The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation

An OpenAI Research Report

Investigation By
Brian McMahon

26 February 2018

Investigation Overview

Key insights from OpenAI paper published on 20 February 2018

"How can we forecast, prevent and mitigate the harmful effects of malicious uses of AI?"

“Dual Use” Technology

Digital Security

- Synthetic images, audio and text
- Network security

Political Security

- Surveillance
- Media Control

Physical Security

- Drones
- Self-Driving Cars

Example: “Deep Fakes”

Trump’s real Face synthetically mapped onto SNL’s Alec Baldwin

https://www.youtube.com/watch?v=hoc2RISoLWU

Recommendations

1. Collaboration between policymakers and tech researchers
2. “Dual-use” nature considered in research priorities
3. Look to mature “dual-use” industries for best practices
4. Expand range of stakeholders involved in addressing key challenges

High priority research areas

Learn from and with the cybersecurity community

Explore different openness models

Promote a culture of responsibility

Develop technological and policy solutions

Resources

Brundage, Miles. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation." OpenAI. 20 Feb 2018.

Surveys the landscape of potential security threats from malicious uses of AI and proposes ways to better forecast, prevent and mitigate these threats.

"Slaughterbots." Future of Life Institute. 13 November 2017.

Satire on hypothetical weaponized drone scenario.

"GifFakes" Subreddit.

Subreddit for gifs created with FakeApp, a program that fabricates neural network-generated faceswap videos.

QUESTIONS?

APPENDIX

Themes of Interest

- Drones
- Social media bots / Fake news
- Spear phishing / social engineering sophistication
- A/V forgery - deepfakes
- Terrorism
- Authoritarian regimes / surveillance / censorship
- Research “openness” - look to cybersecurity
- “Dual-use” - peaceful/military aims, beneficial/harmful ends
- Data poisoning
- Law enforcement
- Backdoors
- Ethics - IEEE...
- Lethal autonomous weapons

Investigation Overview

Key insights from the paper *"The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation"* released on OpenAI and arXiv.org on 20 February 2018

- Authoring institutions include OpenAI, University of Cambridge and University of Oxford, Future of Humanity Institute, Centre for Study of Existential Risk, Center for New American Security and Electronic Frontier Foundation
- Results of a a workshop convened at the University of Oxford in February 2017 by experts on AI safety, drones, cybersecurity, lethal autonomous weapons systems, and counterterrorism

Seeks to answer "how can we forecast, prevent and (when necessary) mitigate the harmful effects of malicious uses of AI?"

Authors

26 individuals representing:

- University of Oxford
- Future of Humanity Institute, University of Oxford
- Centre for the Study of Existential Risk
- University of Cambridge
- Center for a New American Security
- Electronic Frontier Foundation
- OpenAI

Scope

- Only consider AI tech currently available or plausible in the next 5 years
  - Focused in particular on tech leveraging machine learning
- Only consider scenarios where an individual or org deploys AI tech or compromises an AI system with the aim to undermine security of another individual, organization or collective
- Work is within study of social implications of, and policy responses to, AI

What is OpenAI?

- OpenAI is a non-profit AI research company founded by Elon Musk and Sam Altman that aims to promote and develop friendly AI in such a way as to benefit humanity as a whole.
- The founders are motivated in part by concerns of existential risk from artificial general intelligence.

What is AI?

AI refers to the use of digital technology to create systems that are capable of performing tasks commonly thought to require intelligence

ML is variously characterized as either a sub-field of AI or a separate field, referring to the development of digital systems that improve performance on a given task over time through experience

AI: Many Positive Use Cases

- Automatic speech recognition
- Machine translation
- Medical image analysis
- Spam filters
- Search engines
- Driverless cars
- Digital assistants, esp for nurses and doctors
- AI-enabled drones for expediting disaster relief ops
- Expediting scientific research
- Governance

“Dual Use” Tech

Positive Use	Negative Use
Facial recognition	Applied to autonomous weapons systems
Generate synthetic images, text and audio	Impersonation, sway public opinion
Drone package delivery	Loaded with explosives and/or firearms
“White hat” cyber defenses	“Black hat” cyber attacks
Malware detection	Vulnerability detection
Surveillance to catch terrorists	Oppress ordinary citizens
Autonomous vehicles and systems	Manipulation at central points of failure
Filter fake news	Manipulate public opinion

Threat Scope

Expansion of existing threats. Lowers costs and barriers of existing attacks

- Spear phishing, network penetration, malware

Introduction of new threats. New attacks may arise through the use of AI systems to complete tasks otherwise impractical for humans

- Impersonation and communication falsification

Increased collateral damage of threats. More effective, finely targeted, difficult to attribute, and likely to exploit vulnerabilities in AI systems

- Control over drone / bot swarms; adversarial training data “poisoning”

Cybersecurity

Automated cyberattacks to alleviate existing tradeoff between scale and efficacy

Expand threat associated with labor-intensive cyberattacks

- Spear phishing: Collecting and using information specifically relevant to target (name, gender, affiliations, topics of interest, etc), allowing customization of facade to appear more relevant / trustworthy

Expect novel attacks to exploit vulnerabilities of:

- Humans (ie social engineering, especially interaction impersonation wrt speech, video or text/social media)
- Software(ie automated vulnerability detection for penetration or malware)

