshop attendees, was the erosion of democracy. The feared abuses in this category were more numerous than any other. They ranged from the deliberate to the unintentional. In the first category was AI content generated by a malicious actor with the intent to deceive or suppress voting.[31] Other fears stemmed from chatbots hallucinating incorrect responses about the political process, such as voting locations.

Breaking these fears down into discrete categories allows researchers to identify scenarios where potential sources of information may be more accessible to enable measurement of changes within an ecosystem. For example, in the United States, voter suppression is illegal. This means that official information on incidents of voter suppression, such as providing false information on polling locations, is available—albeit at the state level.[32] It would also be possible to survey voters to understand who had planned to vote, what information they consulted, and if this led them to the correct place to do so or not. If content created using GenAI can be identified as the reason for people not voting, the impact could be measured by the number of people who would have voted had they received reliable information. This would be a subset of the larger question on the impact of GenAI on democracy, but it would be a potentially less complex and more realistically measurable one.

Democracy itself is a vastly complex system where many systems come together, from the proliferation of different social media platforms with varying standards of enforcement to the traditional media and their diverse audiences, to the local and national candidates and parties, to the mechanics of voting themselves. Establishing a single baseline for the health of democracy is a correspondingly complex undertaking: for example, Freedom House assesses the state of freedom in a country according to twenty-five distinct factors, including freedom of expression, but also freedom of association, the independence of the judiciary, and the conduct of elections.[33] To understand how GenAI might be affecting democracy overall would require many different types of studies, assessing different factors, likely over time. As such, research into this question would require collaboration between many teams with different specializations. A project such as this could yield immensely valuable results, but the complexity of the undertaking should not be underestimated.

Conclusions

Measuring the impact of something in the information environment is challenging at the best of times. That doesn't mean it shouldn't be tried, but rather that attempts need to be realistic and explicit about what can be achieved. The more evident and direct the use of GenAI and more specific the observable change in relation to its use, the greater the chances of measuring that impact.

Conversely, where the use of GenAI is one among many processes within a complex system—especially if that system is, in turn, affecting another complex system—the more challenging those possible changes are to measure. Known use cases of GenAI in closed systems are easier to measure. Whereas complex systems introduce many variables, making it more difficult to pinpoint causation.

The studies with the greatest chance of credible and informative findings will be those which can effectively answer the four foundational questions that emerged from our workshop. How will the study reliably detect GenAI content? How will it define the baseline against which change is measured? Is the system to be studied sprawling or controlled? And how can the complexity of the system be assessed and mitigated?

Understanding the role that GenAI plays in information ecosystems will be crucial as the technology advances, and ever more AI providers enter the field. The foundational questions identified in our workshop are designed to inspire further research that seeks to address the question of identifying and measuring the impact of GenAI on society. This paper offers a path to identifying and measuring these changes that can be adopted by researchers and applied by policymakers in assessing the quality of attempted measurements.

About the Author

Samantha Lai is a senior research analyst at the Carnegie Endowment for International Peace's Information Environment Project. Her work has been published in the Journal of Online Trust and Safety, Lawfare, and TechTank; and featured by the OECD, NPR, and Politico, among others.

Ben Nimmo is a threat intelligence investigator at OpenAI, contributing in his personal capacity. He was a co-founder of the Atlantic Council's Digital Forensic Research Lab (DFRLab), and later served as Graphika's first head of investigations.

Derek Ruths is an associate professor of computer science at McGill University. He joined the faculty in 2009 after completing his PhD in computer science at Rice University. His ongoing work considers the problem of characterizing and predicting the large-scale dynamics of human behavior in online social platforms.

Alicia Wanless is the director of the Information Environment Project at the Carnegie Endowment for International Peace, which aims to foster evidence-based policymaking for the governance of the information environment.

Notes

[1] Pooja Chhabria, “The Big Election Year: How to Protect Democracy in the Era of AI,” World Economic Forum, January 29, 2024, https://www.weforum.org/agenda/2024/01/ai-democracy-election-year-2024-disinformation-misinformation/; Ali Swenson and Kelvin Chan, “Election Disinformation Takes a Big Leap With AI Being Used to Deceive Worldwide,” AP News, March 14, 2024, https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd.

