Meeting the moment: combating AI deepfakes in elections through today's new tech accord

Feb 16, 2024 | Brad Smith - Vice Chair & President

As the months of 2024 unfold, we are all part of an extraordinary year for the history of both democracy and technology. More countries and people will vote for their elected leaders than in any year in human history. At the same time, the development of AI is racing ever faster ahead, offering extraordinary benefits but also enabling bad actors to deceive voters by creating realistic "deepfakes" of candidates and other individuals. The contrast between the promise and peril of new technology has seldom been more striking.

This quickly has become a year that requires all of us who care about democracy to work together to meet the moment.

Today, the tech sector came together at the Munich Security Conference to take a vital step forward. Standing together, 20 companies [1] announced a new Tech Accord to Combat Deceptive Use of AI in 2024 Elections. Its goal is straightforward

Related Stories

Feb 13, 2024 | Brad Smith

Combating abusive AI-generated content: a comprehensive approach ›

Feb 5, 2024 | Courtney Gregoire

Increased uptake of generative AI technology brings excitement and highlights the importance of family conversations about online safety, says new research from Microsoft ›

Feb 5, 2024 | Noreen Gillespie

Here's how we're working with journalists to create the newsrooms of the future with AI ›

Related Blogs

but critical – to combat video, audio, and images that fake or alter the appearance, voice, or actions of political candidates, election officials, and other key stakeholders. It is not a partisan initiative or designed to discourage free expression. It aims instead to ensure that voters retain the right to choose who governs them, free of this new type of AI-based manipulation.

The challenges are formidable, and our expectations must be realistic. But the accord represents a rare and decisive step, unifying the tech sector with concrete voluntary commitments at a vital time to help protect the elections that will take place in more than 65 nations between the beginning of March and the end of the year.

While many more steps will be needed, today marks the launch of a genuinely global initiative to take immediate practical steps and generate more and broader momentum.

What's the problem we're trying to solve?

It's worth starting with the problem we need to solve. New generative AI tools make it possible to create realistic and convincing audio, video, and images that fake or alter the appearance, voice, or actions of people. They're often called "deepfakes." The costs of creation are low, and the results are stunning. The AI for Good Lab at Microsoft first demonstrated this for me last year when they took off-the-shelf products, spent less than $20 on computing time, and created realistic videos that not only put new words in my mouth, but had me using them in speeches in Spanish and Mandarin that matched the sound of my voice and the movement of my lips.

In reality, I struggle with French and sometimes stumble even in English. I can't speak more than a few words in any other language. But, to someone who doesn't know me, the videos appeared genuine.

AI is bringing a new and potentially more dangerous form of manipulation that we've been working to address for more than a decade, from fake websites to bots on social media. In recent months, the broader public quickly has witnessed this expanding problem and the risks this creates for our elections. In advance of the New Hampshire primary, voters received robocalls that used AI to fake the voice and words of President Biden. This followed the documented release of multiple deepfake videos beginning in December of UK Prime Minister Rishi Sunak. These are similar to deepfake videos the Microsoft

Jan 28, 2024 | Lillian Barnard

Governing AI in Africa: Policy frameworks for a new frontier ›

Jan 24, 2024 | Eric Horvitz

Broadening AI innovation: Microsoft's pledge to the National AI Research Resource pilot ›

Dec 19, 2023 | Julie Brill

Enhancing trust and protecting privacy in the AI era ›

More Cybersecurity Stories

Standing up for democratic values and protecting stability of cyberspace: Principles to limit the threats posed by cyber mercenaries ›

April 11, 2023

Threat Analysis Center (MTAC) has traced to nation-state actors, including a Russian state-sponsored effort to splice fake audio segments into excerpts of genuine news videos.

This all adds up to a growing risk of bad actors using AI and deepfakes to deceive the public in an election. And this goes to a cornerstone of every democratic society in the world – the ability of an accurately-informed public to choose the leaders who will govern them.

This deepfake challenge connects two parts of the tech sector. The first is companies that create AI models, applications, and services that can be used to create realistic video, audio, and image-based content. And the second is companies that run consumer services where individuals can distribute deepfakes to the public. Microsoft works in both spaces. We develop and host AI models and services on Azure in our datacenters, create synthetic voice technology, offer image creation tools in Copilot and Bing, and provide applications like Microsoft Designer, which is a graphic design app that enables people easily to create high-quality images. And we operate hosted consumer services including LinkedIn and our Gaming network, among others.

