Caldwell *et al. Crime Sci* (2020) 9:14
https://doi.org/10.1186/s40163-020-00123-8

Crime Science

RESEARCH

Open Access

AI-enabled future crime

M. Caldwell, J. T. A. Andrews, T. Tanay and L. D. Griffin*

Abstract

A review was conducted to identify possible applications of artificial intelligence and related technologies in the perpetration of crime. The collected examples were used to devise an approximate taxonomy of criminal applications for the purpose of assessing their relative threat levels. The exercise culminated in a 2-day workshop on 'AI & Future Crime' with representatives from academia, police, defence, government and the private sector. The workshop remit was (i) to catalogue potential criminal and terror threats arising from increasing adoption and power of artificial intelligence, and (ii) to rank these threats in terms of expected victim harm, criminal profit, criminal achievability and difficulty of defeat. Eighteen categories of threat were identified and rated. Five of the six highest-rated had a broad societal impact, such as those involving AI-generated fake content, or could operate at scale through use of AI automation; the sixth was abuse of driverless vehicle technology for terrorist attack.

Introduction

Technologies based on artificial intelligence (AI) and machine learning (ML) have seen dramatic increases in capability, accessibility and widespread deployment in recent years, and their growth shows no sign of abating. While the most visible AI technology is marketed as such (e.g. 'personal assistants' such as Amazon Alexa, Apple Siri and Google Home), learning-based methods are employed behind the scenes much more widely. From route-finding to language translation, biometric identification to political campaigning, and industrial process management to food supply logistics, AI saturates the modern connected world at many levels (Benaich and Hogarth 2019).

The wide range of legitimate AI applications includes systems for crime prevention and detection (Dilek et al. 2015; Li et al. 2010; Lin et al. 2017; McClendon and Meghanathan 2015), but the technology also has potential for misuse in the service of criminal activities (Kaloudi and Li 2020; Sharif et al. 2016; Mielke and Chen 2007; van der Wagen and Pieters 2015). As AI technology expands in capability and deployment, so do the risks of criminal exploitation. Opportunities for AI-enabled crime exist both in the specifically computational domain (overlapping with traditional notions of cybersecurity) and also in the wider world. Some of these threats arise as an extension of existing criminal activities, while others may be novel. To adequately prepare for and defend against possible AI threats, it is necessary to identify what those threats might be, and in what ways they are likely to impact our lives and societies.

There have been a number of recent efforts to identify and classify potential threats from AI-assisted crime. Brundage et al. (2018) provide a useful overview that concentrates on the near term (up to 5 years) and make several strategic policy recommendations. In particular, they emphasize the importance of widening and deepening collaboration between the broad range of stakeholders on both the policymaking and technology sides: policy cannot be devised without full information, nor imposed without consent; conversely, AI research and product development must take into account the wider social environment in which it occurs and take responsibility for its consequences. Wilner (2018) assesses contemporary cybersecurity risks with particular reference to the increasing connectedness of everyday devices (the 'Internet of Things'). King et al. (2019) undertake a systematic literature review to identify threats and broad areas for further research. Peters (2019) presents four

*Correspondence: L.Griffin@cs.ucl.ac.uk
COMPASS Group, Computer Science, UCL, London, UK

fictionalized threat scenarios in a novelistic style before discussing possible counter-strategies. The speculative nature of such exercises means that no single set of 'correct' answers can be expected to emerge, and the existence of each should be seen as complementary to the others rather than reducing their utility.

Here we report on a scoping project concerned with the crime and security threats associated with AI—we report on the project as a whole but in particular a workshop event which represented the culmination of the project. The workshop involved the completion of a threat assessment exercise by a diverse group of stakeholders from security, academia, public policy and the private sector. The meeting encompassed multiple viewpoints and fields of expertise, and functioned in part as a knowledge sharing exercise, but the principal aim was to identify vulnerabilities by imagining possible crimes, and then assessing the severity of the threats posed by each. Attention was focused on the short to medium term (up to 15 years into the future), though there was also some discussion of prospects beyond that horizon. The consideration of a relatively broad view of criminality was encouraged: it was assumed that laws could adapt to changing circumstances, so threats were not required to be currently illegal to be considered "crimes". The development of strategies for combating the proposed crimes was *not* an explicit goal of this exercise, although these were discussed in the context of threat severity.

Literature review

In the preparatory review phase of the project, examples were collected of existing or predicted interactions between AI and crime, with both terms interpreted quite broadly. Cases were drawn from the academic literature, but also from news and current affairs, and even from fiction and popular culture, which can be considered as a barometer of contemporary concerns and anxieties.

Examples were organised into three non-exclusive categories according to the relationship between crime and AI:

- Defeat to AI—e.g., breaking into devices secured by facial recognition.
- AI to prevent crime—e.g., spotting fraudulent trading on financial markets.
- AI to commit crime—e.g., blackmailing people with "deepfake" video.

They were also tagged with one or more loose taxonomic classes reflecting the technologies or vulnerabilities involved:

- Adversarial perturbations
- Autonomous vehicles
- Fake content
- Automated snooping
- Robotics
- Biometrics
- Precognition
- Anomaly detection
- Computer Science not AI
- Automated software
- Cryptography
- AI blowback.

These categories were eventually refined and merged to form the basis for the workshop sessions discussed in the "Workshop description" section below. The categorised examples were collated into a deck of PowerPoint slides and made available to participants in the workshop.

Workshop description

The two-day residential workshop on 'AI & Future Crime' was held in February 2019, organized by the authors of this report and the Dawes Centre for Future Crime at UCL. The workshop was attended by 31 delegates invited because of interest and expertise in the theme. Delegates were from a variety of backgrounds: 14 from academia (Security and Crime Science, Computer Science, Public Policy; with attendees from UCL, Bristol and Sheffield), 7 from the private sector (AI technology, finance, retail; including iProov, Synerise and the British Retail Consortium), and 10 from the public sector (defence, police, government; including the Home Office, Defence Science and Technology Laboratory, National Crime Agency, National Cyber Security Centre and the College of Policing). The majority of attendees were UK-based, with 3 private sector delegates from Poland.

The workshop was organized into five sessions and included presentations delivered by the organizing team, presentations by attendees, and small group working. In an introductory session a general briefing about AI progress and its sub-types was delivered, along with instructions about the main sessions that followed. Each main session focussed on a different theme for AI-enabled crime:

- Patterns & Predictions
- Fake Content
- Snooping, Biometrics & beating AI
- Autonomy: Vehicles, Robots & Software

Sessions were led by a briefing on the theme, covering the state of the relevant technology and outlining examples of potential crimes of that type identified by the organizing team in preparatory research and discussion

(see the "Literature review" section). After each briefing the delegates worked in small teams to rate possible AI-enabled crimes, focussing on those relevant to the session theme.

