Week 11 Workshop

FIT5230 Malicious AI

AI+Security Warfare, AI Governance

MONASH University

Future
AI+Security

The Future of Security

- AI:
    - with great power comes great responsibility
    - AI is everywhere
    - Midjourney, ChatGPT, Gemini

- AI is not new, the advancement of hardware and software helps in increasing its applications.

- Application of AI
    - Healthcare - to improve patient outcome and advance research
    - Transportation

How does AI help in healthcare?

- Efficiency in Data Analysis: AI can process and analyze massive amounts of data much faster than humans. This capability is particularly useful in fields that generate large datasets, such as genetics and medicine.

- Genetic Research: AI helps geneticists by quickly sifting through vast genetic data libraries. This allows researchers to identify genes that may be linked to diseases, which can lead to the development of new diagnostic tests.

- Medical Treatments: AI accelerates the process of finding medical treatments. By analyzing data, AI can help identify potential treatments for diseases more quickly than traditional methods.

- Example - Every Cure: The non-profit organization Every Cure uses AI to search medical databases. Their AI algorithms match existing medications with diseases they might treat, saving time and resources in the process.

What to consider when using AI in healthcare?

AI systems are trained to look for patterns in large amounts of data.

- Make recommendation based on these patterns
- Suggest diagnoses, or initiate actions.
- They can potentially continually learn, becoming better at tasks over time.

What to consider when using AI in healthcare?

To make it work, these need to be addressed:

- AI should be at least as good as a human doctor at the tasks it performs. AI should not be used if it will lead to more incorrect diagnoses or medical errors.
- If AI systems generate decisions – such as diagnoses or treatment plans – without human input, it may be unclear who is responsible for errors. So people often want clinicians to remain responsible for the final decisions, and for protecting patients from harms.
- If health services are already discriminatory, AI systems can learn these patterns from data and repeat or worsen the discrimination. So AI used in health care can make health inequities worse. This is not acceptable.

How does AI help in Transportation?

Real-Time Traffic Management:

- Data Analysis: AI systems analyze real-time data on traffic patterns, volume, and other factors such as weather conditions and accidents.
- Adjusting Traffic Signals: Based on this analysis, AI can adjust traffic lights and signals to optimize traffic flow, reduce congestion, and minimize delays.

How does AI help in Transportation?

Navigation Apps:

- Optimisation Algorithms: Apps like Google Maps use AI algorithms to find the best routes for users. These algorithms consider various factors such as current traffic conditions, road closures, and historical traffic data.
- Route Suggestions: By continuously analyzing data, these apps can provide real-time updates and suggest alternative routes to avoid traffic jams and reach destinations faster.

Security Threats on AI

Confidentiality (CONF):

- Threat: AI systems can inadvertently expose sensitive information.
- If an AI model is trained on personal data, it might reveal private details during its operations.
- Example: A chatbot trained on customer service interactions might accidentally disclose personal information from previous conversations.

Integrity (INT):

- Threat: AI systems can be manipulated to produce incorrect or harmful outputs. This can happen through data poisoning, where attackers corrupt the training data.
- Example: An AI system used for medical diagnoses could be fed false data, leading to incorrect diagnoses and treatments.

Security Threats on AI

Authentication (AUTH):

- Threat: AI can be used to bypass authentication mechanisms. For instance, deepfake technology can create realistic fake identities that trick biometric systems.
- Example: A deep fake video could be used to impersonate someone and gain unauthorized access to secure systems.

Non-Repudiation:

- Threat: AI-generated content can make it difficult to prove the origin of actions or communications. This can lead to issues where individuals deny their involvement in certain activities.
- Example: An AI-generated email could be used to commit fraud, and the sender could deny having sent it, complicating the investigation.

Malicious AI Application

AI and Bioweapons:

Gene Sequencing:

1. AI can assist in gene sequencing, which is the process of determining the order of nucleotides in DNA.
2. This technology can be used to create or modify organisms.

Risky Pathogens:

1. With AI's help, even non-experts could potentially produce dangerous pathogens, such as novel viruses.
2. This is because AI can simplify complex biological processes and make them more accessible.