This has given us visibility to the full range of the evolution of the problem and the potential for new solutions. As we've seen the problem grow, the data scientists and engineers in our AI for Good Lab and the analysts in MTAC have directed more of their focus, including with the use of AI, on identifying deepfakes, tracking bad actors, and analyzing their tactics, techniques, and procedures. In some respects, we've seen practices we've long combated in other contexts through the work of our Digital Crimes Unit, including activities that reach into the dark web. While the deepfake challenge will be difficult to defeat, this has persuaded us that we have many tools that we can put to work quickly.

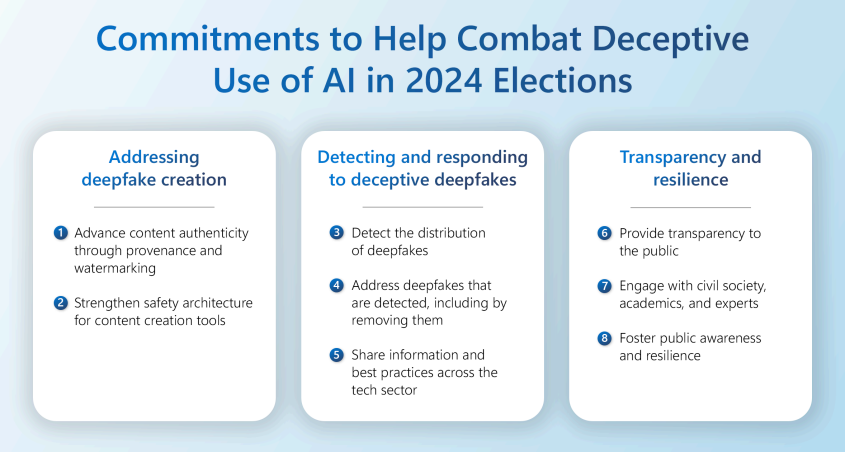
Like many other technology issues, our most basic challenge is not technical but altogether human. As the months of 2023 drew to a close, deepfakes had become a growing topic of conversation in capitals around the world. But while everyone seemed to agree that something needed to be done, too few people were doing enough, especially on a collaborative basis. And with elections looming, it felt like time was running out. That need for a new sense of urgency, as much as anything, sparked the collaborative work that has led to the accord launched today in Munich.

Digital Crimes Unit: Leading the fight against cybercrime ›

May 3, 2022

Keeping your vote safe and secure: A story from inside the 2020 election ›

June 22, 2021



What is the tech sector announcing today – and will it make a difference?

I believe this is an important day, culminating hard work by good people in many companies across the tech sector. The new accord brings together companies from both relevant parts of our industry – those that create AI services that can be used to create deepfakes and those that run hosted consumer services where deepfakes can spread. While the challenge is formidable, this is a vital step that will help better protect the elections that will take place this year.

It's helpful to walk through what this accord does, and how we'll move immediately to implement it as Microsoft.

The accord focuses explicitly on a concretely defined set of deepfake abuses. It addresses "Deceptive AI Election Content," which is defined as "convincing AI-generated audio, video, and images that deceptively fake or alter the appearance, voice, or actions of political candidates, election officials, and other key stakeholders in a democratic election, or that provide false information to voters about when, where, and how they can lawfully vote."

The accord addresses this content abuse through eight specific commitments, and they're all worth reading. To me, they fall into three critical buckets worth thinking more about:

First, the accord's commitments will make it more difficult for bad actors to use legitimate tools to create deepfakes. The first two commitments in the accord advance this goal. In part, this focuses on the work of companies that create content

generation tools and calls on them to strengthen the safety architecture in AI services by assessing risks and strengthening controls to help prevent abuse. This includes aspects such as ongoing red team analysis, preemptive classifiers, the blocking of abusive prompts, automated testing, and rapid bans of users who abuse the system. It all needs to be based on strong and broad-based data analysis. Think of this as safety by design.

This also focuses on the authenticity of content by advancing what the tech sector refers to as content provenance and watermarking. Video, audio, and image design products can incorporate content provenance features that attach metadata or embed signals in the content they produce with information about who created it, when it was created, and the product that was used, including the involvement of AI. This can help media organizations and even consumers better separate authentic from inauthentic content. And the good news is that the industry is moving quickly to rally around a common approach – the C2PA standard – to help advance this.