[2] All participants were given the choice to be named or anonymous for this paper given security and institutional concerns. As a result, there is a slight discrepancy between the number of names listed as contributors and the number of participants in the workshop.

[3] Will Douglas Heaven, “What Is AI?,” *MIT Technology Review*, July 10, 2024, https://www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/

[4] Kim Martineau, “What Is Generative AI?,” IBM, April 20, 2023, https://research.ibm.com/blog/what-is-generative-AI.

[5] Elise Thomas, “’Hey, Fellow Humans!’: What Can a ChatGPT Campaign Targeting Pro-Ukraine Americans Tell Us About the Future of Generative AI and Disinformation?,” Institute for Strategic Dialogue, December 5, 2023, https://www.isdglobal.org/digital_dispatches/hey-fellow-humans-what-can-a-chatgpt-campaign-targeting-pro-ukraine-americans-tell-us-about-the-future-of-generative-ai-and-disinformation/; “Deepfake It Till You Make It: Pro-China Actors Promote AI-Generated Video Footage of Fictitious People in Online Influence Operation,” Graphika, February 7, 2023, https://public-assets.graphika.com/reports/graphika-report-deepfake-it-till-you-make-it.pdf; Jack Brewster, “How I Built an AI-Powered, Self-Running Propaganda Machine for $105,” *Wall Street Journal*, April 12, 2024, https://archive.md/2024.04.14-095216/https:/www.wsj.com/politics/how-i-built-an-ai-powered-self-running-propaganda-machine-for-105-e9888705.

[6] Natalie Wade, “State Department Remarks on Ukraine-US Are Digitally Faked,” AFP Fact Check, June 5, 2024, https://factcheck.afp.com/doc.afp.com.34V36KW; Paul Myers, et al., “A Bugatti Car, a First Lady and the Fake Stories Aimed at Americans,” BBC Verify and BBC News, July 2, 2024, https://www.bbc.com/news/articles/c72ver6172do.

[7] Heather Chen and Kathleen Magramo, “Finance Worker Pays Out $25 Million After Video Call With Deepfake ‘Chief Financial Officer,’” CNN, February 4, 2024, https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html; Renee DiResta and Josh A. Goldstein, “How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth,” Harvard Kennedy School Misinformation Review, August 15, 2024, https://misinforeview.hks.harvard.edu/article/how-spammers-and-scammers-leverage-ai-generated-images-on-facebook-for-audience-growth/.

[8] “Frank AI,” Frank on Fraud, accessed October 3, 2024, https://frankonfraud.com/wp-content/uploads/2024/07/FrankAI.m4v; Siqi Chen (@blader), “Where we are with ai today,” X, August 9, 2024, 9:09 PM, https://x.com/blader/status/1822077697805447405.

[9] Max Matza, “Fake Biden Robocall Tells Voters to Skip New Hampshire Primary Election,” BBC, January 22, 2024, https://www.bbc.com/news/world-us-canada-68064247.

[10] To retrieve relevant research articles, we conducted a systematic keyword-based search on search terms including “generative ai” + “disinformation”, “misinformation”, “information”, “harms”. Articles were excluded if they did not have a substantial focus on harms related to the information environment (taking for example articles that focused primarily on GenAI’s economic or labor impact.) The remaining papers were used to identify commonly-perceived harms and identify gaps in literature on measuring the impact of GenAI.

[11] Chiara Longoni, et al., “News From Generative Artificial Intelligence Is Believed Less,» in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 97-106, 2022, https://dl.acm.org/doi/pdf/10.1145/3531146.3533077; Cristian Vaccari and Andrew Chadwick, “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News,” *Social media+ society* 6, no. 1 (2020), https://journals.sagepub.com/doi/10.1177/2056305120903408; Sarah, R. Kreps, Miles McCain, and Miles Brundage, “All the News That’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation,” *Journal of experimental political science* 9, no. 1 (2022): 104–117, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3525002; Soubhik Barari, Christopher Lucas, and Kevin Munger, “Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media,” *OSF Preprints*, January 13, 2021, https://osf.io/preprints/osf/cdfh3; Giovanni Spitale, Nikola Biller-Andorno and Federico Germani, “AI Model GPT-3 (Dis) informs Us Better Than Humans,” *Science Advances* 9, no. 26 (2023), https://www.science.org/doi/10.1126/sciadv.adh1850; (Max) Hui Bai, et al., “Artificial Intelligence Can Persuade Humans on Political Issues,” *OSF Preprints*, February 4, 2023, https://osf.io/preprints/osf/stakv; Gillian Murphy, et al., «Face/Off: Changing the face of movies with deepfakes.» *Plos one* 18, no. 7 (2023), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0287503; Donghee Shin, Amy Koerber, and Joon Soo Lim, “Impact of Misinformation From Generative AI on User Information Processing: How People Understand Misinformation From Generative AI,” *New Media & Society* (2024), https://journals.sagepub.com/doi/abs/10.1177/14614448241234040?journalCode=nmsa; Matthew Groh, et al., «Human Detection of Political Speech Deepfakes Across Transcripts, Audio, and Video,» *Nature Communications* 15, no. 1 (2024): 7629, https://www.nature.com/articles/s41467-024-51998-z.