Malicious AI Application

AI in Warfare:

Scenario Planning:

1. Military powers can use AI to design and simulate warfare scenarios.
2. AI can analyze vast amounts of data to create detailed and strategic plans.
3. Missile Guidance Systems: AI is used to enhance the accuracy and effectiveness of missile guidance systems.
4. Submarine Detection: AI can analyze sonar and other sensor data to detect submarines that are operating covertly. This helps in maintaining maritime security and tracking potential threats.

Ethical Considerations:

If these AI tools are used without ethical oversight, the consequences can be severe.

Malicious AI Application

AI in Warfare:

Scenario Planning:

1. Military powers can use AI to design and simulate warfare scenarios.
2. AI can analyze vast amounts of data to create detailed and strategic plans.
3. Missile Guidance Systems: AI is used to enhance the accuracy and effectiveness of missile guidance systems.
4. Submarine Detection: AI can analyze sonar and other sensor data to detect submarines that are operating covertly. This helps in maintaining maritime security and tracking potential threats.

Ethical Considerations:

If these AI tools are used without ethical oversight, the consequences can be severe.

AI vs Human

- Will AI surpasses human, or even replace human one day?

- data→pattern, meaning, knowledge
  - humans: natural intelligence (NI)
    - lack capability to find complex patterns

  vs
  - machines: artificial intelligence (AI)
    - process faster/simultaneously, repetitive,remember more

- Ultimate Goal of AI:
  - human-like intelligence +
  - super-human intelligence capabilities
  - memory ↑
  - speed, parallelism ↑

AI vs Human

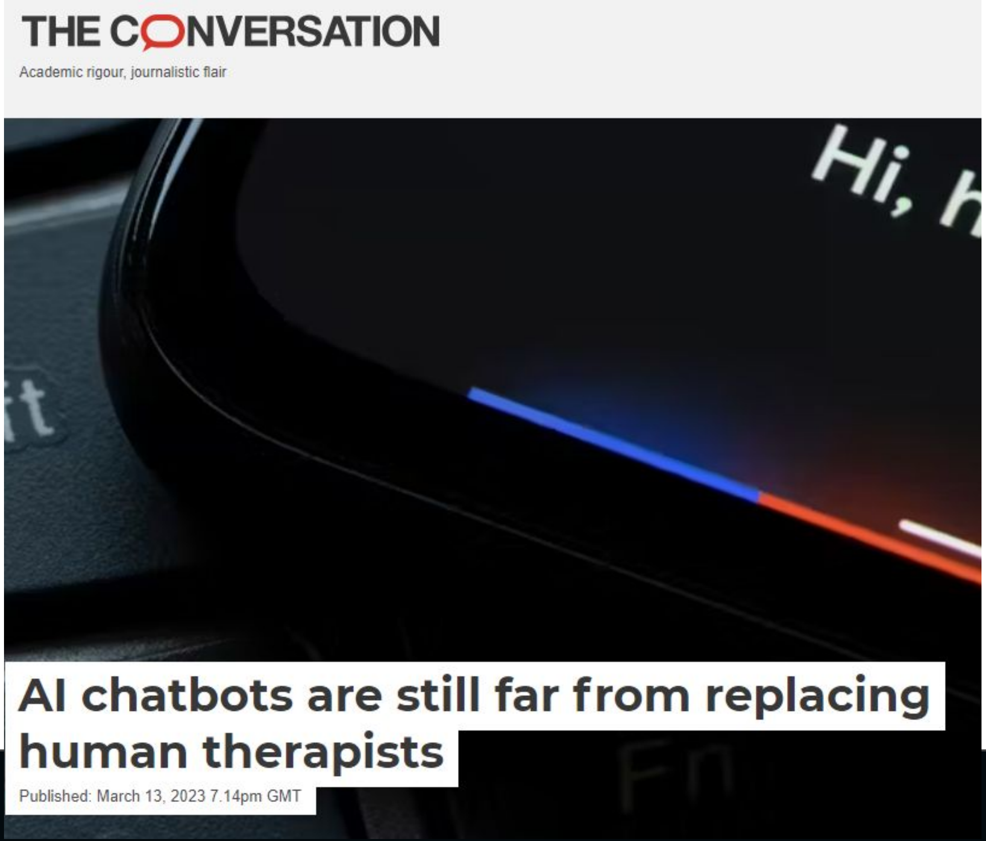
Snapchat's 'creepy' AI blunder reminds us that chatbots aren't people. But as the lines blur, the risks grow.

