INTELIGÊNCIA ARTIFICIAL E CRIME: UM GUIA PARA CRIMINÓLOGOS

ARTIFICIAL INTELLIGENCE AND CRIME: A PRIMER FOR CRIMINOLOGISTS

Keith J Hayward[1]

Matthijs M Maas[2]

Resumo[1]**:** Este Artigo introduz o conceito de Inteligência Artificial (IA) para os estudiosos da Criminologia. Fazendo uma revisão geral deste fenômeno (incluindo breves explicações de áreas importantes relacionadas como "aprendizado de máquina", "aprendizado profundo" e "aprendizado por reforço"), o estudo aborda a possível utilização da IA pelos criminosos, incluindo o que aqui chamamos de "crimes com IA", "crimes na IA" e "crimes pela IA". Nas seções do Artigo, objetiva-se destacar o potencial da IA como fenômeno criminógeno, tanto em termos de ampliação de crimes já existentes quanto de facilitação de novas transgressões no meio digital. Na parte final do texto, voltamos nossa atenção para as principais formas pelas quais o paradigma da IA está transformando as práticas policiais, tanto de vigilância social como de justiça criminal, através de modalidades de monitoramento difuso baseadas em previsão e prevenção. Ao longo do Artigo, apresentamos uma variedade de exemplos programáticos que, em conjunto, esperamos que sirvam como um guia útil para criminólogos interessados no "nexo tecnologia-crime".

Palavras-chave: Inteligência Artificial. Criminologia dos Megadados. Crimes Cibernéticos. Criminologia Digital. Monitoramento Onipresente. Aprendizado de Máquina. Tecnologia e Crime.

Abstract: This article introduces the concept of Artificial Intelligence (AI) to a criminological audience. After a general review of the phenomenon (including brief explanations of important cognate fields such as "machine learning", "deep learning", and "reinforcement learning"), the paper then turns to the potential application of AI by criminals, including what we term here "crimes with AI", "crimes against AI", and "crimes by AI". In these sections, our aim is to highlight AI's potential as a criminogenic phenomenon, both in terms of scaling up existing crimes and facilitating new digital transgressions. In the third part of the article, we turn our attention to the main ways the AI paradigm is transforming policing, surveillance, and criminal justice practices via diffuse monitoring modalities based on prediction and prevention. Throughout the paper, we deploy an array of programmatic examples which, collectively, we hope will serve as a useful AI primer for criminologists interested in the "tech-crime nexus".

[1] *Professor of Criminology, Copenhagen Centre for Criminology, University of Copenhagen, Denmark*.

[2] *Senior Research Fellow (Law & AI), Institute for Law & AI*.

[1] Artigo originalmente publicado em 2021, em *Crime, Media, Culture*, Vol. 17(2), 209-233. Os Autores desejam esclarecer que este texto foi escrito originalmente em 2020, e, portanto, é anterior aos muitos e rápidos avanços tecnológicos que ocorreram no campo da Inteligência Artificial nos últimos cinco anos, bem como seus impactos na disciplina da Criminologia. Tradução e Notas do Tradutor (NT) de Artur de Brito Gueiros Souza, com o auxílio de ferramentas de inteligência artificial.

Keywords: Artificial Intelligence. Big Data criminology. Cybercrime. Digital Criminology. Inescapable Surveillance. Machine Learning. Technology and Crime.

1. INTRODUÇÃO

Em pé atrás de uma mesa no tipo de palco gigantesco de conferências preferido por executivos da Microsoft e por futurologistas de tecnologia com *headset*, o cientista da computação Zeyu Jin se dirige ao público presente ao evento *Adobe MAX 2016* em San Diego, Califórnia. Jin é um dos designers por trás do "*Project Voco*", um protótipo de software de edição e geração de áudio que permite aos usuários fabricar som digital. Apelidado de "*Photoshop para voz*", o *Voco* está na vanguarda de um conjunto de tecnologias que, basicamente, permite que qualquer pessoa com um *laptop* forje ou manipule gravações de voz. Uma atmosfera silenciosa se instala enquanto o software é exibido em uma tela de vídeo do tamanho de um anúncio de outdoor... O silêncio não dura muito.

