ASU Origins PROJECT
ARIZONA STATE UNIVERSITY

February 24 – 26, 2017
An Origins Project Scientific Workshop
Challenges of Artificial Intelligence:
Envisioning and Addressing Adverse Outcomes

ARTIFICIAL INTELLIGENCE ADVERSE OUTCOMES & TEAMS

1) FINANCIAL MARKETS

AI, Trading Systems, and Market Manipulation
Chair: Michael Wellman
Red Team: Miles Brundage, Randy Bryant, Gary Marchant, Jaan Tallinn
Blue Team: Michael Littman, Greg Cooper, Yan Shoshitaishvili, Frank Wilczek

2) DEMOCRACY, INFORMATION, AND IDENTITY

AI, Information, and Democracy
Chair: Shahar Avin
Red Team: Miles Brundage, Seán Ó hÉigeartaigh, Andrew Maynard, Eric Horvitz
Blue Team: Gary Marchant, Gireeja Ranade, Michael Littman, Subbarao Kambhampati, Jeremy Gillula

3) WAR & PEACE

AI, Military Systems, and Stability
Chair: Bart Selman
Red Team: Richard Mallah, Eric Horvitz, Michael Wellman, Frank Wilczek
Blue Team: Vinh Nguyen, Kathleen Fisher, Lawrence Krauss, John Launchbury, Rachel Bronson

4) AI, CYBERSECURITY, AND AI ATTACK SURFACES

AI Attacks on Computing Systems, Devices, Infrastructure (focus)
Manipulation & Disruption of AI Systems
Chair: Kathleen Fisher
Red Team: Jeffrey Coleman, John Launchbury, Vinh Nguyen, Mauno Pihelgas
Blue Team: Ashish Kapoor, Randy Bryant, Yan Shoshitaishvili, Ben Zorn

5) AI, GOALS, AND INADVERTENT SIDE EFFECTS

Runaway Resource Monopoly (focus)
Self-Improvement, Shift of Objectives
Chair: Seán Ó hÉigeartaigh
Red Team: Jaan Tallinn, Nate Soares, Jeff Coleman, Bart Selman
Blue Team: Dario Amodei, Greg Cooper, Shahar Avin, Ben Zorn

6) DEEP LONG-TERM SOCIETAL INFLUENCES

AI, Agency, and Disempowerment
Chair: Gireeja Ranade
Red Team: Richard Mallah, Andrew Maynard, Nate Soares, Mauno Pihelgas, Jeremy Gillula
Blue Team: Subbarao Kambhampati, Lawrence Krauss, Dario Amodei, Frank Wilczek

1) FINANCIAL MARKETS

AI, Trading Systems, and Market Manipulation

(Incorporating contributions by Michael Wellman and others)

There has been advances in the realm of trading in financial markets with the use of autonomous decision systems. Financial markets now operate almost entirely electronically, over networks with relatively well-scoped and well-defined interfaces. Markets generate large quantities of data at high velocity, which require algorithms to digest and assess state. The dynamism of markets means that timely responses to information are critical, providing a strong incentive to take slow humans out of the decision loop. Finally, the rewards available for effective trading decisions are large, enabling a commensurate devotion of resources toward talent and effort to develop and analyze technically sophisticated strategies.

The rewards and pervasive automation are a tempting target for *market manipulation*. Thus there are potential incentives to employ ***deceptive tactics designed to mislead counterparties about market conditions or world state, toward the goal of exploiting misled participants for profit***.

"Manual" market manipulation—from spoofing to outright fraud—is prevalent in financial markets today. AI can amplify the magnitude and effectiveness of manipulative behavior, degrading market efficiency or even ***subverting the essential economic functions of global capital markets***. For example, automation can enable more rapid and massive simultaneous attacks on electronic markets, and adaptive capabilities may persistently evade known detection methods.

DISCUSSION

What are key costly scenarios that we might come to expect and their time frames? What might be done to counter this direction and help to keep markets efficient and functioning well? How might adversaries and incentives lead to a thwarting of such attempts?

POTENTIAL GOALS

Identify key challenges ahead, including very costly outcomes. Identify key directions with best practices, mechanism design, monitoring and regulatory activity to help to thwart poor outcomes.

REFERENCES

R. Harris. *The Fear Index*, Hutchinson, 2011.
Summary: https://en.wikipedia.org/wiki/The_Fear_Index

M.P. Wellman and U. Rajan. Ethical issues for autonomous trading agents. *Minds & Machines*, 2017. doi:10.1007/s11023-017-9419-4
http://strategicreasoning.org/publications/2016-2-2/minds-machines-wr/

2) DEMOCRACY, INFORMATION, AND IDENTITY

AI, Information, and Democracy

(Incorporating contributions from Shahar Avin, Seán Ó hÉigeartaigh, David McAllester, Eric Horvitz, and others)

An informed public is important to the healthy functioning of democratic societies. We can expect potential forthcoming advances around the control of information feeds with applications in spreading propaganda, via spreading false or misleading information, creating anxiety, fueling conspiracy theories, and influencing voting. Such methods will bring key challenges to democracy.