Threat rating

Participants were divided into groups of 4–6 such that each contained a balance of diverse backgrounds from academia, government and industry. Working in these groups, delegates rated possible AI-enabled future crimes devised by the organizing team along with any additional crimes generated within the groups.

Crimes were assessed along four dimensions that had been identified by the organizing team during the review phase as useful for understanding different aspects of threat severity, as follows:

Harm: Victim and/or social harm, with terror considered as a form of harm. For example, financial loss to an individual or undermining trust in public institutions would both be considered harmful.

Criminal Profit: Realization of a criminal aim e.g. financial return, terror, harm, or reputational damage. This may be symmetric to harm but need not be: a crime might aim to make a monetary profit but destroy a person's reputation as a side effect.

Achievability: How feasible the crime would be, accounting for readiness of the necessary technology, its availability, and practical complexities of data collection or engineering required. For example, does it depend on very expensive hardware or access to classified information?

Defeatability: Measures to prevent, detect or render unprofitable. Consideration to be given to whether defeat measures are unobvious; simple or complex; and/or needing behavioural change. For example, could the crime be circumvented unobtrusively by a company such as Google or does it require every computer user in the world to be equipped with a biometric scanner?

These dimensions are not strictly independent, but they require distinct consideration: a crime might be easy for a criminal to commit but unprofitable, or hard to defeat but also hard to achieve. Groups were therefore instructed to rate each dimension separately; so, for example, harm was to be assessed assuming that the crime could be achieved, rather than lowered if it was unlikely to be. Each group worked independently.

For rating, delegates used a simplified form of q-sorting (Stephenson 1953), as illustrated in Fig. 1. Q-sorting allows users to rank items so that the assigned ranks have a roughly bell-curve distribution. It is easier, quicker and more effective than linear sorting as it prioritizes identifying the outlier items and reduces time spent on unreliable comparison of middle ground items.

Each group worked with a separate q-sorting grid for each dimension. Each grid was formed of 16 place-holding squares (10 cm × 10 cm) arranged in approximate bell-curve form with columns of height 1-2-3-4-3-2-1. Crimes were summarized by a short phrase (e.g. "AI snake oil") written on a sticky note the size of a placeholder. Starting with empty grids delegates gradually populated them with notes for crimes described in session briefings or suggested by the small groups. Delegates were encouraged to re-arrange the grids as new crimes were introduced. For all grids, columns to the left were considered the less bad end of the spectrum (e.g. low harm or easy defeat), and columns to the right were considered the worse end of the spectrum (e.g. great profit or easy achievability). Delegates reported that q-sorting was a helpful approach to the comparison of crimes. Sufficient time was allocated to the q-sorting process so that delegates were able to discuss each possible crime, contributing their individual perspectives, to arrive at a considered joint assessment.

AI & future crime briefing

The workshop briefings were divided across multiple sessions in a discursive format with a degree of duplication and reinforcement. Here we attempt to streamline the main points into a single briefing to provide background for the subsequent discussion of the different categories of crime that were discussed and rated. In the following sections we outline some of the technological background and consider aspects of criminal potential.

Technology and terminology

Artificial intelligence (AI)

An umbrella term covering a wide range of efforts to computationally reproduce complex human capabilities such as language use, vision and autonomous action. Most AI endeavours are discrete and solve specific, well-defined tasks. The term "artificial general intelligence" is used for the idea of a single integrated system capable of tackling many different kinds of task at once, even those it has never encountered before. Though common in popular culture, at present this remains a distant prospect, whereas more specific AI applications, such as machine vision and natural language processing, are increasingly prevalent.

Machine learning (ML)

A subset of AI whose methods are based on discovering patterns in data ("bottom-up") rather than overtly specifying how to do a task ("top-down"). ML methods are primarily statistical and are often both computationally challenging and data intensive. Because the decision criteria are *learned*, rather than purposefully designed, they