Depois que Jin demonstra a facilidade com que consegue "fabricar" a voz do comediante americano Keegan Michael-Key, o público de blogueiros e empreendedores emergentes explode com uma fervora salva de palmas. Jin, empolgado com a resposta, ergue os braços em retribuição aos aplausos. No entanto, mesmo antes das palmas diminuírem, uma nota de advertência é soada pelo parceiro de comédia de Key, Jordan Peele, que diz em tom jocoso: "Você pode se meter em grandes problemas por algo assim". Jin não se abala. "Não se preocupe", ele responde, acrescentando: "Na verdade, nós pesquisamos como evitar falsificações... não é para coisas más".[2]

Não é para coisas más. A resposta dada por Jin não é nada surpreendente. Na verdade, por décadas, os cientistas da computação ficaram tão seduzidos pelo potencial ilimitado das novas tecnologias, que os efeitos negativos decorrentes desses sistemas foram sendo minimizados ou, muitas vezes, completamente ignorados.[3] Conhecida como "tecno-otimismo", essa falha em efetivamente equilibrar recompensa-e-risco ficou ilustrada com a famosa frase "*Não seja mau*" — um antigo lema do código de conduta corporativo do *Google*. Contudo, assim como o *Google* discretamente abandonou aquele lema, depois das polêmicas envolvendo

[2] http://www.youtube.com/watch?v=QUK6rEUZAcA.

[3] É claro que essa não é uma "posição universal". No mesmo período, um grupo de cientistas da computação — especialistas em segurança cibernética — muito fez para promover uma "mentalidade de segurança" mais cautelosa". (SCHNEIER, 2008).

a empresa com casos de censura, evasão fiscal e escândalos de violação de privacidade, os desenvolvedores de tecnologia passaram a perceber que os sistemas de Inteligência Artificial (IA) não só permitem práticas policiais inaceitáveis, mas também aportam novas formas nocivas de atividades criminosas.[4]

Efetivamente, se muitos cientistas da computação são compreensivelmente "culpados" ao concentrar seus esforços apenas no lado tecnológico do que aqui poderíamos chamar de "*nexo tecnologia-crime*", então o mesmo pode ser dito, na direção contrária, para os *criminólogos*. **NT1** De fato, enquanto a nossa disciplina continua a evoluir no conhecimento sobre crime e punição, ela o faz em grande parte desconectada dos muitos desafios sociais que estão sendo impostos pela gigantesca disrupção tecnológica (BROWN, 2006; HAYWARD, 2012; HOLT e BOSSLER, 2014).[5] A maioria dos criminólogos não apenas ignora questões relacionadas à tecnologia em si, como, com poucas exceções, demonstra um deliberado desprezo pelas teorias de outras disciplinas que buscam abrir um espaço para o diálogo entre as ciências sociais e as tecnologias da informação e comunicação. Como resultado, embora nos últimos anos estudiosos tenham começado a pesquisar e alertar sobre os potenciais "usos maliciosos" da IA (BRUNDAGE *et al*., 2018) e suas aplicações ofensivas e defensivas (BROADHURST *et al.*, 2018), criminólogos tradicionais ainda não desempenharam nenhum papel significativo no que KING *et al.* (2019) descreveram, de maneira pertinente, como o emergente campo interdisciplinar do "*IA-Crime*" (IAC).[6]