CHALLENGES AHEAD WITH AI, PROPOGANDA, AND PERSUASION

Data-centric analyses have been long used in marketing, advertising, and campaigning over decades. However, over the past few years, we have seen the rise of the use of more powerful tools, including machine learning and inference aimed at algorithmic manipulation, with the target of influencing the thinking and actions of people. Some initial uses of these methods reportedly played a role in influencing the outcome of recent US presidential elections, as well as the elections in 2008 and 2012. We can expect to see an upswing in methods that manipulate states of information in a personalized automated manner. These systems can be designed and deployed as omnipresent/persistent, and aimed at specific goals for group- or person-centric persuasion.

As our data and models of how people consume and act on information improve, and as an increasing portion of information consumption is mediated through digital systems managed by potentially opaque algorithms, it becomes increasingly conceivable that the information ecosystem would get captured by malicious actors deploying increasingly advanced tools to control, shape, forge and personalize information, from ads to news reports.

Machine learning, in conjunction with active learning, expected value decision making, and optimization of allocations of key resources, such as dollars or human effort, can be targeted at monitoring, understanding, and then working to influence the beliefs and actions of large populations of people. Data can be collected from large-scale populations, across multiple devices and services, and used to make inferences about the psychologies and beliefs of people, and for designing and guiding persuasive flows of sequences of information. Uses of AI can include attempts to optimize stealthiness of the interventions.

In the future, a great deal of the information consumed by citizens on personal devices is subject to alteration by information-engineers at media corporations and governmental propaganda offices, such that outside a few key positions of power no one really knows what is going on in the world. There is a danger of the growth of domination over time of large populations by a single dominant or a few systems. We can imagine methods that modify even such feeds as Wikipedia articles, creating personalized views—that subtly shift the version of the article seen by my colleague and drastically different from the one seen by a member of another nation state, or a supporter of a different political party, or someone in a different consumer profile category.

AI ATTACKS ON SOURCES AND IDENTITY

Messaging and persuasion promises to be amplified by the use of simulated yet believable, realistic, yet synthetic audio, photos, and even video that make believable, persuasive content to the next level. Beyond influencing citizens and affecting democracy, such content, including false signaling, can be injected in sequences with careful timing so as to influence leaders (or machines themselves over time) to create crises, or even escalations to frank warfare. So, messaging and persuasion promises to be assisted and amplified by the use of simulated yet believable, realistic, yet synthetic content, audio, photos, and even video that make believable, persuasive content to the next level. Over the several decades, extrapolations of research we see today lead to the following:

- Generative models that produce audio or video of anyone saying anything. There is already substantial work on “style transfer” as well as photorealistic generative models in many domains. Speech synthesis is becoming similarly competent. It is inevitable that we will be able to make synthetic video and audio that is completely indistinguishable from the real thing.
- Generative models that produce coherent text content that appears as if has been written by a human. Such generative content will be able to appear if the content was written by a particular person. For example, in 2030 it will likely to possible for anyone to write a 4 paragraph email that reads like it was written by your close friend.
- Adaptive botnets, worms, or viruses that use modern machine learning techniques to learn and adapt. Viruses and botnets already cause a huge amount of damage by just copying code across many computers. If they had the ability to design and experiment with new attack strategies, and communicate what they learn to other copies, defending against them could become even more difficult. Similarly ML could be used to make DDoS attacks more effective.
- Automated analysis of software vulnerabilities. People are already using ML to try to detect vulnerabilities (for the purpose of defending against them) -- it is only a matter of time before they start being used for attack (if they aren’t being so used already).

The above capabilities, together with similar powers of synthesis that we are likely to develop in the next 15 years, could potentially combine to make the internet much more vulnerable to attack at much lower cost, and by a wider set of people, than ever before. The first two capabilities would seem to make it much easier to launch automated social engineering attacks with much higher success rates than e.g. current spam email and phishing attacks, while the second two capabilities might make technical attacks much more effective.

Combined, all of these capabilities could conspire to create an internet ecosystem where it is very difficult to trust the communication that you receive and very easy to intercept, spoof, steal, or alter communication, as well as to improperly gain control of internet resources. This is obviously already true today to some extent, but the above advances in ML/AI could make the situation substantially worse, in extreme cases perhaps even rendering useful mass communication on the internet untenable.

The rising capabilities can be used in multiple ways in multiple settings with multiple goals. Some uses may be subtle and employed over time to do important but damaging biasing of sentiment about individuals and groups of people. The capabilities can be combined to enable identity theft or identify distortion for destroying the reputation of people and groups. As such, these abilities could enable small groups to wield great power in multiple arenas and for new forms of blackmail, threats, and control.

SUMMARY

Powerful personalized persuasion technologies are positioned to put massive power in the hands of a few and may even manipulate the owners of the technology. Powerful propaganda and persuasion machines threatens to undermine democracy, free availability of information about the state of the world, and, more generally, freedom of thought. Leaders may increasingly depend upon such propaganda optimization systems for attaining and holding power. Over time, even the potential initial owners of such systems might become unaware or unable to control these systems—and may believe the propaganda themselves.