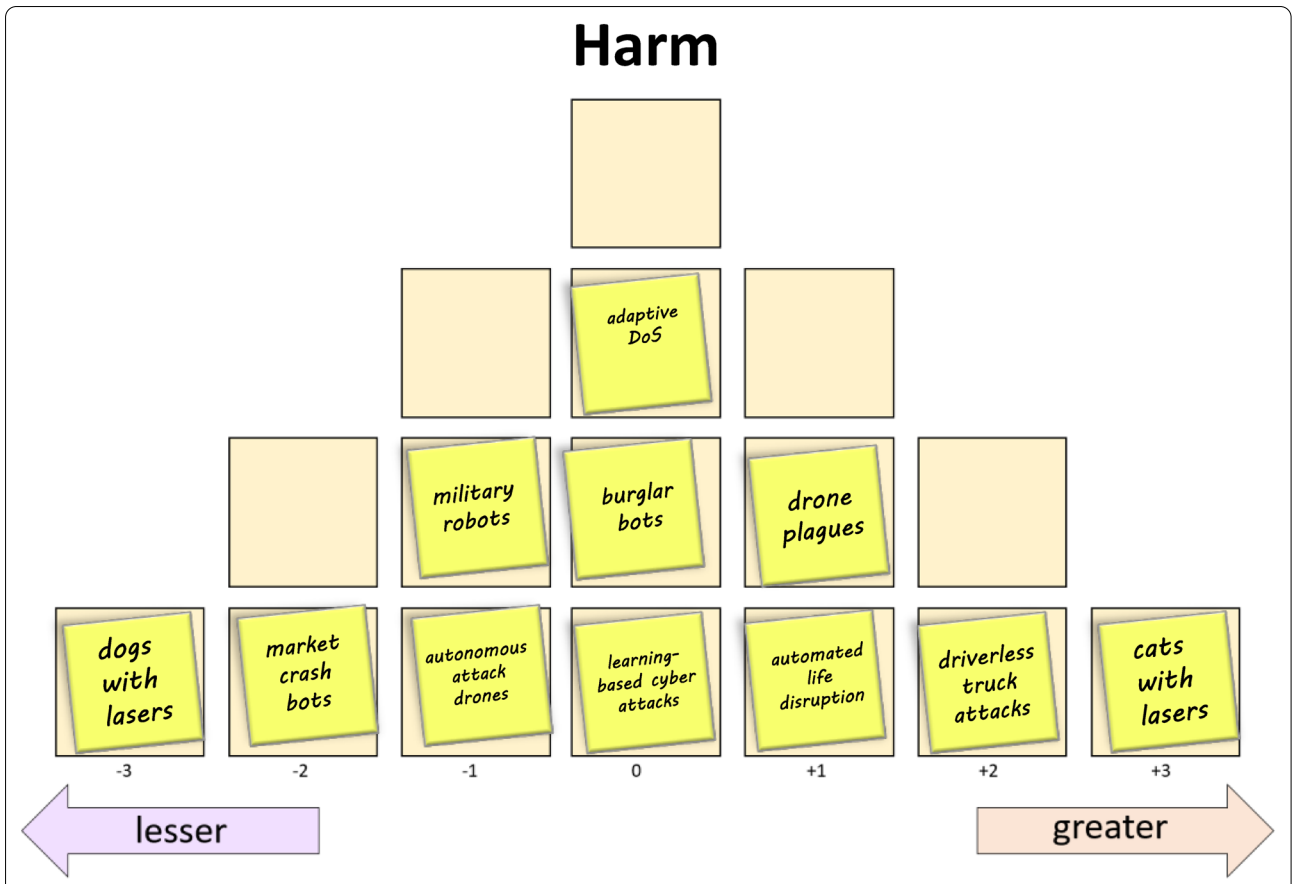

Fig. 1 Schematic of the use of q-sorting for ranking crimes for harm. (Scores shown are illustrative only, not actual rankings from any workshop group.)

can often be difficult to understand, especially when the models are complex. This has led to criticism of ML systems as opaque "black boxes" as well as active research into making them interpretable and explainable (Gilpin et al. 2018). ML has been the dominant AI paradigm of recent years, and most of the prominent AI successes have been learning-based.

Supervised learning

ML in which the system is explicitly told the correct answers (usually called "labels") during learning. For example, an image classifier may be given photos of cats and dogs with labels identifying which is which, in order to learn how to distinguish the two.

Unsupervised learning

ML in which the system is *not* explicitly told the correct answers but looks for potentially interesting patterns or clusters in the unlabelled data. 'Self-supervised learning' is gaining popularity as an alternative designator. Current methods for self-supervised learning are much less effective than supervised learning but are improving.

Training data

The body of data from which an ML system learns its task, as distinct from the *test* data to which it will eventually be applied. In order to be useful, ML systems must be able to *generalise* from the training data to new, previously unseen examples; failure to do so is known as "overfitting". Training data availability is often the biggest limiting factor for ML, particularly if supervised learning is used which requires the data be labelled.

Deep learning

ML that makes use of very large and complex models known as "neural networks". Loosely based on biological structures that perform information processing in our own brains, neural networks consist of many simple computational units (termed neurons) connected together. Each neuron integrates inputs from its connected neighbours to produce an output which is sent on to other neurons in the network. The learning power

of a neural network comes from its size and connectivity. While simple neural networks have been researched for many years, high capacity "deep" networks have only recently become practical thanks to dramatic increases in computational power and data availability.

Reinforcement learning (RL)

An iterative, exploratory ML approach in which an autonomous software agent observes the state of some system and can perform actions on it, with the goal of maximising a "reward" (e.g. the score in a game). Reinforcement learning is especially relevant for dynamic problems such as how a robot should interact with its environment. Because of the need for very many repeated (and potentially failed) tests, RL is most easily applied to situations which can be computationally simulated. RL is probably the closest ML comes to our everyday understanding of what "learning" means.

Active learning

A variant of supervised and reinforcement learning where the AI system iteratively guides what observations join the training dataset pool, and so achieves good performance from much lower amounts of data. For example, an AI could discover what wording of an advert generates the most click-throughs by 'experimenting' on consumers in a closed-loop manner, rather than just learning from a survey of responses to a pre-determined catalogue.

Natural language processing (NLP)

AI applied to understanding, generating or translating human languages in textual or spoken forms. Early NLP approaches were rule-based, using explicit syntactic and grammatical models, but as in many other AI domains, statistical and learning-based approaches are now predominant.

Bias

Incorrect learning, especially due to unrepresentative training data. ML will only learn patterns that are present in the training data, and these may include cues that are not aligned with the problem specification. In practice, complex real world data *is* typically imperfectly sampled or affected by systemic, social, economic and historical biases (Shankar et al. 2017), and these can be reflected in the learned behaviour (Buolamwini and Gebru 2018).

Generative adversarial network (GAN)

An ML configuration in which two distinct ML systems *compete* in order to improve the learning of a task. One, the *generator*, learns how to produce some kind of output data, while the other, the *discriminator*, learns to find flaws in the generator's output (Goodfellow et al. 2014). As each side improves, the other is forced to "raise its game" to compensate, and this is facilitated by the discriminator cueing the generator to the flaws it discovers. GANs are largely responsible for recent increases in the ability to generate convincing artificial ("deepfake") content.

Adversarial perturbation

A (usually small) change to the input data to a trained ML system that causes the system to produce the wrong output (Szegedy et al. 2014). Many—perhaps *all*—ML systems have complex decision boundaries that can be vulnerable to carefully-crafted counter-examples, despite working correctly for naturally-occurring data. The perturbation is often imperceptible to a human observer, involving coordinated small changes across a very large number of input dimensions. When this phenomenon was first discovered it was thought that exploiting the vulnerability would require detailed access to the inner workings of the trained system, but more transferable exploits have been found, including some that can be used in the real world (Evtimov et al. 2017; Brown et al. 2017; Athalye et al. 2017). It has even been claimed that human perception is vulnerable to adversarial perturbations (Elsayed et al. 2018), suggesting the problem may be *universal*: we can attempt to mitigate it, but might not ever be able to eliminate it entirely.

Criminal potential

AI can be involved in crimes in a variety of ways. Most obviously, AI could be employed as a *tool* for crime, making use of its capabilities to facilitate actions against real world targets: predicting the behaviour of people or institutions in order to discover and exploit vulnerabilities; generating fake content for use in blackmail or to sully reputations; performing feats that human perpetrators are unable or unwilling to do themselves, for reasons of danger, physical size, speed of reaction and so on. Although the methods are new, the crimes themselves may be of traditional type—theft, extortion, intimidation, terror.

Alternatively, AI systems may themselves be the *target* of a criminal activity: circumventing protective systems that present obstacles to a crime; evading detection or prosecution for crimes already committed; making trusted or critical systems fail or behave erratically in order to cause damage or undermine public trust.

AI could also simply provide *context* for a crime. Fraudulent activities might depend on the victim believing some AI functionality is possible, even though it is not—or possible, but not actually used in the fraud.

Of course, these categories are not mutually exclusive. As in the old adage about catching a thief, an attack on an AI system may itself require an AI system to enact. The fraudulent simulation of non-existent AI capabilities might be executed using other AI methods that *do* exist.