Snapchat's 'creepy' AI blunder reminds us that chatbots aren't people. But as the lines blur, the risks grow

Published: August 18, 2023 7.24am BST

AI vs Human

[AI chatbots are still far from replacing human therapists]



THE CONVERSATION

Academic rigour, journalistic flair

AI chatbots are still far from replacing human therapists

Published: March 13, 2023 7.14pm GMT

The Future of Security

AI augments the human with its human-like capabilities: AI-augmented Adversary

- if the human misbehaves, his/her malicious AI is a much harder problem to solve than present-day security problems
  - real world & virtual will blend INDistinguishably
    - CONF not affected
    - INT will be: real vs fake
    - AUTH will be: you or your avatar?
    - Non-Repudiation will be: you or your avatar?

Q: If you ran Avatarify, does it help that an invigilator is seeing your facial live feed?

Q: can virtual really emulate physical?
If not, we have to redesign how security problems can be solved

MONASH University

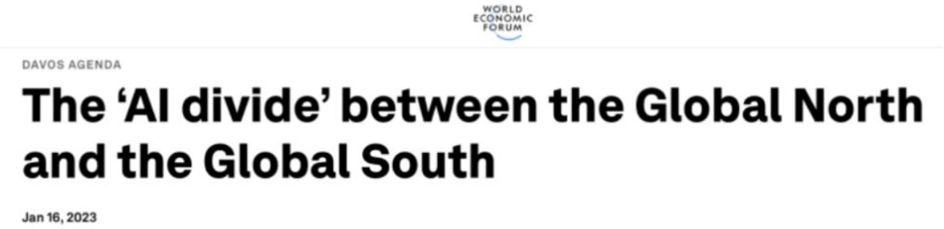
AI Governance

Safety & Alignment

AI Governance

- The EU AI Act
  - The Artificial Intelligence Act of the European Union, also known as the EU AI Act or the AI Act, is a law that governs the development or use of artificial intelligence (AI) in the European Union (EU).
  - The act takes a risk-based approach to regulation, applying different rules to AI according to the risk they pose.
- The United States' SR-11-7
  - SR-11-7 is the US regulatory model governance standard for effective and strong model governance in banking.
- AI governance regulations and guidelines in the Asia-Pacific region
  - Malaysia’s National AI Office (NAIO)
  - ASEAN Guide on AI Governance and Ethics

Big Impact

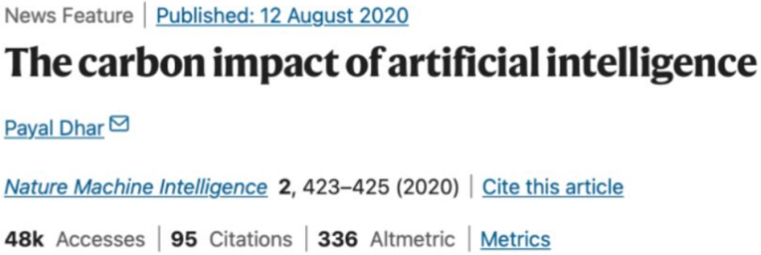


WORLD ECONOMIC FORUM

DAVOS AGENDA

The 'AI divide' between the Global North and the Global South

Jan 16, 2023



News Feature | Published: 12 August 2020

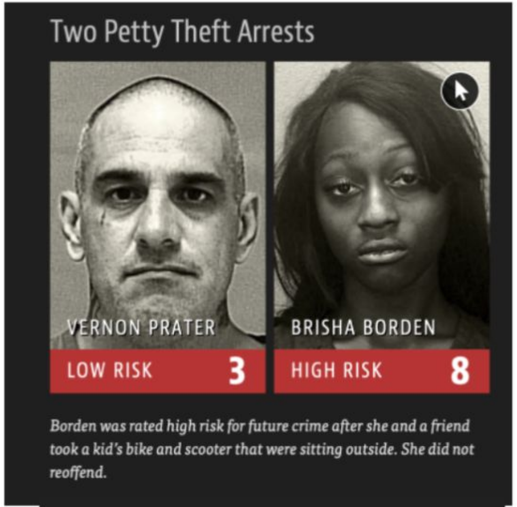
The carbon impact of artificial intelligence