---

[4] Mesmo quando se age com cautela, isso não necessariamente garante o apoio unânime de uma comunidade tecnológica que segue uma cultura de código aberto [*open-source culture*] estabelecida há muito tempo. Em 2019, a organização sem fins lucrativos *OpenAI*, apoiada por Elon Musk, desenvolveu o "GPT-2", um modelo de linguagem capaz de compor prosa coerente (incluindo comunicados de imprensa) a partir de apenas duas frases de contexto (demonstração disponível em https://transformer.huggingface.co/). Inicialmente, a *OpenAI* divulgou publicamente apenas versões menores do seu sistema, expressando preocupação de que a ferramenta completa pudesse ser usada por agentes mal-intencionados "para gerar linguagem enganosa, tendenciosa ou abusiva em grande escala." (RADFORD *et al*., 2019). Esta decisão desencadeou um debate contínuo — e, por vezes, divisivo — na comunidade de IA sobre quando, se é que alguma vez, é apropriado reter a investigação em IA (LEIBOWICZ *et al*., 2019), com outro laboratório a lançar, posteriormente, a sua própria IA de "notícias falsas neurais", para ajudar os investigadores a identificar notícias falsas (ZELLERS *et al*., 2019). Por fim, em novembro de 2019, a *OpenAI* divulgou o modelo completo, juntamente com algumas reflexões sobre futuras estratégias de divulgação responsável para aplicações de IA (SOLAYMAN *et al*., 2019).

NT1 Optou-se por traduzir a expressão *criminologist* por "criminólogo" ao invés de "criminologista", tendo em vista o seu uso mais frequente na criminologia latino-americana. Neste sentido: SHECAIRA, SÉRGIO SALOMÃO. *Criminologia.* 10ª ed. São Paulo: Revista dos Tribunais, 2022.

[5] Definimos "tecnologia" aqui como sendo qualquer combinação de ferramentas, aptidões, processos e técnicas através dos quais a capacidade humana é ampliada (BENNETT MOSES, 2007: 592).

[6] KING *et al*. (2009:2) descrevem a IAC como "uma área relativamente jovem e inerentemente interdisciplinar — abrangendo estudos sociojurídicos e das ciências formais".

Felizmente, essa situação está mudando. Recentemente, a criminologia foi embalada por uma série de novos grupos de pesquisadores preocupados em como o crime está sendo transformado pelo impacto do que GREENFIELD (2017) chama de "novas tecnologias radicais da era da network". Aqui, referimo-nos a novos subcampos do saber, como *criminologia digital* (POWELL *et al*., 2018); *criminologia computacional* (WILLIAMS e BURNAP, 2016) e *criminologia da Big Data* (CHAN e BENNETT MOSES, 2017; SMITH *et al*., 2017); e um corpo crescente de pesquisas sobre *tecnocrimes* (STEINMETZ e NOBLES, 2017), envolvendo criptografia, criptomoedas, comércio ilícito e "*dropgangs*" [entregadores de drogas] na *Dark web*, além de "*stalkerware*" [software espião] (*e.g.,* ALDRIDGE, 2019; KRUITHOF *et al*., 2016; MUNKSGAARD *et al*., 2016; PAOLI *et al*., 2017; PARSONS *et al*., 2019). A recente publicação do impressionante e muito necessário "*Routledge Handbook of Technology, Crime and Justice*", de MCGUIRE e HOLT (2017), evidencia ainda mais o crescente interesse da criminologia nas questões tecnológicas.[7]

Esperamos que este artigo dê ainda mais consistência a esse corpo de produção científica, direcionando a atenção dos estudiosos e pesquisadores para questões criminológicas relacionadas a um aspecto muito particular da tecnologia contemporânea: a IA.