[12] Irene Solaiman, et al., “Evaluating the Social Impact of Generative AI Systems in Systems and Society,” arXiv, 2023, https://arxiv.org/pdf/2306.05949.pdf; Laura Weidinger, et al. “Sociotechnical Safety Evaluation of Generative AI Systems,” arXiv, 2023, https://arxiv.org/pdf/2310.11986.pdf.

[13] Jared M. Spool, “The KJ-Technique: A Group Process for Establishing Priorities,” Center Center, accessed October 3, 2024, https://articles.centercentre.com/kj_technique/.

[14] Several experiments have been conducted to evaluate the effectiveness of AI detection systems. Results vary: David Gewirtz, “I Tested 7 AI Content Detectors – They’re Getting Dramatically Better at Identifying Plagiarism,” ZDNet, August 8, 2024, https://www.zdnet.com/article/i-tested-7-ai-content-detectors-they-re-getting-dramatically-better-at-identifying-plagiarism/; Andrey A. Popkov and Tyson S. Barrett, “AI vs Academia: Experimental Study on AI Text Detectors’ Accuracy in Behavioral Health Academic Writing,” *Accountability in Research* (2024): 1–17, https://www.tandfonline.com/doi/abs/10.1080/08989621.2024.2331757; Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington, “Testing of Detection Tools for AI-Generated Text,” *International Journal for Educational Integrity* 19, no. 1 (2023): 26, https://link.springer.com/article/10.1007/s40979-023-00146-z.

[15] Hany Farid, “Creating, Using, Misusing, and Detecting Deep Fakes,” *Journal of Online Trust and Safety*, September 2022, https://www.tsjournal.org/index.php/jots/article/view/56/36.

[16] Josh A. Goldstein and Renee DiResta, “Research Note: This Salesperson Does Not Exist: How Tactics From Political Influence Operations on Social Media Are Deployed for Commercial Lead Generation,” Harvard Kennedy School Misinformation Review, September 15, 2022, https://misinforeview.hks.harvard.edu/article/research-note-this-salesperson-does-not-exist-how-tactics-from-political-influence-operations-on-social-media-are-deployed-for-commercial-lead-generation/.

[17] Stefan Feuerriegel, Renee DiResta, Josh A. Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz and Nicolas Prollochs, “Research Can Help to Tackle AI-Generated Disinformation,” *Nature Human Behavio*r 7 (2023), 1818–1821, https://epub.ub.uni-muenchen.de/121821/1/NHB__AI-generated_disinformation.pdf.

[18] “Coalition for Content Provenance and Authenticity,” Coalition for Content Provenance and Authenticity, accessed September 23, 2024, https://c2pa.org/.

[19] Kate Knibbs, “Researchers Tested AI Watermarks—And Broke All of Them,” *Wired*, October 3, 2023, https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

[20] Jilenne Gunther, “The Scope of Elder Financial Exploitation: What It Costs Victims,” AARP, June 27, 2023, https://www.aarp.org/pri/topics/work-finances-retirement/fraud-consumer-protection/scope-elder-financial-exploitation/; Natalie C. Ebner, Didem Pehlivanoglu and Alayna Shoenfelt. «Financial Fraud and Deception in Aging,» *Advances in Geriatric Medicine and Research* 5, no. 3 (2023), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10662792/.

[21] “Retraction Watch,” Retraction Watch, accessed December 3, 2024, https://retractionwatch.com/.