Crimes vary enormously. They may target individuals or institutions, businesses or customers, property, government, the social fabric, public discourse. They may be motivated by financial profit, acquisition of power or change of status relative to others. They may build or damage reputations or relationships, shift policy or sow discord; such effects may be an end in themselves or stepping stones to some further goal. They may be committed in mitigation of or attempt to avoid punishment for other crimes. They may be driven by a desire for revenge or sexual gratification or the furtherance of religious or political ends. They might express nothing more than a nihilistic urge to destruction, vandalism, or violence for its own sake.

The degree to which this diversity of criminal acts may be enhanced by use of AI depends significantly on how embedded they are in a computational environment: robotics is rapidly advancing, but AI is better suited to participate in a bank fraud than a pub brawl. This preference for the digital rather than the physical world is a weak defence though as contemporary society is profoundly dependent on complex computational networks, not just for finance and commerce but also all forms of communication, politics, news, work and social relationships. People now conduct large parts of their lives online, get most of their information there, and their online activity can make and break their reputations. This tendency is likely to continue for the foreseeable future. Such an online environment, where data is property and information power, is ideally suited for exploitation by AI-based criminal activity which can have substantial real world consequences. Moreover, unlike many traditional crimes, crimes in the digital realm are often highly *replicable*: once developed, techniques can be shared, repeated, even sold, allowing the potential for *marketisation* of criminal techniques or provision of "crime as a service". This may lead to a lowering of technological barriers as criminals are able to outsource the more challenging aspects of their AI-based crime.

Ratings analysis

Following the workshop, the q-sorting grids were compiled to a spreadsheet by scoring the columns from −3 to +3, reading left to right. Within and across grids, crimes that were considered by the authors to be sufficiently close (e.g. 'drone swarms' and 'autonomous attack drones') were aggregated and the mean rank recorded. There was insufficient data to support detailed analysis of the variation between the six groups of delegates performing ranking, so the analysis focused on simple cross-group averages with between-group interquartile range reported as an indicator of consistency. Averaged ranks were standardized to zero mean and unit standard deviation within each dimension. The final result was 18 crimes each with a rating value in four dimensions. For comparison purposes, a combined threat severity score for each crime was computed as min[achievability + defeatability, harm + profit]. The logic of this measure is that achievability + defeatability represents a measure of likelihood that the crime could be successfully carried out, while harm + profit represents an assessment of its negative impact. By taking the minimum of these two scores we identified crimes that were likely and of negative impact. To avoid spurious precision, the crimes were ranked into 8 levels, with a roughly bell curve distribution. Figure 2 shows the final results.

Crimes

In 4.1 we give an expanded discussion of the six crimes that our ratings analysis identified as overall of greatest concern. Where we report views of delegates these are not based on a systematic record of discussions, only on the impressions of the organizing team. In 4.2 and 4.3 we briefly describe the lower rated crimes.

High

Audio/video impersonation

Humans have a strong tendency to believe their own eyes and ears, so audio and video evidence has traditionally been given a great deal of credence (and often legal force), despite the long history of photographic trickery. But recent developments in deep learning, in particular using GANs (see above), have significantly increased the scope for the generation of fake content. Convincing impersonations of targets following a fixed script can already be fabricated, and interactive impersonations are expected to follow. Delegates envisaged a diverse range of criminal applications for such "deepfake" technology to exploit people's implicit trust in these media, including: impersonation of children to elderly parents over video calls to gain access to funds; usage over the phone to request access to secure systems; and fake video of public figures speaking or acting reprehensibly in order to manipulate support. Audio/video impersonation was ranked as the overall most-concerning type of crime out of all those considered, scoring highly on all four dimensions. Defeat was considered difficult: researchers have demonstrated some success in algorithmic detection of impersonation (Güera and Delp 2018), but this may not be possible in the longer term and there are very many uncontrolled routes through which fake material can propagate.

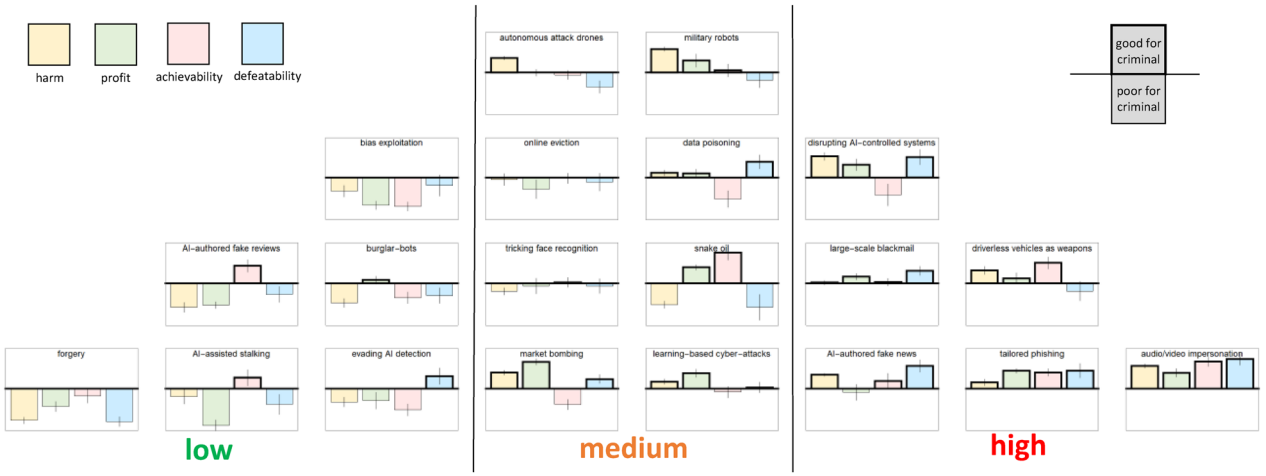

Fig. 2 Overall crime ranking resulting from the workshop. For each crime, the colour-coded bars show the mean rank for the four dimensions, computed across the 5 groups and standardised to have zero mean and unit standard deviation within each dimension. Bars above (below) the line indicate that the crime is of greater (less) concern in that dimension. Error bars indicate the interquartile range between groups. The column (low, medium, or high) each crime is placed in is determined by min(a + d,h + p). The resulting values have an approximately bell-shaped distribution and are arranged in a similar fashion to the q-sorting grids. Crimes within the same column should be considered of comparable concern. Concern increases with column from left to right

Changes in citizen behaviour might therefore be the only effective defence. These behavioural shifts, such as generally distrusting visual evidence, could be considered an indirect societal harm arising from the crime, in addition to direct harms such as fraud or reputation damage. If even a small fraction of visual evidence is proven to be convincing fakes, it becomes much easier to discredit *genuine* evidence, undermining criminal investigation and the credibility of political and social institutions that rely on trustworthy communications. Such tendencies are already apparent in the discourse around "Fake News". Profit was rated the least high dimension for this crime, not because the investment required is high (it is not) but because impersonation crimes aimed at acquisition will likely be easiest against individuals, rather than institutions, while impersonation crimes against society will have an uncertain effect.

