Online Dating & Discovery Association

World Romance Scam Prevention Day: Challenges of Generative AI

To mark World Romance Scam Prevention Day, the Online Dating and Discovery Association (ODDA) is using the month of October to raise awareness of the issue and highlight what our members are doing to remove scammers from their sites.

In this article, we look at the role of Artificial Intelligence (AI) which can be a powerful tool in keeping users safe but which is also being used by scammers.

The ODDAs Simon Newman (SN) talks to David Lissmyr (DL) from content moderation specialists SightEngine.

sightengine

SN: The use of Generative AI has been a game changer in many ways, but we've also seen scammers use it to create fake profile pictures for example. What are some of the problems Generative AI poses to online platforms?

DL: Generative AI significantly amplifies challenges that were already existing on online platforms by enabling the rapid production of custom content at scale. This includes the creation of illegal content such as Child Sexual Abuse Material (CSAM), fraud, misinformation, impersonation and violation of copyright laws.

Scammers can use GenAI to rapidly create convincing online personas, and thereby increase risks to digital safety and integrity.

SN: As with any emerging technology, there are often questions about whether existing laws are fit for purpose. Where does the law lie with images and content created by Generative AI?

DL: This is a new and evolving field. We are still waiting to see how regulation will adapt to the challenges posed by GenAI.

Existing laws that deem certain real images illegal also apply to their AI-generated counterparts, thereby maintaining legal continuity. In that sense, things haven't changed much. However, new requirements will emerge around media labelling (when and how should AI-generated images be labelled as such?), around source verification (when and how should platforms verify the source or authenticity of media?), and around liability.

SN: How do other online platforms deal with AI generated content?

DL: Online platforms currently do not systematically flag AI content, but attempt to empower users by making it easier for them to report AI-generated media they suspect to be fake.

Companies such as Meta, Google or TikTok have announced options for users to self-report AI content they share and have announced labels they will start adding to media that is AI-generated.

SN: How do you spot AI generated content. What can you tell us about Watermarking and Automated Detection?

DL: People resort to one of three methods to spot AI-generated media: watermarking, human evaluation or automated detection.

Watermarking involves embedding data in an image to signify it is AI-generated, like the emerging C2PA standard. However, it remains easy to generate images that do not contain watermarks, for instance by using models that do not include watermarking, or by stripping the watermark. The C2PA watermark being a meta-data, it is automatically removed when an image is uploaded to most online platforms.

Human evaluation, which assesses the authenticity of an image visually, is becoming less reliable as AI technology improves. We have tested thousands of people on this with our 'AI or not' test, and have seen that humans achieve an

accuracy of ~70% on average which, while better than a random guess (50% accuracy) is not sufficiently reliable.

Automated detection is the third option, and this is what we are investing on at Sightengine. We learn to recognise the subtle patterns left by GenAI models in the images and videos produced. This technology is rapidly evolving and represents a critical area of development for maintaining integrity in digital media.

SN: How well does detection of AI generated content work?

DL: A team of researchers from the University of Rochester and the University of Kansas did a very large study on this. They created a massive dataset of AI images across a range of topics (arts, photography, illustrations...) along with corresponding real images, and tested various solutions. They found out Sightengine had the highest accuracy at 98.3%. The next closest competitor achieved 86%. Despite this high mark, we will continue investing in research and development in this area and strive to continue improving our solutions.

SN: As we're looking at romance fraud this month, can the use of AI help us identify potential scammers?

DL: Yes, AI can aid in many different ways. Firstly, AI can detect images and videos used by scammers. Until recently, it was fairly common practice that scammers would re-use the same images over and over again, making them easier to spot through a technique called 'near-duplicate detection algorithms', even if the image or video has been modified or edited. Now scammers are using GenAI to create new profile images on demand. This is where AI detection becomes crucial.

Secondly, AI can help in detecting patterns indicative of scammers. This includes patterns in their exchanges with other users, such as how personal information is shared, or how scammers will attempt to move the conversation to other platforms. By identifying these behavioural patterns, AI helps in pre-emptively flagging potential fraud, safeguarding users from scams.

SN: As a company specialising in content moderation, what tips would you to suggest to online platforms working in the dating and social discovery space?

DL: While AI can help quickly detect AI-generated media and detect behaviours indicative of scamming, we believe platforms should continue investing in other areas, namely user verification, reporting mechanisms, user education and having reactive support. We also believe collaboration between platforms can be very valuable here.

SN: Finally, tell us a bit more about Sightengine.

DL: Sightengine has been a Content Moderation company for 11 years now. We provide online platforms, social media, dating apps with tools to automatically filter and analyse user generated content, be it images, videos or texts. Our objective is to make it simple and fast for platforms to leverage AI for Trust & Safety.

SN: Thank you for your time David.

You can find more about Sightengine on their website:
www.sightengine.com

Can you tell the difference between AI-generated images and real images? Test your skills and compare yourself to others here:
https://sightengine.com/ai-or-not