Com efeito, para observadores leigos, a IA pode ser um conceito difícil de compreender — um fenômeno que parece estar em toda parte, mas, ao mesmo tempo, parece estranhamente opaco. Na cultura popular e nas reportagens da imprensa, a IA costuma ser tema de narrativas fantasiosas sobre "robôs assassinos" ou "sistemas distópicos de vigilância". Contudo, no cotidiano das pessoas, a IA tende a funcionar em um nível muito mais mundano, impulsionando desde Smart-TVs até aplicativos de tradução de idiomas. Talvez seja essa universalidade que confunda as pessoas, ao menos porque cada futuro imaginativo de IA evoca seu próprio conjunto particular de preocupações sobre segurança, ética, legalidade e responsabilidade. Para que a sociedade possa superar essa confusão, o que se faz necessário são respostas claras para perguntas simples, tais como: "O que é exatamente a IA?". "Quais são suas capacidades e limites?". E, mais importante para criminólogos: "Quais são as consequências da sua proliferação e utilização na sociedade, seja como instrumento para fins delituosos ou ilegítimos ou como mecanismo de segurança e controle social?"

[7] Naturalmente, reconhecemos o trabalho pioneiro sobre *cibercrime* realizado por SHELIA BROWN, YVONNE JEWKES, DAVID WALL, MAJID YAR e outros. Contudo, neste momento, gostaríamos de enfatizar a diferença fundamental entre as primeiras pesquisas sobre "crimes online" e o mundo potencialmente mais vasto da criminalidade associada às novas tecnologias radicais [disruptivas].

Para responder a essas perguntas, o presente estudo será desenvolvido em três partes. Começaremos com uma breve e acessível introdução à IA. Como não escrevemos para cientistas da computação, evitaremos entrar em detalhes técnicos e questões computacionais, oferecendo, ao invés disso, uma visão geral mais ampla, voltada especificamente para criminólogos e outros cientistas sociais eventualmente interessados. Na Parte 2, abordaremos as possíveis utilizações da IA por criminosos, classificando-as em "crimes com IA", "crimes na IA" e "crimes pela IA". A Parte 3 tratará da utilização das tecnologias de IA pelas forças policiais, agências de justiça criminal e governos, bem como de que forma a ampliação da vigilância granular [detalhada], onipresente e preditiva redefinirá a cultura e a configuração futura dos ambientes urbanos.

Importa destacar que essas três áreas não esgotam o espaço para futuras investigações criminológicas em relação à IA. Por exemplo, uma área adicional que merece ênfase diz respeito às potenciais inovações *metodológicas* dos sistemas de IA para auxiliar nos estudos dos fenômenos criminológicos. Isso, sem dúvida, oferecerá uma grande oportunidade para criminólogos, mas também exigirá uma cuidadosa análise crítica. No entanto, considerando que o escopo deste Artigo é lançar as bases para o estreitamento do engajamento entre a criminologia e a IA, nossa preocupação aqui não se estende aos métodos digitais do porvir.

Em termos amplos, nossa visão crítica é melhor descrita como dicotômica. Por um lado, nossa justificativa para escrever este Artigo é tranquilizar os leitores sobre a IA, desmistificar o seu conceito e sugerir que, como criminólogos, temos muito a oferecer a esse novo domínio tecnológico. Por outro lado, o estudo também procura deixar claro que, se aplicados de forma maliciosa ou sem a devida diligência [*due diligence*], a utilização generalizada da IA pode causar danos indescritíveis [*unfathomable*]. Portanto, embora não seja nossa intenção assustar o leitor sobre a IA, fazemos este alerta em parte para chamar a atenção para algumas das tendências mais preocupantes da IA. Para alcançar esses objetivos, o Artigo toma por base em uma série de exemplos programáticos que esperamos que possam servir como um guia útil para os criminólogos interessados.

2. O QUE É INTELIGÊNCIA ARTIFICIAL?

"De forma geral, o maior perigo da Inteligência Artificial", brincou o teórico de IA ELIEZER YUDKOWSKY (2008), "é que as pessoas acreditam rápido demais que a compreendem" (p. 308). De fato, segundo a jornalista KELSEY PIPER (2018), "A conversa sobre IA é cheia de

confusão, desinformação e de pessoas falando umas das outras — em grande parte porque usamos o acrônimo 'IA' para nos referir a muitas coisas". Em um esforço para superar a confusão decorrente da proliferação semântica, esta seção objetiva esclarecer a terminologia-chave da IA e algumas das principais funções e limitações dos atuais sistemas de IA.