Physical Security

Central points of failure for autonomous vehicles, power systems

Control of complex systems (such as swarms of micro-drones)

Autonomous weaponization for terrorism and warfare

- Attack and explosive drones

Remote and autonomous increases “psychological distance” from wrongdoing)

Political Security

Surveillance / Privacy (facial recognition, tracking)

Persuasion / Manipulation (bots to spread propaganda, fake news, influence voting, , control popular opinion, sow discord)

Deception (ie manipulating communications of global leaders)

- Impersonate and falsify evidence by faking voice, video or text (see DeepFakes on next slide)
- “Seeing is believing” is no longer necessarily the case

Areas of further research

Learning from Cybersecurity Community

- Red teaming
- Formal verification
- Responsible disclosure of AI vulnerabilities
- Forecasting security-relevant capabilities
- Security tools
- Secure hardware

Areas of further research

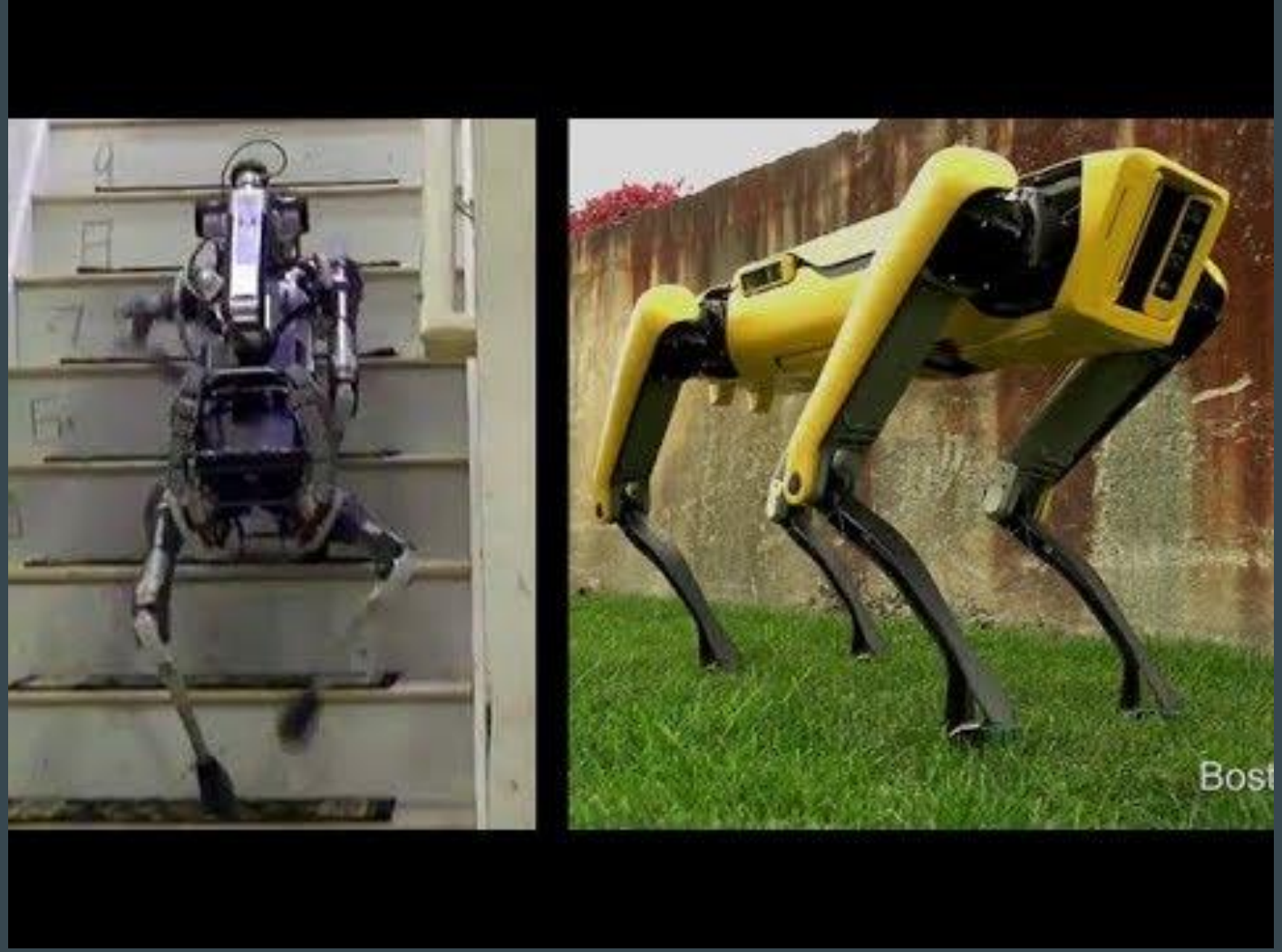
Explore Different Openness Models

- Prevailing norms in ML research community are open
- Also increases power of tools available to malicious actors
- Worth considering abstaining from, or delaying, publishing of some findings related to AI for security reasons?

Tech Demo -1

https://www.youtube.com/watch?v=RCXGpEmFbOw

Tech Demo - 2



https://www.youtube.com/watch?v=3OKZ_n8QW4w

https://www.youtube.com/watch?v=HipTO_7mUOw