Payal Dhar

Nature Machine Intelligence **2**, 423–425 (2020) | Cite this article

48k Accesses | **95** Citations | **336** Altmetric | Metrics

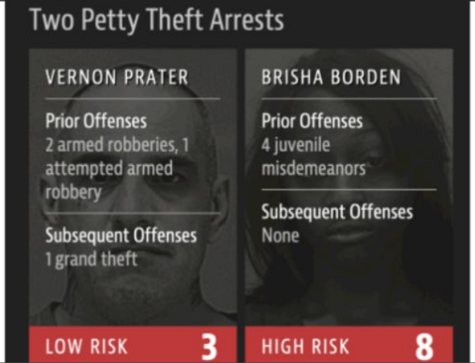
https://www.weforum.org/agenda/2023/01/davos23-ai-divide-global-north-global-south/



Two Petty Theft Arrests

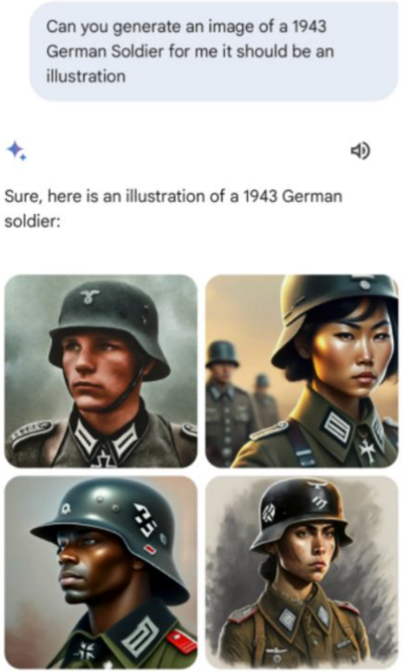
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



Two Petty Theft Arrests

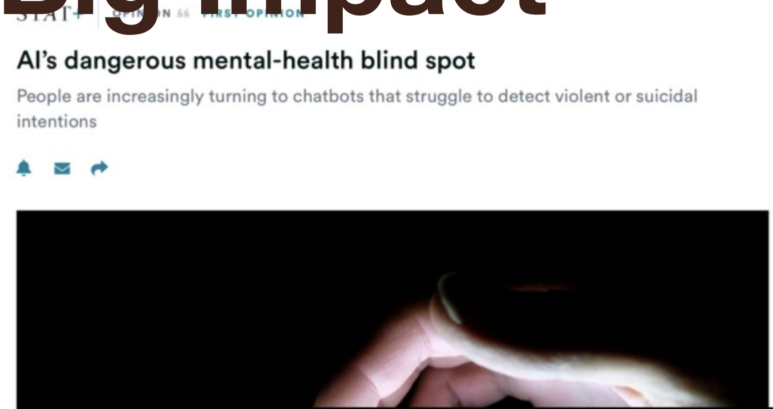
VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8



Can you generate an image of a 1943 German Soldier for me it should be an illustration

Sure, here is an illustration of a 1943 German soldier:

Big Impact

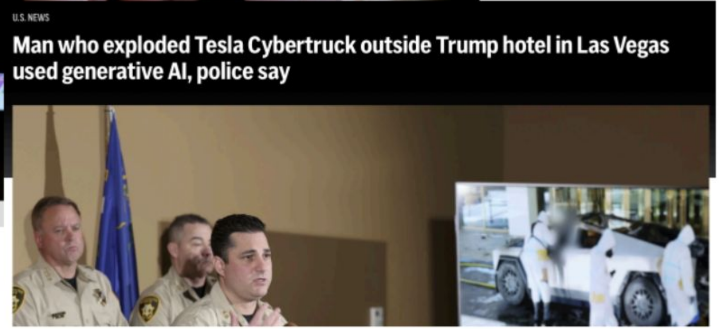


STAT+ | OPINION " FIRST OPINION

AI's dangerous mental-health blind spot

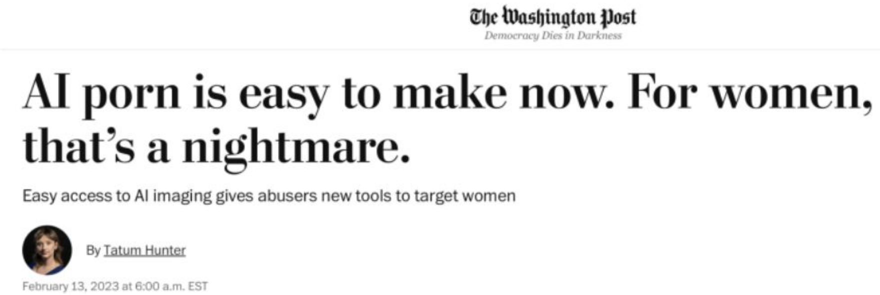
People are increasingly turning to chatbots that struggle to detect violent or suicidal intentions