2.1 NOÇÕES CONCORRENTES DE "INTELIGÊNCIA" ARTIFICIAL

"Inteligência" dentro do paradigma da IA é um conceito filosófico e científico complexo e profundamente contestado. LEGG e HUTTER (2007a), por exemplo, contabilizam mais de 70 definições. Uma forma de superar essa confusão conceitual é aderir à influente estrutura diádica de STUART RUSSELL e PETER NORVIG (2009: 5), que gira em torno de duas questões filosóficas inter-relacionadas:

1. *O que buscamos criar em IA*? A manifestação de certos processos internos de pensamento ou, simplesmente, comportamentos/resultados externos específicos?
2. *Como medimos o desempenho de um sistema de IA*? O objetivo é simplesmente reproduzir/imitar seres humanos ou superá-los, alcançando desempenho "ótimo" com relação a certos resultados?

Essa dicotomia vale a pena ser explorada, ao menos porque bem ilustra como o pensamento sobre a IA se desenvolveu ao longo do tempo.

A primeira pergunta basicamente orienta a maior parte dos comentários gerais sobre se os robôs poderiam ou não possuir o "correto padrão" de elaboração de pensamentos ("senciência"/"consciência").[8] Contudo, atualmente a maior parte das pesquisas sobre IA se afastou das tentativas de reproduzir noções internas de inteligência, focando em critérios mais externos (mensuráveis). De fato, essa foi a questão central no frequentemente mal compreendido "Teste de Turing", que, afinal de contas, foi concebido como um "Jogo de

[8] Por ex., em 2011, um pequeno robô chamado *Qbo* passou no "teste do espelho", exclamando: "*Oh! Este sou eu. Legal!*", ao ser confrontado com o seu reflexo (ACKERMAN, 2011). Essas experiências são frequentemente anunciadas na mídia como evidência de robôs "autoconscientes". Mas, para a maioria dos pesquisadores, esses testes revelam mais sobre as deficiências dos exames psicológicos do que o fornecimento de qualquer evidência efetiva de IA "senciente".

Imitação" (TURING, 1950: 460), e não como um teste comprovativo da senciência metafísica dos computadores[9]. **NT2**

Isso leva à segunda pergunta sobre medir o desempenho da IA. Enquanto pesquisas anteriores eram guiadas quase exclusivamente pela definição humano-cêntrica — e esse entendimento ainda alimenta algumas análises socio-científicas adjacentes de IA da atualidade[10] — grande parte da ciência da computação contemporânea considera a "humanidade" da IA como algo irrelevante ou, no máximo, de valor simbólico. Afinal, o valor de um algoritmo de negociação de ações de alta frequência não é que ele possa jogar conversa fora com os colegas no cafezinho, mas, sim, que ele possa ser eficaz na negociação de derivativos. E, ainda assim, o ponto central não é que o algoritmo tenha *exatamente* o mesmo desempenho de um *trader* humano, mas que ele possa superar a performance humana nessa tarefa.

De fato, em muitos casos marcantes recentemente registrados, o desempenho sobre-humano dos sistemas de IA resultou até em estratégias explicitamente não-humanas que nunca passariam no "Teste de Turing". Foi o que ocorreu com o famoso "Movimento 37" da *AlphaGo*, inicialmente incompreensível, mas que acabou sendo decisivo para vencer a partida disputada contra Lee Sedol [renomado jogador profissional sul-coreano de *Go*], ou das estratégias de

---

[9] O teste de Turing foi superado em 2014 por "Eugene Goostman", um *chatbot* que convenceu 33% dos juízes de que era humano ao se passar por um estudante ucraniano.

