THE DIGITAL TRUST REVOLUTION:

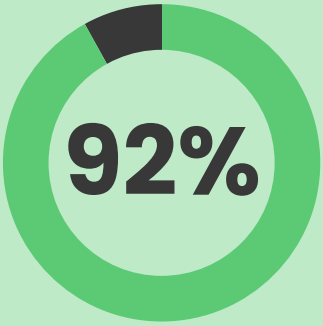
How AI is changing consumer behavior

New research highlights AI-generated content's impact on consumer habits.

AI-generated content is becoming increasingly prevalent online, raising all manner of ethical and societal concerns. From deepfakes to AI-written articles, the technology is advancing at a rapid pace, leaving governments, companies and consumers grappling with its implications.

But our new survey highlights the public's increasing awareness of AI's growing presence on their favorite digital platforms – and they have concerns.

AI-generated content isn't new, but its capabilities have expanded exponentially in recent years. From generating realistic images and video to writing coherent articles, AI is becoming a powerful tool for content creation. But it isn't stopping here, so how prepared is the world to take on its challenges? We spoke to Sam Gregory, award-winning journalist, Executive Director at WITNESS and an expert on generative AI to find out.



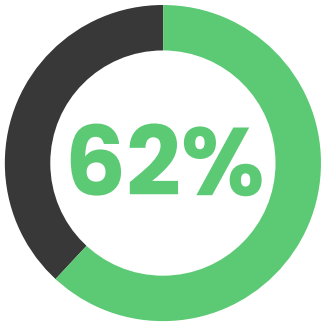
of respondents aged 35-44 agree* with the statement:

'The widespread presence of Artificial Intelligence-generated content impacts my trust of what I read and see online.'

Survey conducted by Censuswide on 7-11 July 2023, of 1000 US national representatives for content moderation agency WebPurify.

Censuswide is a member of ESOMAR – a global association and voice of the data, research and insights industry. It complies with the MRS code of conduct based on the ESOMAR principles.

The public has noticed *a large rise* in AI-generated content on its favorite platforms



of consumers have noticed an increase in the amount of AI-generated content on the platforms they use, including deepfakes and other fake content, with nearly 2 in 5 (**37%**) noticing a large increase.

More men than women are noticing the difference

74% of male respondents reported noticing an increase

51% of female respondents said the same

With younger generations either exposed, or noticing the increase, more

91% aged 35-44

76% aged 25-34

47% aged 55+

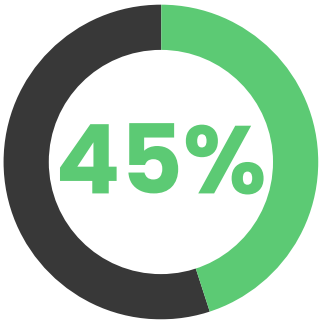
The ability to spot AI-generated content *will only get harder*

"Even I don't feel well-equipped in this area," says Sam. "I think a lot of the way we've been talking to people about AI-based content sort of assumes that they're going to spot these clues within it that will give it away. And I think the people who are overly confident are the ones who now know these clues - for instance, the distorted hands or a robotic-sounding voice. But I'm very aware that those clues are just the current Achilles heel of the algorithm."

Sam says that the figure from the survey is very resonant. In the consultations WITNESS runs globally, Sam says people often feed back that they not only feel poorly equipped to spot AI-generated content, but they also lack the tools to do so.

"I look at the trajectory [of AI's growth] and make a declaration that without signals of provenance, better detection and good media literacy, we're not going to be able to do this."

But there are many who *don't feel equipped* to discern between human and AI-generated content at all



of respondents do not feel well-equipped to discern between human-generated and AI-generated content, with 1 in 7 (**14%**) who do not feel well-equipped at all.