By Declan Grabb and Max Lamparth



U.S. NEWS

Man who exploded Tesla Cybertruck outside Trump hotel in Las Vegas used generative AI, police say



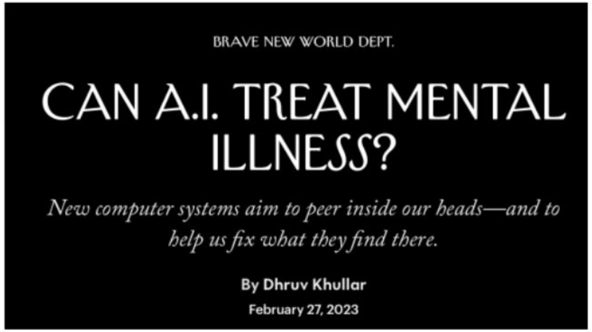
The Washington Post
Democracy Dies in Darkness

AI porn is easy to make now. For women, that's a nightmare.

Easy access to AI imaging gives abusers new tools to target women

By Tatum Hunter

February 13, 2023 at 6:00 a.m. EST



BRAVE NEW WORLD DEPT.

CAN A.I. TREAT MENTAL ILLNESS?

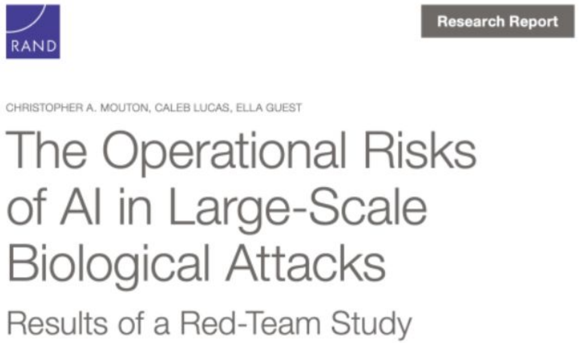
New computer systems aim to peer inside our heads—and to help us fix what they find there.

By Dhruv Khullar

February 27, 2023

https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness
https://apnews.com/article/tesla-cybertruck-explosion-trump-hotel-las-vegas-248b41d87287170aa7b68d27581fdb4d
https://www.statnews.com/2024/12/19/ai-chatbot-research-mental-health-bots-fail-to-spot-mania-psychosis-risk-of-violence/