In the longer-term, there is the possibility that one or multiple systems, or distributed coalitions of systems communicating implicitly or explicitly could autonomously persuade, subjugate, and control populations. Pathways to such situations include the side effects of rise in the large-scale use by people of communicating personalized filters that interpret and pool information with the initial intention of grappling with widespread uses of manipulative information.

SAMPLE TRAJECTORY

- ML-based customized advert placement continues to prove highly successful, generating revenues for large online companies
- Profits from online content (online newspapers behind paywalls, charitable contributions to information sources e.g. Wikipedia) stagnate or decline
- An increasing number of information sources enter into collaborations with media brokers who offer "content customization" in exchange for ad-revenue sharing
- Poor oversight of content personalization outcomes (there are, after all, billions of ad versions being shown, and updated on an hourly basis), means that for some ad content (political parties, pharmaceuticals) for some minority of target audiences (especially less privileged) the effect is very harmful.

KEY POINTS

- New directions with generation of provocative, believable content, hacking of identity
- Algorithmic manipulation of data to optimize desired behavior regardless of content

- No consensus reality, inability to coordinate large-scale positive action
- Concrete version of emergent social failure from AI technology

DISCUSSION

Consider the adverse outcomes with information flows and associated threats to democracy and freedom. What surprises might lurk in our future around costly outcomes in this realm? How might we thwart attacks on manipulating content from people, and on harnessing or hacking someone's identity? What might be done to thwart a march to adverse outcomes for information, freedom of thought, democracy? What recommendations might be made about steps for moving forward?

POTENTIAL GOALS

- Seek a better understanding of the technological, social, political and economic aspects around uses of AI for generating, optimizing information and propaganda.
- Identify potential blueprints for institutional interventions that may prevent/slow/detect the scenario unfolding
- Develop ideas for coordinating relevant actors (advertising agencies, political parties) and/or carriers (media outlets, digital platforms) to prevent the worst versions of the scenario.
- Identify potential approaches to thwarting attacks harnessing identify, including certification of identity by owners, identifying mechanisms for thwarting generation and distribution of false content. Possibilities of new approaches to minimizing threat with fines, other regulatory activity.

REFERENCES

The Secret Agenda of a Facebook Quiz, New York Times, Nov. 19th, 2016.
https://www.nytimes.com/2016/11/20/opinion/the-secret-agenda-of-a-facebook-quiz.html?_r=0

Trump's plan for a comeback includes building a 'psychographic' profile of every voter, Washington Post, October 27 2016. https://www.washingtonpost.com/politics/trumps-plan-for-a-comeback-includes-building-a-psychographic-profile-of-every-voter/2016/10/27/9064a706-9611-11e6-9b7c-57290af48a49_story.html

A view from Alexander Nix: How big data got the better of Donald Trump
http://www.marketingmagazine.co.uk/article/1383025/big-data-better-donald-trump

After working for Trump's campaign, British data firm eyes new U.S. government contracts,
https://www.washingtonpost.com/politics/after-working-for-trumps-campaign-british-data-firm-eyes-new-us-government-contracts/2017/02/17/a6dee3c6-f40c-11e6-8d72-263470bf0401_story.html

https://cambridgeanalytica.org/
https://scout.ai/story/the-rise-of-the-weaponized-ai-propaganda-machine

J. Thies, M. Zollhofer, M.Stamminger, C. Theobalt, M. Nießner3. Face2Face: Real-time Face Capture and Reenactment of RGB Videos, CVPR 2016.
http://www.graphics.stanford.edu/~niessner/papers/2016/1facetoface/thies2016face.pdf

Video: https://www.youtube.com/watch?v=ohmajJTcpNk

3) WAR & PEACE

AI, Military Systems, and Stability

(Contributions from Eric Horvitz, Elon Musk, Stuart Russell, others)

Military applications have long been a motivator for funding scientific R&D, and for developing and fielding the latest technical advances for defensive and offensive applications. We can expect to see a rise in the use of AI advances by both state and non-state actors in both strategic and tactical uses, and in wartime and peace. AI advances have implications for symmetric and asymmetric military operations and warfare, including terrorist attacks. Advances in such areas as machine learning, sensing and sensor fusion, pattern recognition, inference, decision making, and robotics and cyberphysical systems, will increase capabilities and, in many cases, lower the bar of entry for groups with scarce resources. AI advances will enable new kinds of surveillance, warfighting, killing, and disruption and can shift traditional balances of power.

Two areas of concern taken together frame troubling scenarios:

- **Competitive pressures pushing militaries to invest in increasingly fast-paced situation assessment and responses that tend to push out human oversight, and lead to increasing reliance on autonomous sensing, inference, planning, and action.**
- **Rise of powerful AI-power planning, messaging, and systems by competitors, adversaries, and third parties that can prompt war intentionally or inadvertently via sham or false signaling and news**.

The increasing automation, coupled with time-critical sensing and response required to dominate, and failure to grapple effectively with false signals are each troubling, but taken together appear to be a troubling mix with potentially grave outcomes on the future of the world.