Women feel less equipped (or confident) than men

60% of female respondents do not feel well-equipped to spot the difference

29% of male respondents said the same

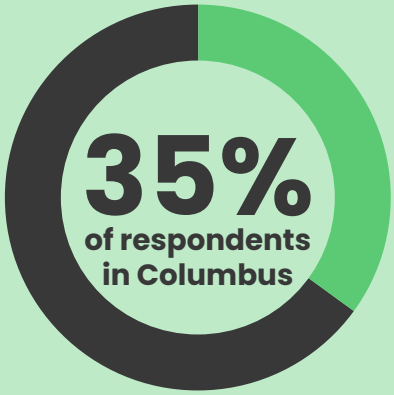
Unsurprisingly, age plays a significant role in people's confidence in spotting AI-generated content

75% of respondents aged 55+ do not feel well-equipped

4% of respondents aged 35-44 said the same

Geography *matters*

Over 4 in 5 (**83%**) respondents in New York feel well-equipped to discern between human-generated and AI-generated content, whereas over a third (**35%**) of those in Columbus said the same.



Many believe it's *a platform's responsibility* to detect and remove the AI-generated content

70%

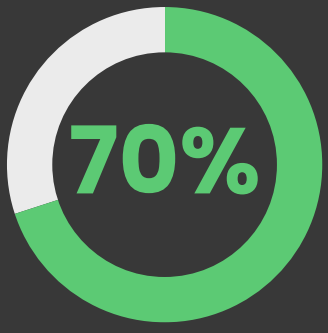
of respondents agree with the statement:

'It is a platform's (website or app) responsibility to detect and remove harmful AI-generated content, such as deepfakes'

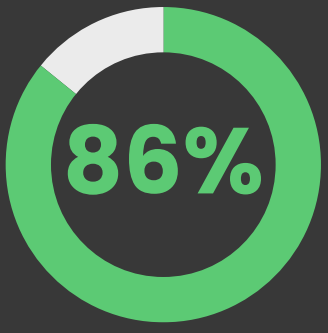
75%

of respondents believe more should be done to protect users from potential risks of AI-generated content.

Platform trust is significantly at risk, if they don't act now



of respondents say that the widespread presence of AI-generated content is impacting their trust in what they see and read online.



of respondents aged 35-44 said they would be less likely to return to a platform or website if they experienced AI-generated content, such as deepfakes

But platforms that invest in identifying and labeling AIGC will gain more trust

66% of respondents
would feel more comfortable using platforms that have measures to control or limit AI-generated content, or that require it to be clearly labeled.

With younger generations driving the call for platforms to take a leading role in labeling

94% of respondents aged 35-44
say this would make them feel more comfortable using platforms, followed by over 7 in 10 (**73%**) of those aged 25-34, and 2 in 5 (**44%**) of those aged 16-24.

Is it the platform's responsibility? What about the developers or consumers?

The pipeline of *responsibility*

Sam argues that the responsibility for managing AI-generated content extends beyond just the platforms or apps where the content is displayed. He introduces the concept of a "pipeline of responsibility," suggesting that accountability should also include the foundational models that generate the AI content and even the APIs that provide access to these models.

"We often talk about platforms' responsibility, but we might go further back to the foundation model," he explains. "Particularly if we are trying to understand that something was made with AI."

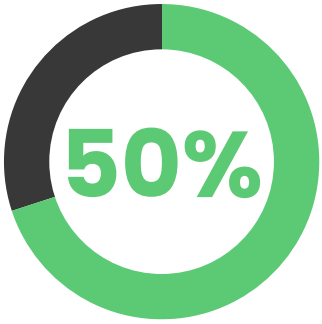
VP of Trust & Safety, Alexandra Popken, agrees: "Platform enforcement is inherently imperfect. Certainly, platforms have a responsibility to adapt their policies and enforcement measures and give users more privacy-respecting signals about what they're consuming, but there are also the AI developers and source models that have a role to play, and online users who need to be better educated on digital and media literacy."

"Certainly, platforms have a responsibility to adapt their policies and enforcement measures... but there are also the AI developers and source models that have a role to play."