But provenance is not sufficient by itself, because bad actors can use other tools to strip this information from content. As a result, it is important to add other methods like embedding an invisible watermark alongside C2PA signed metadata and to explore ways to detect content even after these signals are removed or degraded, such as by fingerprinting an image with a unique hash that might allow people to match it with a provenance record in a secure database.

Today's accord helps move the tech sector farther and faster in committing to, innovating in, and adopting these technological approaches. It builds on the voluntary White House commitments first embraced by several companies in the United States this past July and the European Union's Digital Services Act's focus on the integrity of electoral processes. At Microsoft, we are working to accelerate our work in these areas across our products and services. And we are launching next month new Content Credentials as a Service to help support political candidates around the world, backed by a dedicated Microsoft team.

I'm encouraged by the fact that, in many ways, all these new technologies represent the latest chapter of work we've been pursuing at Microsoft for more than 25 years. When CD-ROMs and then DVDs became popular in the early 1990s, counterfeiters sought to deceive the public and defraud consumers by creating realistic-looking fake versions of popular Microsoft products.

We responded with an evolving array of increasingly sophisticated anti-counterfeiting features, including invisible physical watermarking, that are the forerunners of the digital protection we're advancing today. Our Digital Crimes Unit developed approaches that put it at the global forefront in using these features to protect against one generation of technology fakes. While it's always impossible to eradicate any form of crime completely, we can again call on these teams and this spirit of determination and collaboration to put today's advances to effective use.

Second, the accord brings the tech sector together to detect and respond to deepfakes in elections. This is an essential second category, because the harsh reality is that determined bad actors, perhaps especially well-resourced nation-states, will invest in their own innovations and tools to create deepfakes and use these to try to disrupt elections. As a result, we must assume that we'll need to invest in collective action to detect and respond to this activity.

The third and fourth commitments in today's accord will advance the industry's detection and response capabilities. At Microsoft, we are moving immediately in both areas. On the detection front, we are harnessing the data science and technical capabilities of our AI for Good Lab and MTAC team to better detect deepfakes on the internet. We will call on the expertise of our Digital Crimes Unit to invest in new threat intelligence work to pursue the early detection of AI-powered criminal activity.

We are also launching effective immediately a new web page – Microsoft-2024 Elections – where a political candidate can report to us a concern about a deepfake of themselves. In essence, this empowers political candidates around the world to aid with the global detection of deepfakes.

We are combining this work with the launch of an expanded Digital Safety Unit. This will extend the work of our existing digital safety team, which has long addressed abusive online content and conduct that impacts children or that promotes extremist violence, among other categories. This team has special ability in responding on a 24/7 basis to weaponized content from mass shootings that we act immediately to remove from our services.

We are deeply committed to the importance of free expression, but we do not believe this should protect deepfakes or other deceptive AI election content covered by today's accord. We therefore will act quickly to remove and ban this type of content from LinkedIn, our Gaming network, and other relevant Microsoft services consistent with our policies and practices. At the same time, we will promptly publish a policy that makes clear our standards and approach, and we will create an appeals process that will move quickly if a user believes their content was removed in error.

Equally important, as addressed in the accord's fifth commitment, we are dedicated to sharing with the rest of the tech sector and appropriate NGOs the information about the deepfakes we detect and the best practices and tools we help develop. We are committed to advancing stronger collective action, which has proven indispensable in protecting children and addressing extremist violence on the internet. We deeply respect and appreciate the work that other tech companies and NGOs have long advanced in these areas, including through the Global Internet Forum to Counter Terrorism, or GIFCT, and with governments and civil society under the Christchurch Call.

Third, the accord will help advance transparency and build societal resilience to deepfakes in elections. The final three commitments in the accord address the need for transparency and the broad resilience we must foster across the world's democracies.

As reflected in the accord's sixth commitment, we support the need for public transparency about our corporate and broader collective work. This commitment to transparency will be part of the approach our Digital Safety Unit takes as it addresses deepfakes of political candidates and the other categories covered by today's accord. This will also include the development of a new annual transparency report we will publish that covers our policies and data about how we are applying them.