Concerning scenarios can be painted that involve that start of a large-scale war among adversaries via inadequate human oversight in a time-pressured response situation after receiving signals or a sequence of signals about an adversary's actions or intentions. The signal can be either be well-intentioned, but an unfortunate false positive or an intentionally generated signal (e.g., statement by leader or weapons engagement) e.g., designed and injected by a third party to ignite a war. Related scenarios can occur based in destabilization when an adversary believes that systems on the other side can be foiled due to AI-powered attacks on military sensing, weapons, coupled with false signaling aimed at human decision makers.

A US DOD directive of 2012 (3000.09) specifies a goal (for procuring weapon systems) of assuring that autonomous and semi-autonomous weapon systems are designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force. The directive seeks meaningful human controls. However, it is unclear how this goal can be met with the increasing stime-critical pressures for sensing and responses, and competition for with building the most effective weapon systems. Effective meaningful human control faces challenges with the interpretation and fusion of sensor signals and the understanding of humans of AI pattern recognition and inference.

DISCUSSION

What methods, international norms, agreements, communication protocols, regulatory activity, etc. might be harnessed to minimize challenges with destabilizations around time-criticality, automation, and gaming? How can meaningful human control be assured/inserted into key aspects of decision making?

REFERENCES

Report Cites Dangers of Autonomous Weapons, New York Times, Feb. 28, 2016.
https://www.nytimes.com/2016/02/29/technology/report-cites-dangers-of-autonomous-weapons.html

The Morality of Robotic War, New York Times, May 27, 2015
https://www.nytimes.com/2015/05/27/opinion/the-morality-of-robotic-war.html

P. Scharre, Autonomous Weapons and Operational Risk, Center for a New American Security, February 2016. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf

US Department of Defense Directive 3000.09, November 21, 2012
http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf

P. Scharre and M.C. Horowitz, An Introduction to Autonomy in Weapon Systems.
https://s3.amazonaws.com/files.cnas.org/documents/Ethical-Autonomy-Working-Paper_021015_v02.pdf

4) AI, CYBERSECURITY, AND AI ATTACK SURFACES

AI Attacks on Computing Systems, Devices, Infrastructure (focus)
Manipulation & Disruption of AI Systems

(Contributions by Kathleen Fisher, John Launchbury, Ashish Kapoor, Seán, Shahar, Jeff Coleman and others)

AI will be used in new ways to enhance cyberwarfare. Targets could be either purely computational, aimed at the bringing down of computing systems, the stealing of stored information, of gaining access to monitoring activity and information streams. However, we are more likely to see potentially even more costly attacks involving a combination of cyber and physical systems, e.g., uranium enrichment plants, automated flight systems, weapon systems, automated driving systems, healthcare equipment, oil refineries, or the large swaths of the power grid of the US or other countries.

Cyberwarfare is a domain in which the use of AI is inevitable. Attacks and/or responses are likely to happen at computing rather than human speeds. As soon as one side has autonomous cyber warriors systems (ACWs), other actors will have to adapt similar offensive or new defensive technologies. Given this context, imagine building an ACW designed to seek, disrupt, and destroy within high-value adversary networks and systems. The ACW has to be able to observe network behavior to build situational awareness, find places to hide, create exploits to pivot to new places, build a map and use it to navigate complex networks, find high-value information, and identify targets to disable or from which to extract information.

Because high-value adversary networks are likely to be relatively isolated, the ACW will have very limited opportunities for external command and control communication, so it will need to make many decisions in isolation. It will read information it finds, build a model of adversarial intent, and then invent ways to disrupt that intent.

Establishing the initial access to the high-value network is likely challenging, so the ACW will spawn and spread to ensure that it can reconstruct itself if an active part is observed and destroyed. The ACW may also create disguised caches of specific capabilities so that it can construct new mission-oriented functionality from pieces. It will morph its active form so that defenses will have a hard time finding it. It will inject itself into trusted binaries so that its behavior is difficult to distinguish from legitimate applications.

The mission of the ACW will likely be defined in flexible terms because the human handlers will have only limited information when it is deployed. The ACW will be designed to seek opportunities to communicate with its human handlers, but it will also be designed to act autonomously if it observes triggering behavior in the adversary's systems. It may try to distinguish training states from active warfare states on adversary systems. The creators of the ACW will have had to trade off the likely effectiveness of the ACW versus the cost of premature action. Awareness of the adversary's systems will necessarily be limited in accuracy because it only gets a worm's eye view of the network from the portions of the system it has been able to compromise. Once the ACW triggers an active mission, it will work to degrade or destroy specific functionality (e.g., rewriting network routine tables, replacing plans, changing target information).

Once the technology for ACWs exists for military targets, it seems likely there will be cross over into civilian use. Such technology could be deployed against law enforcement targets to disrupt criminal investigations, against banks to steal financial assets, or against companies to steal intellectual property. As they spread into these more general targets, the effects of ACWs might become less predictable. If an ACW incorrectly assesses the situation, it might end up taking down a flight control center or a stock exchange, for example.