ALEXANDRA POPKEN, VP OF TRUST & SAFETY

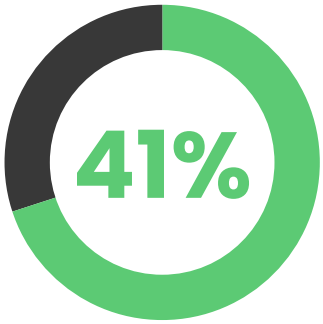
How *confident* **are people in platforms'** *ability to handle* **AI-generated content?"**

Some People are Confident



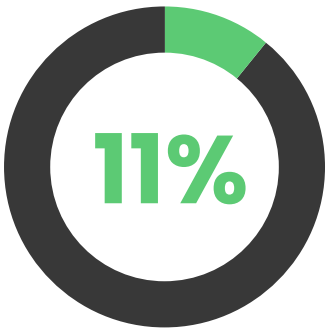
of respondents are confident in the ability of platforms to detect and handle AI-generated content, such as deepfakes, with a quarter (**25%**) who are very confident.

But Not Everyone Agrees



of respondents are not confident in the ability of platforms to detect and handle AI-generated content

A Minority Lack Confidence



Only 1 in 9 (**11%**) are not confident at all in the ability of platforms to detect and handle AI-generated content.

What are platforms doing *to moderate* the influx of AI-generated content?

TikTok
TikTok is enforcing the labeling of hyperreal AI-generated content in its community guidelines, and will take down videos without labels. 'You must disclose AI-generated content that shows realistic scenes' is stated in its community guidelines. 'Labels that help others know the difference between real and fictional content.'

Meta
Meta's own generative tools will mark assets generated using AI automatically. The subsequent labels state 'Image generated by Meta AI' and explain what generative AI is and how to know when posts use AI.

X (formerly known as Twitter)
X uses community notes to crowdsource context on potentially misleading posts, including AI-generated content.

As it relates to labeling or watermarking AI-generated content, Sam says there are many complexities to this. "The challenge is that we're not really clear what labeling or watermarking means," he says. "The devil is in the details. It's really important to recognize that binary visible labeling isn't going to be effective in the long run. The idea that we're going to say a piece of content is AI-generated or not and have a little watermark on a piece of media won't be effective because of the ways that we know people can very easily remove those watermarks."

AI is often just one part of a more complex media production process. Therefore, a simple "yes or no" label won't be effective and will also be inaccurate.

Instead, labeling should focus on how the content was made rather than who made it or why, to protect privacy and freedom of expression. This is particularly important in a global context where the identity of the content creator may be sensitive information.

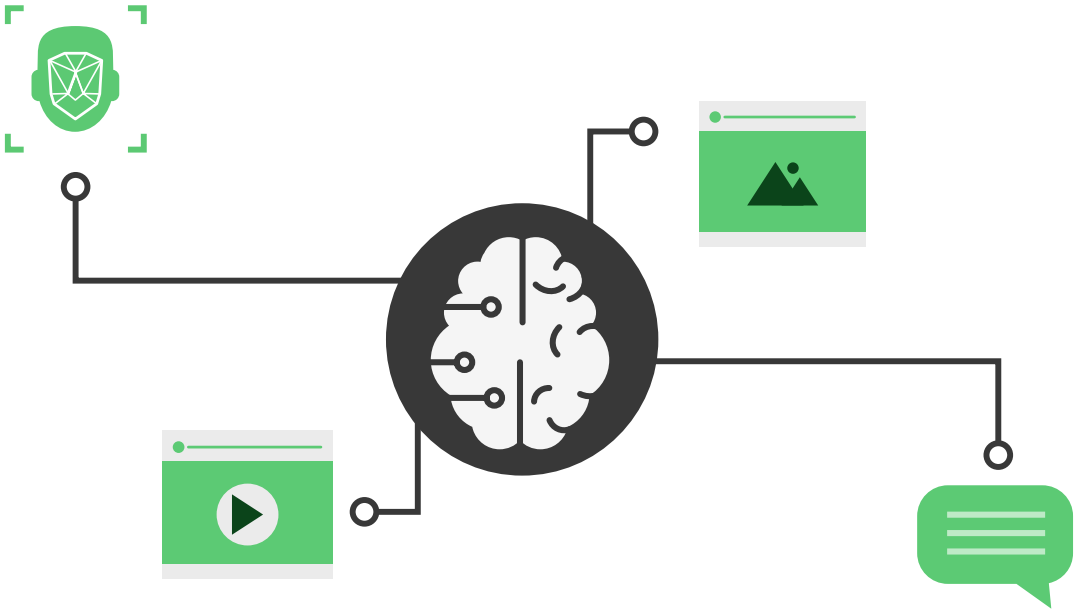
Sam suggests that while visible labels like watermarks have their place, much of the important disclosure information might be "invisible, existing as machine-readable data" that can be accessed through platforms or search engines. This nuanced approach would be more effective in helping people understand the nature of AI-generated content.