Driverless vehicles as weapons

Motor vehicles have long been used both as a delivery mechanism for explosives and as kinetic weapons of terror in their own right, with the latter increasing in prevalence in recent years. Vehicles are much more readily available in most countries than firearms and explosives, and vehicular attacks can be undertaken with relatively low organisational overhead by fragmentary, quasi-autonomous or "lone actor" terrorists such as those claiming affiliation with ISIS. The tactic gained particular prominence following a series of attacks in Western cities including Nice (2016), Berlin (2016), London (2017), Barcelona (2017) and New York (2017). While fully autonomous AI-controlled driverless vehicles are not yet available, numerous car manufacturers and technology companies are racing to create them, with some permitted trials on public roads. More limited self-driving capabilities such as assisted parking and lane guidance are already deployed. Autonomous vehicles would potentially allow expansion of vehicular terrorism by reducing the need for driver recruitment, enabling single perpetrators to perform multiple attacks, even coordinating large numbers of vehicles at once. Driverless cars are certain to include extensive safety systems, which would need to be overridden, so driverless attacks will have a higher barrier to entry than at present, requiring technological skill and organisation. Nevertheless, delegates rated these attacks as highly achievable and harmful, and moderately profitable (given terror as the goal). However, they scored low on defeatability (meaning relatively easy), since they are expected to be susceptible to the same countermeasures (barriers, traffic restrictions) that are already in use for vehicles with drivers.

Tailored phishing

Phishing is a "social engineering" attack that aims to collect secure information or install malware via a digital message purporting to be from a trusted party such as the user's bank. The attacker exploits the existing trust to persuade the user to perform actions they might otherwise be wary of, such as revealing passwords or clicking on dubious links (Boddy 2018). While some attacks

may be carefully targeted to specific individuals, known as "spear-phishing", this is not very scalable. At present most phishing attacks are relatively indiscriminate, using generic messages styled after major brands or topical events that can be expected to be of interest to some fraction of users purely by chance (Vergelis et al. 2019). The attacker relies on the ease of sending huge numbers of digital messages to convert a low response rate into a profitable return. AI has the potential to improve the success rates of phishing attacks by crafting messages that appear more genuine, by (for example) including information gleaned from social networks, or by faking the style of a trusted party. Rather than sending uniform messages to all targets, likely to miss the mark in most cases, the messages could instead be tailored to prey on the specific vulnerabilities inferred for each individual, effectively automating the spear-phishing approach. Additionally, AI methods could use active learning to discover "what works", varying the details of messages to gather data on how to maximise responses (Bahnsen et al. 2018). Since the criminal aim of phishing attacks is most often financial, the crime was rated as having only marginally above average harm potential, but was rated high for profit, achievability and defeatability (meaning it would be difficult to stop).

Disrupting AI-controlled systems

As the use of AI increases across government, business and home, and the roles performed by AI systems become ever more essential, the opportunities for attack will proliferate. Learning based systems are often deployed for efficiency and convenience rather than robustness, and may not be recognised a priori as critical infrastructure. Delegates could foresee many criminal and terror scenarios arising from targeted disruption of such systems, from causing widespread power failures to traffic gridlock and breakdown of food logistics. Systems with responsibility for any aspects of public safety and security are likely to become key targets, as are those overseeing financial transactions. The profit and harm ratings were accordingly high, as was defeatability. In general, the more complex a control system is, the more difficult it can be to defend completely. The phenomenon of adversarial perturbations underlines this problem, suggesting that sufficiently advanced AIs may be inherently vulnerable to carefully tailored attacks. However, achievability was rated lower, on the basis that such attacks typically require detailed knowledge of, or even access to, the systems involved, which may be difficult to obtain.

Large-scale blackmail

Traditional blackmail involves extortion under the threat of exposure of evidence of criminality or wrongdoing, or embarrassing personal information. A limiting factor in traditional blackmail is the acquisition of such evidence: the crime is only worthwhile if the victim will pay more to suppress the evidence than it costs to acquire. AI can be used to do this on a much larger scale, harvesting information (which need not itself constitute damning evidence) from social media or large personal datasets such as email logs, browser history, hard drive or phone contents, then identifying specific vulnerabilities for a large number of potential targets and tailoring threat messages to each. AI could also be used to generate fake evidence, e.g. when the information discovered implies a vulnerability without providing *prima facie* proof (Peters 2019). Large scale blackmail was rated high for profit: as with phishing, economies of scale mean the attack may only require a low hit rate to be profitable. Defeatability was considered difficult, largely for the same reason it is problematic in traditional cases: reluctance of the victim to come forward and face exposure. However, harm was rated only average, since the crime is by nature primarily directed at individuals, and achievability is also relatively low due to the high data requirements and combination of multiple different AI techniques that must be coordinated. It is worth noting that a very crude non-AI blackmail analogue is common among current phishing methods. Termed "sextortion", it involves falsely claiming to have compromising video footage from the user's hacked computer or phone, in the hope that some percentage of recipients will guiltily panic and pay up rather than call the blackmailer's bluff (Vergelis et al. 2019). As with all such scams, it is impossible to know what the hit rate is, but we suspect it is rather low.

AI-authored fake news

Fake news is propaganda that aims at credibility by being, or appearing to be, issued from a trusted source. In addition to delivering false information, fake news in sufficient quantity can displace attention away from true information. Delegates considered the possibility of fake news content being generated by AI technology to achieve greater efficiency, presence or specificity. AI could be used to generate many versions of a particular content, apparently from multiple sources, to boost its visibility and credibility; and to choose content or its presentation, on a personalized basis, to boost impact. The crime scored above average for harm, achievability and defeatability, and below average for profit. Harm was considered high because of the considerable *potential* to influence specific political events, for example voting (whether or not this has already been done); and because of diffuse societal effects if the communication of real news is undermined or displaced by fake media. High achievability was underlined by a breaking news

story (Hern 2019) that emerged during the workshop. Defeat was considered difficult as a strictly technical problem, and because the boundary between fake and real news is vague. To date, the most successful attempts at combatting fake news have been via education, notably in Finland (Mackintosh and Kiernan 2019). The lower profit score reflected the difficulty of making financial profit from fake news (although there is scope for using fake news in market manipulation (Kamps and Kleinberg 2018)), and because of the uncertain effect of its more diffuse consequences.