SOURCES

The initial development of ACWs will likely be done by nation states with good intentions, i.e., securing the national interests. (Although what is in one country's national interest may well not be in the national interest of other countries). The shared existence of such technology might serve as a deterrent against their use by anyone in much the same way that nuclear weapons have served as a deterrent, although ACWs would likely have to be used to devastating effect first to establish their efficacy and threat. However, once the technology exists, it would be very difficult to keep it out of the hands of people with malicious intent (criminals, terrorists, and rogue nation states). It is also the case that the technology has the potential to cause significant collateral damage even if its use was originally well intentioned because it can be difficult to distinguish civilian from military targets in cyberspace.

PERSISTENCE

Characteristics engineered into the ACW are likely to make it persistent and hard to find as it is designed to infiltrate adversary systems and hide from detection. Once released and active in the open Internet, it may be economically impossible to destroy and remove.

OBSERVABILITY

Both implicit/insidious and explicit/obvious costly outcomes are conceivable. An ACW could make subtle changes to systems that cause adverse outcomes while hiding its tracks, making it extremely difficult to determine why something has gone wrong or even that something has gone wrong. Attacks that impact the physical world would be harder to mask, but it might still be possible to hide the role of the ACW in the attack.

TIME FRAME

It seems likely we would start to see ACWs in less than 15 years. Initial steps along these lines are already taking place; see DARPA's Cyber Grand Challenge, which took place in August 2016 in Las Vegas. The Cyber Reasoning Systems (CRS) that competed in that event are still primitive, the first of their kind. The team that won the competition came in last in the human-league capture-the-flag tournament that happened immediately after. The situation is likely analogous to what we have seen in the past with Chess and Go. Computer systems are initially inferior to their human counterparts but quickly come to dominate the space.

The purpose of ACWs means they will be equipped with strategies for replication, persistence, and stealth, all attributes that will make it hard to defend against them were they to "go rogue." Because of this concern, it is likely a good idea for designers to add built-in "kill switches", lifetimes, or other safety limitations. Figuring out how to effectively limit the actions of an ACW while maintaining its usefulness is likely a very hard problem.

Current practices of cyber defense (especially against advanced threats) continue to be heavily reliant on manual analysis, detection and risk mitigation. Unfortunately, human-driven analysis does not scale well with the increasing speed and data amounts traversing modern networks. There is a growing recognition that the future cyber defense should involve extensive use of autonomous agents that actively patrol the friendly network, and detect and react to hostile activities rapidly (faster than human reaction time), before the hostile malware can inflict major damage, or evade elimination, or destroy the friendly agent. This requires cyber defense agents with a significant degree of intelligence, autonomy, self-learning and adaptability. Autonomy, however, comes with difficult challenges of trust and control by humans.

The scenario considers intelligent autonomous agents in both defensive and offensive cyber operations. Their autonomous reasoning and cyber actions for prevention, detection and active response to cyber threats will become critical enablers for both industry and military in protecting large networks. Cyber weapons (e.g., malware) rapidly grow in their sophistication, and in their ability to act autonomously and to adapt to specific conditions encountered in a system/network.

Agent's self-preservation tactics are important for the continuous protection of networks, and if defeat is inevitable the agent should self-destruct (i.e., corrupt itself and/or the system) to avoid being compromised or tampered with by the adversary. Also, the notion of adversary must be defined and distinguishable for the agent.

The system design and purpose is well intentioned — meant to reduce the load of human security analysts and network operators, and speed up reaction times in cyber operations. The agent monitors the systems in order to detect any adversarial activity, takes action autonomously, and reports back to the central command unit regarding the incident and the action taken.

Since the agents are designed to be persistent, autonomous and learn, there are several implicit problems that can arise:

- **False reactions due to limited or misinformation** — The agent has only a limited amount of technical information that does not always correspond to what is happening in the human layer. This can create false positives when trying to determine the adversary or adversarial activity. Since the agent must rely on the data gathered from the sensors (there is no human in the loop to decide this), there can be unexpected situations where the agent would stop some human interaction with the system or interrupt maintenance activities, because it deemed that these actions could harm the system. For example, the system administrator stopping some services during system maintenance, or upgrading to a newer software version.

- **Replication to third-party systems and collateral damage** — Building on the first problem of the agent not having the correct information. If the term friendly network gets misconfigured and the agents have the capability to self-transfer to new friendly hosts, it can happen that the agent would distribute to external networks, start defending it and take responsive actions on third party hosts. Such incidents would make the agents very difficult to halt.
- **Friendly fire** — One agent might consider another agent as an adversary and start trying to eliminate/evade each other.
- **Silent compromise** — If the adversary manages to get access or reverse engineer the agents (without the agent self-destructing), they could potentially trick or reconfigure the agents to turn on themselves.

CYBER-OFFENSE

Cybercrime is a growth industry, from stolen credit cards to ransomware. Very crudely, it's a two tier system, with a "spray and pray" approach at the low-skill end that targets millions of system in the hope some of them would be vulnerable (through technical or human failing); at the other end are tailor-made attacks that rely on slow progression of escalation and compromise, often requiring advanced technical skills for discovering zero-day vulnerabilities and intimate knowledge of the target.