[22] Bernard Marr, “How Generative AI Will Change the Jobs o Doctors and Healthcare Professionals,” *Forbes*, March 13, 2024, https://www.forbes.com/sites/bernardmarr/2024/03/13/how-generative-ai-will-change-the-jobs-of-doctors-and-healthcare-professionals/.

[23] Herbert B. Dixon Jr., “The ‘Deepfake Defense’: An Evidentiary Conundrum,” American Bar Association, June 11, 2024, https://www.americanbar.org/groups/judicial/publications/judges_journal/2024/spring/deepfake-defense-evidentiary-conundrum/.

[24] Thomas Frank, “Conundrum: Why People Do Not Listen to Evacuation Orders,” *Scientific American*, September 27, 2019, https://www.scientificamerican.com/article/conundrum-why-people-do-not-listen-to-evacuation-orders/.

[25] Max Zahn, “Many People Evacuating Hurricane Ian Face Dire Financial Choices,” ABC News, September 29, 2022, https://abcnews.go.com/Business/people-evacuating-hurricane-ian-face-dire-financial-choices/story?id=90630756; Jonathan Franklin, “Some Don’t Evacuate, Despite Repeated Hurricane Warnings, Because They Can’t,” NPR, August 31, 2023, https://www.npr.org/2022/09/28/1125448849/why-people-dont-evacuate-hurricane-ian-florida; Edris Alam, “Reasons for Non-evacuation and Shelter-Seeking Behaviour of Local Population Following Cyclone Warnings Along the Bangladesh Coast,” *Progress in Disaster Science* 21 (2024), https://www.sciencedirect.com/science/article/pii/S2590061723000340.

[26] Alam, “Reasons for Non-evacuation.”

[27] Steve Almasy, “A Week After Helene Struck the Southeast, Power Outages and Impassable Roads Stymie Recovery as Death Toll Reaches 202,” CNN, October 3, 2024, https://www.cnn.com/2024/10/03/us/helene-recovery-roads-water-power/index.html.

[28] Alam, “Reasons for Non-evacuation,”; Jennifer Collins et al., “Hurricane Risk Perceptions and Evacuation Decision-Making in the Postvaccine Era of COVID-19 in US Coastal States Impacted by North Atlantic Hurricanes,” *Weather, Climate, and Society* 16, no. 1 (2023): 51–65; Stanley K. Smith and Chris McCarty, “Fleeing the Storm (s): An Examination of Evacuation Behavior During Florida’s 2004 Hurricane Season,” *Demography* 46 (2009): 127–145.

[29] Henry Adjer, Giorgio Patrini, Francesco Cavalli and Laurence Cullen, “The State of Deepfakes: Landscape, Threats, and Impact,” Deeptrace Labs, September 2019, https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[30] Kirsten Zeiter et al., “Tweets That Chill: Analyzing Online Violence Against Women in Politics: Report of Case Study Research in Indonesia, Colombia, and Kenya,” National Democratic Institute, June 14, 2019, https://www.ndi.org/sites/default/files/NDI%20Tweets%20That%20Chill%20Report.pdf.

[31] Shannon Bond, “A Political Consultant Faces Charges and Fines for Biden Deepfake Robocalls,” NPR, May 23, 2024, https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative.

[32] “Voter Fraud, Voter Suppression, and Other Election Crimes,” usa.gov, accessed October 15, 2024, https://www.usa.gov/voter-fraud.

[33] “Freedom in the World,” Freedom House, https://freedomhouse.org/report/freedom-world#Data/.

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 170 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Information Environment Project

The information environment is integral to democracy. This is the space where people process information to make sense of the world using tools from alphabets to artificial intelligence to produce outputs from the spoken word to virtual reality and whatever comes along in the future. Manipulation of the information environment threatens the legitimacy of democracy if citizens are increasingly unable to make free and informed decisions. Our understanding of this complex system is still emerging at the same time as conflicts within the information environment erode its integrity. In response, democracies around the world are increasing control over their national information ecosystems. But with little evidence to inform policymaking, they risk backsliding into authoritarianism or having their interventions backfire as trust in public institutions is degraded by information pollution. Carnegie's Information Environment Project is a multistakeholder effort to help policymakers understand the information environment, think through the impact of efforts to govern it, and identify promising interventions to foster democracy.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org