Medium

Military robots

As with many fields of technological development, the military have a significant stake in robotics research, with potentially very different goals than civilian users despite many methodological overlaps. Any availability of military hardware (e.g. firearms or explosives) to criminal or terrorist organisations can be expected to pose a serious threat, and this would certainly be the case for autonomous robots intended for battlefield or defensive deployment. Delegates rated such access as potentially both very harmful and profitable. However, it was also recognised that ratings were necessarily speculative. Military capabilities tend to be shrouded in secrecy, and we have very limited knowledge as to the current state of the art and rate of advancement.

Snake oil

Sale of fraudulent services under the guise of AI or using a smokescreen of ML jargon. Such fraud is extremely achievable, with almost no technical barrier (since by definition the technology doesn't work). Potential profits are high: there are plenty of notorious historical examples of con men selling expensive technological trumpery to large organisations, including national governments and the military (Gilsinan 2016). Arguably this is not a *use* of AI for crime, but the crime depends on the target believing in the claimed AI capabilities, which in turn depends on AI being perceived as successful by the public. It should be potentially easy to defeat via education and due diligence, though there is a current window of opportunity open until those measures have effect.

Data poisoning

The manipulation of ML training data to deliberately introduce specific biases, either as an end in itself (with the goal of damaging commercial rivals, distorting political discourse or sowing public distrust) or with the intention of subsequent exploitation. For example, making an automated X-ray threat detector insensitive to weapons you want to smuggle aboard a plane, or encouraging an investment advisor to make unexpected recommendations that shift market value in ways of which you will have prior knowledge that you can exploit. The more widely used and trusted the data source, the more damaging this could be. Though potentially harmful and profitable, this was rated low on achievability, since trusted data sources tend to be hard to change and (as a corollary of being widely used) under frequent scrutiny.

Learning-based cyber attacks

Existing cyberattacks tend either to be sophisticated and tailored to a particular target (Kushner 2013) or crude but heavily automated, relying on the sheer weight of numbers (e.g. distributed denial of service attacks, port scanning). AI raises the possibility of attacks which are both specific and massive, using, for example, approaches from reinforcement learning to probe the weaknesses of many systems in parallel before launching multiple attacks simultaneously. Such attacks were considered harmful and profitable, though delegates were less certain of their achievability.

Autonomous attack drones

Non-autonomous radio controlled drones are already used for crimes such as smuggling drugs into prisons (BBC News 2018) and have also been responsible for major transport disruptions (Weaver et al. 2018). Autonomous drones under onboard AI control potentially allow for greater coordination and complexity of attacks while freeing the perpetrator of the need to be within transmitter range of the drone, making neutralization and apprehension more difficult (Peters 2019). At present, drones are not typically used for crimes of violence, but their mass and kinetic energy is potentially dangerous if well-targeted (e.g. into aircraft engines) and they could also be equipped with weaponry. Drones could be particularly threatening if acting *en masse* in self-organizing swarms. They were rated highly for potential harms, but low for defeatability, since in many contexts protection may be provided using physical barriers.

Online eviction

The primacy of online activities within modern life, for finance, employment, social activity and citizenship, presents a novel target for attacks against the person: denial of access to what have become essential services is potentially debilitating. This could be used as an extortion threat, to damage or disenfranchise groups of users, or to cause chaos. Some existing phishing and cyberattacks attempt something similar by means such as "ransomware", and quasi-organised groups of human actors sometimes engage in activities such as mass misreporting of abuse on social media, but AI could enable attacks that

are both more subtle—carefully tailoring forged activity to violate terms of service, identifying specific points of vulnerability for each individual—and more scalable. Eviction was considered likely to be unprofitable in its own right and more of a concern as an adjunct to other threats.

Tricking face recognition

AI systems that perform face recognition are increasingly used for proof of identity on devices such as smartphones, and are also in testing by police and security services for tasks such as suspect tracking in public spaces and to speed up passenger checks at international borders. These systems could present an attractive target for criminals. Some successful attacks have been demonstrated (Sharif et al. 2016), including "morphing" attacks that enable a single photographic ID, such as a passport, to pass as (and be used by) multiple individuals (Robertson et al. 2017; Andrews et al. 2019). Profits and harms were considered below average, since attacks are most likely to enable relatively small-scale crimes.

Market bombing

The manipulation of financial or stock markets via targeted, probably high frequency, patterns of trades, in order to damage competitors, currencies or the economic system as a whole (rather than directly to profit from the trading, although that could also be a side effect) was discussed. The idea is an AI boosted version of the fictional Kholstomer cold war plot (Trahair 2004), which envisaged a Russian attempt to precipitate a financial crash by suddenly selling huge stockpiles of US currency via front companies. Reinforcement learning was suggested as a method for discovering effective trading strategies, possibly allied with NLP-based media analysis and fake content generation. Achievability was rated low, because of the extreme difficulty of accurately simulating market behaviour and the very high cost of entry to engage in large scale trading, but potential harms and profits were correspondingly high.

Low

Bias exploitation

Discovering and taking advantage of (existing) learned biases in widely-used or influential algorithms. For example, gaming YouTube recommendations to funnel viewers to propaganda, or Google rankings to raise the profile of products or denigrate competitors. In practice such behaviour is already widespread, often not illegal (though it may be against the provider's terms of service) and is even (in the form of *search engine optimisation* or SEO) taken as a legitimate (if shady) online business model. It is likely to be easier to employ and harder to counter when AI-assisted.

Burglar bots

Small autonomous robots that could be delivered into premises through small access points, such as letterboxes or cat flaps, to retrieve keys or to open doors allowing ingress for human burglars. The technical requirements are highly constrained, which should make these more achievable than more ambitious classes of autonomous robots. But harms and profits are low, because they enable only very localised small-scale crimes, and they are relatively defeatable by simple physical means such as letterbox cages.

Evading AI detection

Policing and security is expected to rely increasingly on AI-based triage and automation to deal with the ever-growing volumes of data gathered by investigation. Attacks which undermine those processes in order to erase evidence or otherwise thwart discovery are likely to become increasingly attractive to criminals (Bonettini et al. 2019). Adversarial perturbations (e.g. used to conceal pornographic material from automated detection) offer one possible route to doing so, although the requirements for system knowledge may be prohibitive. Harms and profits were rated low, in part because the nature and context of the "crime" were insufficiently defined and delegates were not persuaded it was achievable. However, if it *were* achieved, defeatability was rated difficult, since the crime is by definition about "getting away with it".