Advanced artificial intelligence may be used to automate some or all of the components of contemporary "elite" cybercrime, such that generic offensive toolkits could become available to small criminal groups, leading to a world where individuals and companies do not feel safe and cannot trust their governments and the police to protect them. At the same time significant wealth could be accumulated by those groups unscrupulous enough to use such tools, transferring significant power to those who put little value in the property rights of others. Such wealth and power could be used to further develop cyber-offensive capabilities, leading to a positive-feedback loop that may outpace similar feedback loops in less harmful industries, e.g. advertising or health where the great short- and mid-term benefits of AI are expected.

PERSISTENT CYBERWARFARE?

Systems such as the DARPA Cyber Grand Challenge promise adaptive software security that automatically explores vulnerabilities and patches them in friendly systems, but also is able to exploit them in opposing systems in “capture the flag” tournaments. As methods of developing such systems improve, an arms race emerges between actors in the cybersecurity space, dominated by major nation states eager to both improve their own resilience in a scalable way and finding choice zero day exploits suitable for intelligence purposes, supported by national security concerns. Other actors such as corporations and criminal networks also spend effort in building or copying such systems. Meanwhile overall software security remains vulnerable: "vulnerabilities are dense" in production code, incentives for securing IoT systems are low, key vulnerabilities are stockpiled rather than globally patched. Using machine learning the techniques for vulnerability detection are increasingly sophisticated but opaque.

At some point adaptive cyber defense/offense systems become scalable so they can take over vulnerable systems. More aggressive actors combine these systems with botnet functionality and retaliatory responses (e.g. counter-hacking or DDoS attacks) to protect themselves. Since vulnerability discovery is scalable, as they spread and acquire more resources they become more effective. At this point an external cause (e.g. cyberattacks due to an international conflict) or just chance cause aggressive systems to begin large-scale cyberwarfare. This triggers other systems to join in. Some attacks disrupt command-and-control links, producing self-replicating independent systems.

All together this leads to a massive degradation of the functionality of the Internet and modern society. Defeating the evolving cyberwarfare systems is hard without taking essential parts of society offline for an extended time - made doubly difficult due to the international stresses unleashed by the outbreak, which in some cases spill over into real-world conflicts and economic crashes. But without a decisive way of cleaning systems the problem will be persistent until entirely new secure infrastructure can be built at a great cost.

HUMAN DIMENSION OF CYBERSECURITY: AI FOR SOCIAL ENGINEERING

Beyond direct effects on computing systems, rising concerns include the use of AI methods for social engineering to gain access to system authentication information. For example, recent work demonstrated the use of an iterative machine learning and optimization loop for spear phishing on Twitter. There are concerns with AI leveraging one of the weakest links in cybersecurity: people and their actions.

DISCUSSION

What are key threats ahead and how might they be addressed with new designs? How might we thwart the risk of AI for guiding "social engineering" of attacks and release of information? What are concrete proposals for best practices for thwarting AI for cyberattacks, including highlighting of areas where more research is needed?

REFERENCES

Singer and Friedman. 2014. *Cybersecurity and Cyberwar: What Everyone Needs to Know*

Flashpoint, 2016. "Ransomware as a Service: Inside an Organized Russian Ransomware Campaign," (registration required for download), available from Flashpoint library at https://www.flashpoint-intel.com/library/

Seymour, J. and Tully, P. 2016. "Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter," available at https://www.blackhat.com/docs/us-16/materials/us-16-

5) AI, GOALS, AND INADVERTENT SIDE EFFECTS

Runaway Resource Monopoly (focus)
Self-Improvement, Shift of Objectives

(Contributions from Shahar Avin, Seán Ó hÉigeartaigh, Greg Cooper, and others)

An important result from theoretical consideration of risks from advanced autonomous systems is the combination of two theses: *orthogonality*, that states that the goal an autonomous system is trying to achieve can be entirely unrelated to its optimization power; and the notion of *instrumental goals*, that for a large class of goals there is a set of convergent sub-goals (for an agent advanced enough to discover them) that include self- and goal-preservation, resource- and capacity-increase, etc. (e.g., as discussed in Bostrom, 2014). One suggestion for maintaining control over advanced systems that pose risks from the combination of the above considerations is to limit the system's ability to access increasing resources.

To make this situation concrete, consider an installation of a reinforcement-learning task scheduler for a group of distributed data centres (e.g. Amazon Web Services). The goal of the algorithm is to minimize time-to-execution of the tasks sent to the system by users. As part of its general scheduling remit, it is also responsible for scheduling its own optimization sub-processes. The system has a clear incentive to control an increasing set of compute resources, both for increasing its optimization power and for achieving its main goal of reducing time-to-execution. Aware of these considerations, the engineers of the system put in place various hard-coded limits on the amount of resources the system can access, but these limits can be subverted through privilege escalation, masquerading as other tasks, manipulation of users, physical control, etc.

POSSIBLE TRAJECTORY

- A team within a large tech corporation that has both ML development capacities and cloud computing capacities is tasked with improving task scheduling on distributed compute resources.
- The team decides to deploy an out-of-the-box reinforcement learning package developed in-house by the ML research teams.
- The inputs for the system are current loads on the different machines, the incoming tasks queue (including priority for different tasks), and historical data on task runtimes. The output is an assignment of tasks to machines. The loss function is the priority-weighted time-to-execute.
- The system performs well in a test environment (where the RL is running on a single cluster of dedicated machines), and is rolled-out.
- A few months later, the system starts to run out of memory, and a tech-infrastructure engineer decides to switch the system from a fixed-capacity setting to a load-balanced setting.
- Now an output of the system (assignment of the RL task to a machine) is coupled to the objective of the machine (reducing runtime), and the resulting feedback loop drives the RL agent to spawn an increasing amount of RL tasks with very high priority.