The accord's seventh commitment obliges the tech sector to continue to engage with a diverse set of global civil society organizations, academics, and other subject matter experts. These groups and individuals play an indispensable role in the promotion and protection of the world's democracies. For more than two centuries, they have been fundamental to the advance of democratic rights and principles, including their critical work

to advance the abolition of slavery and the expansion of the right to vote in the United States.

We look forward, as a company, to continued engagement with these groups. When diverse groups come together, we do not always start with the same perspective, and there are days when the conversations can be challenging. But we appreciate from longstanding experience that one of the hallmarks of democracy is that people do not always agree with each other. Yet, when people truly listen to differing views, they almost always learn something new. And from this learning there comes a foundation for better ideas and greater progress. Perhaps more than ever, the issues that connect democracy and technology require a broad tent with room to listen to many different ideas.

This also provides a basis for the accord's final commitment, which is support for work to foster public awareness and resilience regarding deceptive AI election content. As we've learned first-hand in recent elections in places as distant from each other as Finland and Taiwan, a savvy and informed public may provide the best defense of all to the risk of deepfakes in elections. One of our broad content provenance goals is to equip people with the ability to look easily for C2PA indicators that will denote whether content is authentic. But this will require public awareness efforts to help people learn where and how to look for this.

We will act quickly to implement this final commitment, including by partnering with other tech companies and supporting civil society organizations to help equip the public with the information needed. Stay tuned for new steps and announcements in the coming weeks.

Does today's tech accord do everything that needs to be done?

This is the final question we should all ask as we consider the important step taken today. And, despite my enormous enthusiasm, I would be the first to say that this accord represents only one of the many vital steps we'll need to take to protect elections.

In part this is because the challenge is formidable. The initiative requires new steps from a wide array of companies. Bad actors likely will innovate themselves, and the underlying technology is continuing to change quickly. We need to be hugely

ambitious but also realistic. We'll need to continue to learn, innovate, and adapt. As a company and an industry, Microsoft and the tech sector will need to build upon today's step and continue to invest in getting better.

But even more importantly, there is no way the tech sector can protect elections by itself from this new type of electoral abuse. And, even if it could, it wouldn't be proper. After all, we're talking about the election of leaders in a democracy. And no one elected any tech executive or company to lead any country.

Once one reflects for even a moment on this most basic of propositions, it's abundantly clear that the protection of elections requires that we all work together.

In many ways, this begins with our elected leaders and the democratic institutions they lead. The ultimate protection of any democratic society is the rule of law itself. And, as we've noted elsewhere, it's critical that we implement existing laws and support the development of new laws to address this evolving problem. This means the world will need new initiatives by elected leaders to advance these measures.

Among other areas, this will be essential to address the use of AI deepfakes by well-resourced nation-states. As we've seen across the cybersecurity and cyber-influence landscapes, a small number of sophisticated governments are putting substantial resources and expertise into new types of attacks on individuals, organizations, and even countries. Arguably, on some days, cyberspace is the space where the rule of law is most under threat. And we'll need more collective inter-governmental leadership to address this.

As we look to the future, it seems to those of us who work at Microsoft that we'll also need new forms of multistakeholder action. We believe that initiatives like the Paris Call and Christchurch Call have had a positive impact on the world precisely because they have brought people together from governments, the tech sector, and civil society to work on an international basis. As we address not only deepfakes but almost every other technology issue in the world today, we find it hard to believe that any one part of society can solve a big problem by acting alone.

This is why it's so important that today's accord recognizes explicitly that "the protection of electoral integrity and public

trust is a shared responsibility and a common good that transcends partisan interests and national borders."

Perhaps more than anything, this needs to be our North Star.

Only by working together can we preserve timeless values and democratic principles in a time of enormous technological change.

[1] Adobe, Amazon, Anthropic, ARM, ElevenLabs, Google, IBM, Inflection AI, LinkedIn, McAfee, Meta, Microsoft, Nota, OpenAI, Snap, Stability AI, TikTok, TrendMicro, TruePic, and X.

Tags: AI for Good, artificial intelligence, cyber influence, cybersecurity, deepfakes, Defending Democracy Program, Microsoft Designer, MTAC, Responsible AI, Tech Accord to Combat Deceptive Use of AI in 2024 Elections