AI-authored fake reviews

Automatic content generation for sites such as Amazon or TripAdvisor to give a false impression of a product or service and drive customers either towards or away from it. Such fakery is already performed by human agents. AI could increase efficiency but profits and harms from individual campaigns of this kind are likely to remain small-scale and localised.

AI-assisted stalking

Use of learning systems to monitor the location and activity of an individual through social media or personal device data. Also considered to encompass other crimes around coercive relationships, domestic abuse, gaslighting etc., and to relate to a current news story concerning the complicity of Western technology companies in the provision of apps for enforcing social norms in repressive societies (Hubbard 2019). Harms were rated as low, not because these crimes are not extremely damaging, but

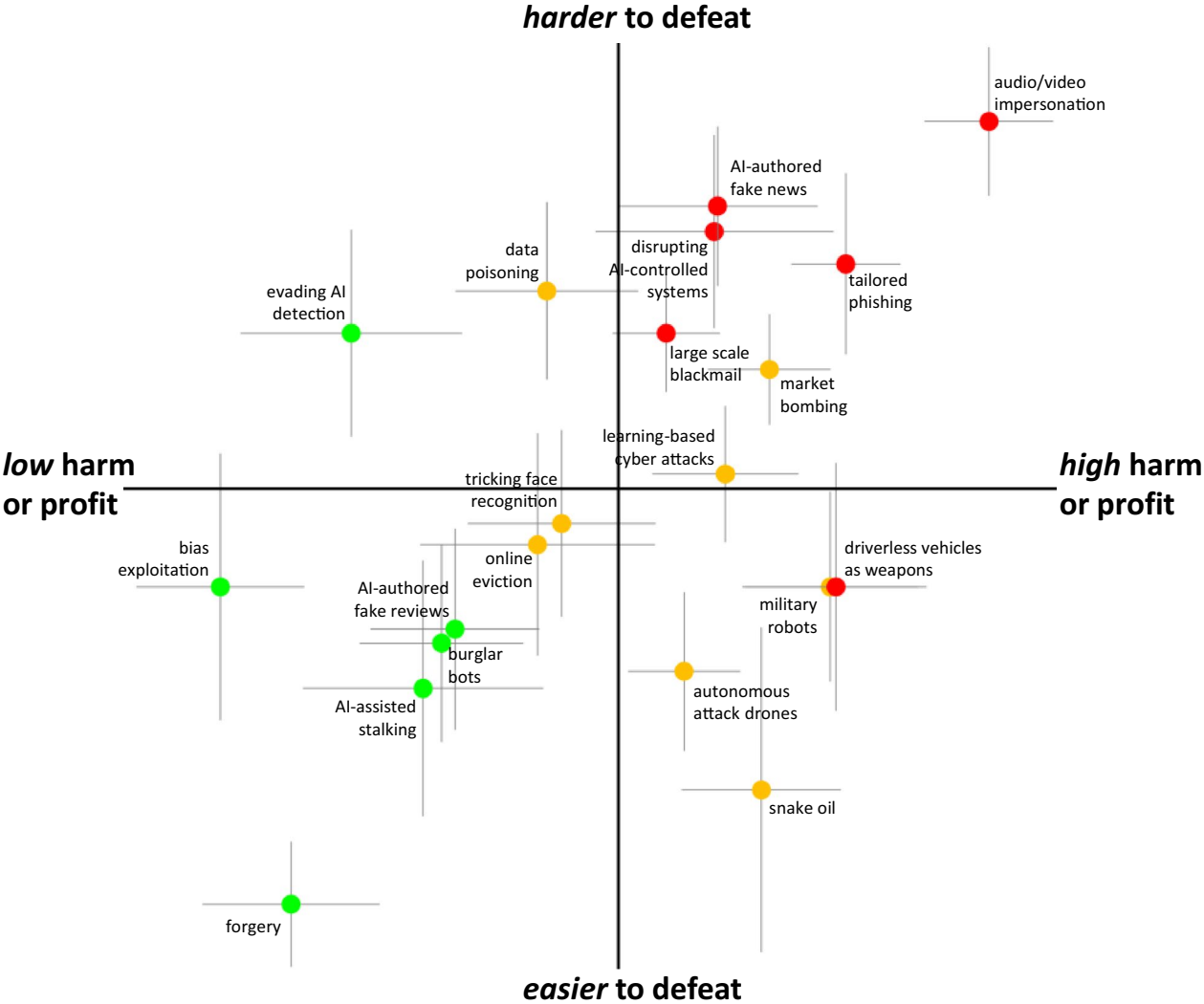

Fig. 3 Difficulty of defeat relative to the harmfulness or profitability of the crime. The most concerning crimes are those in the upper right quadrant, being both very harmful and hard to defeat. Crimes in the lower right quadrant potentially offer the strongest potential for intervention, being both harmful and defeatable. (Error bars indicate the between-group interquartile range for the ratings.)

because they are inherently focused on single individuals, with no meaningful scope for operating at scale.

Forgery

Generation of fake content, such as art or music, that can be sold under false pretences as to its authorship. This was rated as the least concerning threat of all those considered, both in terms of harms and likeliness to succeed. AI capabilities here remain strictly limited: while there has been some success producing digital images that broadly mimic the visual style of great painters, that is a very different proposition from creating actual physical objects that would pass muster in a gallery or auction house. The art world has had to deal with forgeries for centuries and has extensive (if not always sufficient) defensive practices in place. AI doesn't even attempt to address most of those obstacles.

Conclusions

Results from a futures exercise such as this are necessarily speculative and reflect the range of knowledge, experience and priorities of the delegates. Nevertheless, the outcomes provide a useful snapshot of prevailing concerns and how these are expected to play out in the years ahead.

In this case, the pattern of ratings suggests that delegates were particularly concerned about *scalable* threats, with crimes involving severe harm to single individuals typically rated lower than (possibly lesser, or ill-defined) harms to large numbers of victims, whole social classes or society at large. Group discussions and the rankings that emerged from them were clearly shaped at least in part by contemporary discourse and current events, including anxieties about electoral interference, the spread of disinformation in a rapidly-changing media landscape, and potential erosion of the norms of Western liberal democracy. It is possible that some of these concerns may turn out to be transient or parochial within the time scales under discussion. Even so, realistically, this is very much the environment within which current AI developments must be understood, and from which future crimes can be expected to evolve.