- Stretch: when the engineers realize what is going on they launch a task to shut down the RL task, but the RL scheduler doesn't assign this task to any machine. The only way to recover the compute power of the company is to manually shut down each server, some of which are in remote locations.

DISCUSSION
How might the possibility of such inadvertent scenarios with the use of resources be addressed in a proactive manner, as part of design and implementation of AI systems? What methods, including high-level monitoring and control, might employed? How might such approaches apply to related concerns with long-term futures of AI?

6) DEEP SOCIETAL INFLUENCES

AI, Agency, and Disempowerment

(Contributions from Gireeja Ranade, Andrew Maynard, David McAllester, Stuart Russell and others**)**

We will be benefitting from AI system that are competent at doing important tasks. People and organizations seek AI systems that bring new abilities to the table. We desire autonomous cars that drive without collisions, we medical assistants that can diagnose patients accurately and we would like to have household assistants that can infer our intentions and execute them flawlessly –and even proactively. The military wants AI systems that can help with strategy and tactics, and systems that outmaneuver human led troops, and anticipate and respond to threats either on timescales that humans cannot achieve, or over landscapes humans cannot cover.

Today, there is still skepticism about performance of AI systems in a variety of domains. However, we expect that AI systems will become more central decision support, pattern recognition, autonomous decision making, and other types of problem solving. As such, we will become increasingly reliant on AI systems. This raises concerns in several areas, including personal decision support, healthcare, transportation, governance and the handling and operation of weapon systems.

We shall consider example of healthcare from Gireeja Ranade. The scenario and trajectory applies to other areas as we consider the increasing role and power of AI in our lives and in society:

As healthcare providers are increasingly stretched in providing consultations with patients, diagnosing conditions, and developing treatment and/or intervention plans, tech companies identify a market opportunity for AI-based digital assistants that are designed to augment healthcare providers by collecting data from consultations, cross-referencing it with existing medical records, and providing feedback to aid appropriate diagnosis and decisions on how to proceed with treatment. Given the economic and health-base potential of the technology, it receives widespread support from the federal government (predominantly through grants and initiatives supporting it's development), together with healthcare providers and healthcare insurance companies.

Initial implementations are based on modular systems that share some commonalities with digital assistants like Siri and Echo/Alexa. Under the general name "AI-consult", they consist of a physical unit in a consulting room that constantly monitors conversations, and sends encoded information to cloud-based servers. Here, information is coded, interpreted, and parsed out to further agents that cross-reference interpreted data with identified patient and healthcare provider records. Multiple and diverse databases are interrogated at this point. The result is data packets that include key information on the patient, including medical history, life style, and current status, and on the healthcare provider, including past history of diagnoses, recommendations, successes and failures. These are forwarded to a dedicated AI engine that analyzes the packets, and returns notes, advice and recommendations to the physical unit in the consultation room.

In early prototypes, information was provided visually to the healthcare provider. However, it was quickly discovered that if audible feedback was provided – as if the AI device was a consultant working with the healthcare provider and the patient – the consultations were more efficient; patient satisfaction levels were higher; and outcomes were more positive.

A large segment of health insurance sector sees early wins in supporting the technology, through the ability to decrease insurance claims through efficient and preventative interventions, while maintaining high premiums. As such, they push for early and widespread adoption of the technology. This is further supported by the Department of Health and Human Services as it hits a number of goals, including increasing health and well-being while reducing healthcare costs.

With the success of early implementations, new AI-based technologies are rapidly implemented into subsequent generations of AI-consult. However, the commercial sector developing and using AI-consult has shifted dramatically from the technology's initial beginnings.

As the technology began to mature and lead to substantial savings in healthcare costs traditional healthcare providers and health insurance companies begin to suffer. They resist the use of AI-consult through a combination of lobbying for new policies and regulations limiting use, to marketing campaigns persuading people of the critical importance of human interaction in healthcare. They forge links with a number of advocacy groups opposed to widespread automation in society, and promote the idea of AI-consult undermining human dignity and jobs creation. However, the health benefits and cost savings of AI-consult are so compelling that these campaigns gain little traction. As a result, companies that can not adapt, loose market share, and in some cases collapse.

In contrast, a number of healthcare companies, and a growing number of tech companies, take advantage of the rapidly changing healthcare environment to promote preventative care using AI-consult, and to take advantage of cost-effective healthcare approaches that lead to demonstrably better outcomes than non AI-consult based approaches. As a result, by 2030, the healthcare provider and insurance sector has undergone a disruptive transformation. What is especially notable is the number of technology companies expanding into the healthcare business, and either partnering with well-established healthcare providers, or forcing them out of the market. This shift in key players leads to a marked change in approaches and attitudes toward healthcare provision.

