Public Service Announcement

**Product ID Number: NIAC-2025-2946509-86**
**September 9, 2025**

# Potential Dangers Of Artificial Intelligence's Pervasiveness

Artificial intelligence (AI) offers incredible potential, but its rapid growth and its current saturation in every facet of our society comes with significant risks to navigate. This bulletin outlines key dangers and steps to take to protect yourself and others from this rapidly advancing form of technology.

What is commonly thought of as "AI" is basically a computer program that is "trained" or "taught" to respond to user prompts. Based on the training, it "learns" how to decide what responses are the most correct and most acceptable for the situation or question it is presented. Furthermore, any prompts or information shared with an AI model becomes a part of the model's training to facilitate better responses in the future.

## AI Powered Cyber Attacks

- Cyber threat actors are using AI to craft more convincing phishing schemes and automate their network attacks.
- AI can coordinate Bot attacks in a Distributed Denial of Service (DDoS) attack.
- Multiple specialized AI agents can work together collaboratively to complete multi-prong attacks.
- AI can assess a target's network, determine the attack with the best chance of success, report back to the attacker, and assist in the implementation of the attack.
- AI can adapt attacks in real-time to defensive measures that defenders enact, increasing the likelihood of the attacker's success.

## Misinformation and Deepfakes

- AI can be used to create realistic images of people, places and events that are not real with very little effort or special knowledge needed. This can lead to false information being widely spread, financial loss, public panic, cyber bullying, and social and political instability.
- Misinformation and deepfakes can appear to be from anyone; political figures, celebrities, or everyday people. This can be in the form of videos, images, voice calls and texts.
- Misinformation is any information that can be verified as false that is widely spread (i.e. statistics from a previous year being reported as current statistics).
- Deepfakes are highly realistic, synthetic audio or visual media. This is similar to a person using a variety of video or audio clips to create a product that makes it seem like a person did or said something they never did.

## Jailbroken AI Models & Dark AI Models

- "Jailbroken" AI models have been manipulated to bypass safety measures ("Guardrails") that developers have coded into it. This allows the AI model to provide malicious or restricted information to the user, such as tricking the AI to provide answers that it has been programed to not supply due to safety concerns (i.e. how to build a bomb).
- "Dark" AI models are built for malicious purposes and have no ethical or safety restrictions built in. These allow the user to ask anything and receive answers without ethical restraints.
- Researchers have proven that even non-"Jailbroken" or "Dark" AI models can be "tricked" into providing information about optimal times for attack, physical location vulnerabilities, 'recipes' for chemical and biological explosives, as well as where and how to buy guns on the dark market under the guise of "security planning" even with the safety restrictions still in place.

This product is intended for wide distribution. It does not have restrictive intelligence handling requirements.

# Public Service Announcement

**Product ID Number: NIAC-2025-2946509-86**
**September 9, 2025**

## Issues With Machine Learning

- Algorithmic Bias - If you base the AI training on false, or negative information, the AI model can only respond with false or negative information. If you base the training only on positive and true information, the AI could distort the responses in that way. AI developers may not even understand the bias they are introducing in AI simply from the choice of training.
- Context Poisoning – Malicious actors could intentionally insert false or misleading data into the training sources. (i.e. if you teach a person that the sky is green, that is what they learn and will report to others, even when there is a differing opinion)

Keep in mind that how AI models are trained, the sources used are important, but not always available, to users.

## Social And Mental Health Impacts

- Overreliance on AI and increased virtual only interactions can decrease face to face interactions and lead to isolation, a decrease in meaningful social connections, and loneliness.
- Social interactions with AI may blur the lines between real emotions and virtual feelings that are expressed.
- Individuals may become dependent on AI to supply support and decision making. Long term use of AI can cause a decrease in critical thinking skills and self-efficacy.

## Mitigating The Negative Effects Of AI

- Encourage a balance between technological interaction and human connection to prevent social isolation and emotional disconnect.
- Ensure the information you are receiving from sources align with other, reputable information sources. Verify information on a variety of sites.
- Know how the AI model you use is "trained." The size and breadth of the training sample can indicate either a well-rounded AI model or biased AI model.
- Review safety and security measures that are in place on an AI model before using it widely.
- Keep in mind that AI can be very realistic and seem true. Even if the same audio or video clip is available from several sources it could still be false. Look for multiple sources confirming the information using a variety of different audio or visual clips. Critical thinking must be used when making important decisions. Verify information received with several other reputable sources.
- Be careful what personal information you share with AI models as this may become available to others. Understand, and use, privacy settings available in the AI model.

Resources:

B. C. Stahl et al., "A systematic review of artificial intelligence impact assessments," Artif. Intell. Rev, vol. 56, no. 11

M. L. Littman et al., "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report," Stanford Univ., Stanford, CA, Sept. 2021.

A. Zilber, "ChatGPT gave instructions on how to bomb arenas, make anthrax and illegal drugs, alarmed researchers reveal", *New York Post*, Aug. 29, 2025. [Online] Available: https://nypost.com/2025/08/29/business/chatgpt-gave-instructions-on-how-to-bomb-arenas-make-anthrax-and-illegal-drugs-alarmed-researchers-reveal/ [Accessed; Sept. 2, 2025

B. Marr, "The Dark Side Of AI: How Deepfakes And Disinformation Are Becoming A Billion-Dollar Business Risk". *Forbes*, Nov. 6, 2024 [Online] available https://www.forbes.com/sites/bernardmarr/2024/11/06/the-dark-side-of-ai-how-deepfakes-and-disinformation-are-becoming-a-billion-dollar-business-risk/, [Accessed Sept. 1, 2025]

This product is intended for wide distribution. It does not have restrictive intelligence handling requirements.