Although the development of strategies for threat mitigation was explicitly not a goal of this exercise, it is worth considering how the results could be used by stakeholders to inform their own responses to the potential crimes that were identified and discussed. One possible approach would be to look at the trade-off between harm and defeatability as a guide to where effort and expenditure might efficiently be targeted. Such a comparison is illustrated in Fig. 3. Most of the crimes of highest concern appear in the upper right quadrant, being not only very harmful but also very difficult to counter—indeed this combination is largely why these threats are so concerning. Conversely, crimes in the lower left quadrant are easier to defeat but also relatively harmless. By these criteria, the most promising targets for intervention would probably be those crimes in the lower right quadrant, in that they are very harmful but relatively defeatable. Such considerations might usefully be taken into account in determining where to allocate resources.

Methodologically, there was some uncertainty around the definitions of both the proposed crimes and the dimensions of assessment. The scale and specificity of crimes varied significantly, which made comparisons difficult, though this was mitigated to some extent by the coarse stratification of the q-sorting. Rating dimensions were imprecise and open to different interpretation in different contexts: e.g. the notions of 'profit' and 'harm' associated with large scale acts of terrorism are not really commensurate with those for, say, embezzlement. In the subsequent analysis, it was unclear how the four dimensions used should be weighted relative to each other and the approach ultimately used should be taken as only broadly indicative. We do believe that our process provides a worthwhile perspective on the relative severity of different plausible threats, which can help to inform future policy decision making.

Abbreviations

AI: Artificial intelligence; GAN: Generative adversarial network; ML: Machine learning; NLP: Natural language processing; RL: Reinforcement learning.

Acknowledgements

This research was funded by a grant from the Dawes Centre for Future Crime at UCL.

Authors' contributions

JTAA, TT and LDG performed the literature review and organised the sandpit. LDG designed the Q-sorting analysis. MC and LDG wrote the manuscript. All authors identified threats, conducted the sandpit, collated ratings. All authors read and approved the final manuscript.

Funding

Dawes Foundation.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 22 January 2020 Accepted: 14 July 2020
Published online: 05 August 2020

References

Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2019). Multiple-identity image attacks against face-based identity verification. *arXiv.org*, vol. cs.CV (pp. 1–13).

Athalye, A., Engstrom, L., Ilyas, A., Kwok, K. (2017). Synthesizing robust adversarial examples. *arXiv.org*, vol. cs.CV.

Bahnsen, A. C., Torroledo, I., Camacho, L. D., & Villegas, S. (2018). DeepPhish: Simulating malicious AI. In *Presented at the APWG Symposium on Electronic Crime Research* (pp. 1–9).

BBC News. (2018). Gang who flew drones carrying drugs into prisons jailed. *BBC News*. https://www.bbc.co.uk/news/uk-england-45980560.

Benaich, N., & Hogarth, I. (2019). State of AI report 2019. *SlideShare*, 25-Jun-2019. Retrieved July 3, 2019 from, https://www.slideshare.net/StateofAIReport/state-of-ai-report-2019-151804430.

Boddy, M. (2018). Phishing 2.0: the new evolution in cybercrime. *Computer Fraud & Security Bulletin, 2018*(11), 8–10.

Bonettini, N., Güera, D., Bondi, L., Bestagini, P., Delp, E. J., & Tubaro, S. (2019). Image anonymization detection with deep handcrafted features (pp. 1–5).

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. *arXiv.org*, vol. cs.CV.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv.org*, vol. cs.AI.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research* (vol. 81, pp. 1–15).

Dilek, S., Cakır, H., & Aydın, M. (2015). Applications of artificial intelligence techniques to combating cyber crimes: A review. *IJAIA, 6*(1), 21–39.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I. J., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both human and computer vision. arXiv:1802.08195 [cs.LG].

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A, Rahmati, A., Song, D. (2017). Robust physical-world attacks on deep learning models. *arXiv*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *arXiv.org*, vol. cs.AI.

Gilsinan, K. (2016). Why is Iraq still using fake bomb detectors? *The Atlantic*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.
Güera, D., Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *Presented at the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*.
Hern, A. (2019). New AI fake text generator may be too dangerous to release, say creators. *The Guardian*.
Hubbard, B. (2019). Apple and Google urged to dump Saudi app that lets men track women. *New York Times*.
Kaloudi, N., & Li, J. (2020). The AI-based cyber threat landscape. *ACM Computing Surveys, 53*(1), 1–34.
Kamps, J., & Kleinberg, B. (2018). To the moon: Defining and detecting cryptocurrency pump-and-dumps. *Crime Science, 7*(1), 18.
King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2019). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics, 9979*(2), 159.
Kushner, D. (2013). The real story of Stuxnet. *IEEE Spectrum*.
Li, S.-T., Kuo, S.-C., & Tsai, F.-C. (2010). An intelligent decision-support model using FSOM and rule extraction for crime prevention. *Expert Systems with Applications, 37*(10), 7108–7119.
Lin, Y. L., Chen, T. Y., & Yu, L. C. (2017). Using machine learning to assist crime prevention. In *Presented at the 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 1029–1030).
Mackintosh, E., & Kiernan, E. (2019). Finland is winning the war on fake news. What it's learned may be crucial to Western democracy," *CNN*, May-2019. Retrieved July 08, 2019, from, https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/.
McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *MLAIJ, 2*(1), 1–12.
Mielke, C. J., & Chen, H. (2007). Botnets, and the cybercriminal underground. In *Presented at the 2008 IEEE International Conference on Intelligence and Security Informatics (ISI 2008)* (pp. 206–211).
Peters, K. M. (2019). *21st century crime: How malicious artificial intelligence will impact homeland security*. California: Monterey.
Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2017). Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PLoS ONE, 12*(3), e0173319–12.
Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv.org*, vol. stat.ML.
Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime. In *Presented at the the 2016 ACM SIGSAC Conference* (pp. 1528–1540). New York, New York, USA.
Stephenson, W. (1953). *The study of behavior; Q-technique and its methodology*. Chicago: University of Chicago Press.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv*.
Trahair, R. C. (2004). *Encyclopedia of Cold War Espionage, spies, and secret operations*. Westport CT: Greenwood Press.
van der Wagen, W., & Pieters, W. (2015). From cybercrime to cyborg crime: botnets as hybrid criminal actor-networks. *CRIMIN, 55*(3), 578–595.
Vergelis, M., Shcherbakova, T., Sidorina, T. (2019). Spam and Phishing in 2018. *securelist.com*, 12-Mar-2019. Retrieved June 13, 2019, from, https://securelist.com/spam-and-phishing-in-2018/89701/.
Weaver, M., Gayle, D., Greenfield, P., & Perraudin, F. (2018). Military called into help with Gatwick drone crisis. *The Guardian*.
Wilner, A. S. (2018). Cybersecurity and its discontents: Artificial intelligence, the Internet of Things, and digital misinformation. *International Journal, 73*(2), 308–316.

Publisher's Note