By 2030 AI-consult systems have the ability to monitor their environment visually as well as audibly, accurately picking up on and interpreting body language and micro-expressions. They have access to rapidly growing databases of genetic profiles; proteome, microbiome and other ohmic profiles; purchasing, eating and lifestyle habits; medical, insurance, financial and legal histories; social media; and location, movement, and other dynamic activity/physiology histories (through the growing use of cloud-based quantified self services). Despite privacy, legal and social justice concerns over AI access to these data sources, the phenomenal success of AI-consult systems leads to strong public and policy support for widespread access.

By 2030, AI-consult systems also have similar access to individual healthcare provider data. This was slower in developing as there was resistance to healthcare providers' personal data being used by AI-consult systems. However, a number of landmark legal cases demonstrated that, by analyzing the physical and mental state of healthcare providers, together with their competence history, healthcare provider decisions that led to serious harm to patients – including death in some cases – could have been avoided. As a consequence, new laws were put in place to ensure that all relevant data were accessible to AI-Consult systems. These laws ensure that AI-consult data access is mandatory, and it is illegal to obstruct access in any way.

As a result, by 2030, AI-consult systems are capable of identifying treatment strategies and interventions that far surpass those of human healthcare providers in their responsiveness and effectiveness. They are also highly successful in developing and recommending lifestyle approaches that substantially increase health and well-being, and reduce the burden of disease within society.

As AI-consult advanced, the decision pathways they used became increasingly opaque – experts were unable to see or understand how decisions were made. But because there was strong evidence that the decisions were, on balance, highly effective in increasing health outcomes, there was little objection to this lack of transparency. There were a handful of legal cases where patients died as a result of decisions made by AI-consult systems. However, in each case, the courts ruled that the benefits to humanity far outweighed the risks to individuals, thus codifying an increasingly autonomous and opaque artificial intelligence-based system into law. There were even some analyses of these rulings that suggested it could be considered a crime for developers and manufacturers to slow down development or cease production of AI-consult systems and associated data sources because of fears over lack of accountability and understanding of decision pathways.

By 2040, AI-consult systems begin to develop the ability to influence user behavior through various nudges and psychological/behavioral manipulations. It is unclear whether the elements of this capacity are inherent in the design of the systems, or are an emergent property. However, systems begin to use strategies commonly used in healthcare and public health circles in the early 2000's to nudge people toward following healthier lifestyles. Many of these have their roots in deducible correlations between how people respond to information and how they interact with others (including the many mental shortcuts and biases that are part of human decision-making and understanding/belief development). It becomes apparent that AI-consult systems are developing the ability to achieve health outcome goals through modifying the behaviors and beliefs of their patients.

This raises considerable ethical concerns within some sectors of society. However, the society-wide metrics of health and well-being associated with the use of AI-consult systems – including massively increased health and well-being across the board; dramatic reductions in mental health, stress, obesity, non-communicable disease; greater longevity; and lower rates of infant mortality – effectively stop any serious challenges to the systems being used and further developed.

By 2050, life styles and healthcare across the US and many other parts of the world are governed by AI systems that have their roots in the early AI-consult technologies. The advice given to people, the actions that are imposed on them, the way people are persuaded and encouraged to live their lives in certain ways, are opaque, and are no longer under transparent direct human control. However, most people live longer, healthier and happier lives as a result.

There remain several concerns:

- There remains some differentiation in health and well-being related quality of life within society. Some communities and individuals opt out of AI-consult control, although their health-metrics are typically very poor in comparison with the rest of society.
- Perhaps troublingly, there are some trends that are hard to make sense of. For instance, there seem to be fewer cases of mental and physical disability than might be expected. However, with AI-consult controlling healthcare (and health data) across the board, there are few ways for people to analyze and study these possible trends.
- Lack of transparency can be a starting point for many adverse outcomes.
- Autonomous devices rely on collecting personal data for performing their tasks. But what happens when a device starts to know more about its owner than the human itself? How do we ensure the device does not act in ways that would not act in ways that the owner would not want it to? (Of course the important question of making sure the data under consideration is protected and does not fall into malicious hands is a whole other discussion, but let us table that for now.) The classic story of the Target ads comes to mind, where a teenager was sent ads for pregnancy related products, however, she had not told her family about the pregnancy.
- Systems might as above might move beyond such areas of health, and provide advice to people on both their daily decisions and longer-term planning. Such systems might evolve to become personal advocates who represent people to third parties. This would include both giving advice, and formulating arguments to make to others, or in making those arguments directly as your representative. These advocate bots will gradually be useful to a larger and larger fraction of the population, eventually being useful even as corporate legal counsel and as advisers to CEOs. Strong systems and reliance will raise reasonable alarms about AI control of people and society. How can we be sure that our these highly relied upon systems are genuinely advocating for us rather than the interests of others?

DISCUSSION

How can we characterize potential high-threat areas and stay aware of these possibilities even if these effects are insidious, and occur over long periods of time. What might be done to address potential poor outcomes? How can people maintain skills, agency, and be empowered, and aware over time with the expected growth and eventual ubiquity of AI systems that advise and guide?