Recommendations **for platforms**

"The rise of AI-generated content presents a dynamic and novel challenge. At WebPurify, we are leveraging our sophisticated AI capabilities and expertly-trained human moderation team to review and remove emerging risks posed by newer technologies like it," says Alex, and suggests platforms consider the following steps to prepare for this wave:

1 Assess your risks and ethical obligations.

Sam suggests that the people who face disproportionate harm from generative AI are vulnerable and marginalized communities. A human rights impact assessment can provide a helpful and actionable framework for platforms to follow to ensure they're protecting these groups. Furthermore, "platforms integrating AI into their products or even using AI to enforce against AI should consider establishing ethical standards around consumer privacy and security, data model transparency and unbiased training process, regulatory compliance and appropriate use," Alex adds.



2 Update your policies to incorporate AI-generated content.

"Ensure community guidelines reflect appropriate and inappropriate uses of AI-generated content," suggests Alex. "For example, consider prohibiting things like synthetic and manipulated media that are used to deceive, confuse, or harm users." Additionally, if you're integrating generative AI into your platform or product, consider internal ethics guidelines to which you hold yourselves accountable.

3 Invest in automated and manual content moderation.

Alex suggests implementing robust content review and moderation systems to ensure that AI-generated content adheres to platform guidelines. "It's also important to partner with a content moderation provider that is up-to-date on this technology's risks and flexible in their moderation approach," she adds.

4 Educate and engage with your users.

Engage with your user base to gather feedback and address concerns related to AI-generated content. Educate your users about AI-generated content, its limitations, and potential risks. And where possible, equip users with tools and signals to help them discern between AI-generated and human-created content.

5 Partner with your peer set.

"The reality is that platforms are going to face an uptick in harmful AI-generated content – whether that's deepfakes circulating on social media platforms, sophisticated scams in online dating, or malicious phishing attacks powered by AI," Alex cautions. "To the extent that platforms can signal-share, it will make the industry writ large better prepared for the challenges that lie ahead." Alex also recommends that platforms consider an industry-wide consortium to establish best practices and standards around generative AI use and moderation.

• ABOUT WEBPURIFY •

WEBPURIFY

Your Customers Create. We Moderate.

WebPurify empowers communities to be their best with scalable hybrid AI and human content moderation solutions for the world's leading brands.

Ensuring positive user experiences for millions of customers, from marketplaces to the metaverse, with multimedia content seamlessly filtered to any brand's specifications.

Expert Consulting to Get You Started

At WebPurify, we're with you every step of the way—to define, optimize and proactively manage a solution that succeeds for your users and you.

Trusted by Hundreds of the World's Most Respected Organizations

Crayola Microsoft NBCUNIVERSAL Wieden+Kennedy accenture

Got a project in mind? Email: sales@webpurify.com