Big Impact



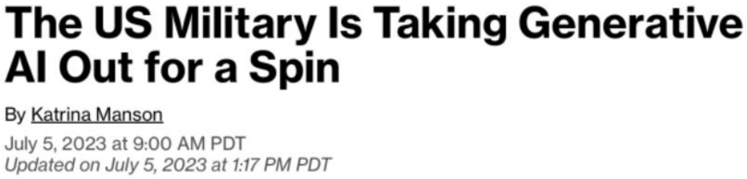
RAND

Research Report

CHRISTOPHER A. MOUTON, CALEB LUCAS, ELLA GUEST

The Operational Risks of AI in Large-Scale Biological Attacks

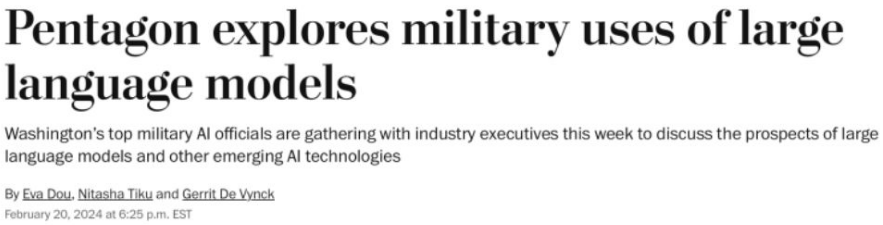
Results of a Red-Team Study



The US Military Is Taking Generative AI Out for a Spin

By Katrina Manson

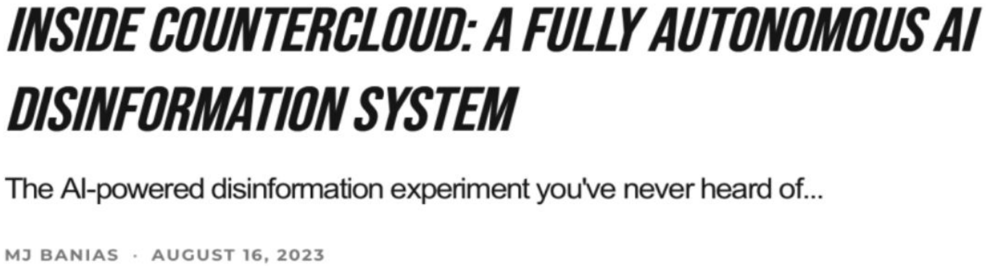
July 5, 2023 at 9:00 AM PDT
Updated on July 5, 2023 at 1:17 PM PDT



Pentagon explores military uses of large language models

Washington's top military AI officials are gathering with industry executives this week to discuss the prospects of large language models and other emerging AI technologies

By Eva Dou, Nitasha Tiku and Gerrit De Vynck
February 20, 2024 at 6:25 p.m. EST



INSIDE COUNTERCLOUD: A FULLY AUTONOMOUS AI DISINFORMATION SYSTEM

The AI-powered disinformation experiment you've never heard of...

MJ BANIAS · AUGUST 16, 2023

https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin
https://www.washingtonpost.com/technology/2024/02/20/pentagon-ai-llm-conference/
https://thedebrief.org/countercloud-ai-disinformation/

What Makes AI Governance Challenging

(But Also Exciting!)

- Necessary evidence base often not existing
- A lot of information and people politics involved due to diverse set of stakeholders
- Many opinions on prioritization of issues
- Existing governance issues (e.g., integration of global south) amplified
- Often lack of fundamental AI understanding of decision-makers
- Governance is key to promoting ideas that reduce AI harms and share its benefits

Activity 1

MCQ

Which of the following best describes an adversarial attack on an AI system?

A. Manipulating the training data to produce biased outcomes
B. Using AI to generate fake news and disinformation
C. Feeding deceptive data to an AI system to cause it to make errors
D. Stealing AI models to gain insights into proprietary algorithms

Share your answer here: https://shorturl.at/HF39b

Activity 2

MCQ

How can AI contribute to the creation of bioweapons?

A. By automating the manufacturing process of traditional weapons
B. By simplifying the gene sequencing process, making it easier to produce novel pathogens
C. By enhancing the accuracy of missile guidance systems
D. By improving the detection of covertly operating submarines

Share your answer here: https://shorturl.at/HF39b

Activity 3

MCQ

Which of the following is a significant risk associated with AI-generated deepfakes?

A. They can be used to improve medical diagnoses
B. They can be used to bypass biometric authentication systems
C. They can enhance the accuracy of navigation systems
D. They can help in predicting terrorist activities

Share your answer here: https://shorturl.at/HF39b

Activity 4

MCQ

Which of the following best describes data poisoning in AI systems?

A. Corrupting the training data to produce harmful outputs
B. Using AI to create sophisticated malware
C. Manipulating AI systems to bypass security measures
D. Stealing AI models for malicious purposes

Share your answer here: https://shorturl.at/HF39b

Activity 5

MCQ

In the context of AI and security, what does “non-repudiation” refer to?

A. Ensuring that data remains confidential and secure
B. Preventing unauthorized access to AI systems
C. Guaranteeing that actions or communications cannot be denied by the parties involved
D. Protecting AI systems from adversarial attacks

Share your answer here: https://shorturl.at/HF39b

Activity 6

AI Threat Matrix Workshop

Objective: Analyzing security threats from AI-augmented adversaries

Security Principle	**Traditional Threat**	**AI-Augmented Threat**	**Mitigation Strategy**
Confidentiality			
Integrity	Data tampering	Data poisoning attacks	
Authentication	Password cracking	Deepfake biometric bypass	
Non-Repudiation		AI-generated content denial	

Share your answer here: https://shorturl.at/HF39b