NT2 De acordo com JACOB TURNER, em um artigo seminal de 1950, Alan Turing questionou se as máquinas poderiam pensar como os humanos. Ele então sugeriu uma experiência chamada "Jogo da Imitação". Neste exercício, em salas separadas, um juiz deve tentar identificar qual entre dois jogadores é um homem e uma mulher, usando apenas perguntas e respostas datilografadas. Se um dos jogadores conseguir ludibriar o juiz, achando que o homem é a mulher, vence o jogo. Turing propôs uma variação do jogo na qual a máquina de IA substitui o homem. Se a máquina conseguir persuadir o juiz não apenas de que é humana, mas também de que é a jogadora feminina, então ela demonstrou inteligência. As versões modernas desse jogo simplificaram a tarefa, pedindo a um programa de computador e a várias pessoas, em salas isoladas, que mantivessem uma conversa digitada de cinco minutos com um painel de juízes em uma sala diferente. Os fiscais têm de decidir se a entidade com a qual estão trocando mensagens é humana ou não; se o computador conseguir enganar uma proporção suficiente de juízes (em geral em torno de 30%), então ele venceu o jogo. Segundo Turner, o grande problema do "Jogo da Imitação" de Turing é que ele testa apenas a capacidade de imitar um humano em uma conversa digitada, mas uma imitação habilidosa *não* equivale a inteligência. De fato, em alguns dos testes mais "bem-sucedidos" de programas concebidos para ganhar o jogo, os programadores criam estratégias para que a máquina exponha as fragilidades que tendemos a associar aos humanos, tais como erros ortográficos. Outra tática adotada pelos programadores nos modernos "Testes de Turing" é usar respostas humorísticas pré-definidas para desviar a atenção diante da falta de respostas convincentes dos seus computadores às perguntas dos juízes. (TURNER, Jacob. *Robot Rules. Regulating Artificial Intelligence.* London: Palgrave MacMillian, 2019, pp. 10-11).

[10] A definição humano-cêntrica é utilizada no estudo da IAC de KING *et al*. (2019). Contudo, consideramos essa definição demasiado restrita; embora os sistemas autônomos "semelhantes aos humanos" possam colocar desafios peculiares, muitas das utilizações mais potentes da IA em termos criminais — ou policiais — envolvem concepções mais amplas de inteligência não-humana e não-antropomórfica.

xadrez "alienígenas" da *AlphaZero* (KNIGHT, 2017). Na IA moderna, na maioria das vezes não se trata de uma questão de "senciência", mas sim da capacidade, não de imitar com precisão a performance humana, mas de superá-la em vários domínios.

2.2 TERMINOLOGIA MODERNA DA INTELIGÊNCIA ARTIFICIAL

Historicamente, houve uma variedade de abordagens distintas ou "tribos" em IA (DOMIGOS, 2015), muitas das quais se inspiraram cultural e intelectualmente em áreas como a lógica, a biologia, a estatística ou a psicologia. Um tipo de sistema de IA que vem sendo usado há décadas é o "IA Simbólico", que executa tarefas seguindo um conjunto de regras explícitas e lógicas "*se-então*" [*if-then rules*] ou "declarações condicionais". Por exemplo, o piloto automático de uma aeronave manterá o avião dentro de faixas seguras pré-definidas em termos de altitude/velocidade. Essas regras foram pré-programadas por especialistas humanos com base no conhecimento por eles dominado. A IA simbólica sustenta esses chamados "sistemas especialistas" [*expert systems*], que são tão amplamente utilizados que muitas vezes nem pensamos neles como IA (SCHARRE, 2019).

Todavia, na última década, três desenvolvimentos — avanços em *Big Data*, potência de processamento de dados e inovações algorítmicas — levaram ao surgimento do "aprendizado de máquina" [*machine learning* ou ML], que é uma abordagem de IA mais dinâmica e menos "frágil". A ML envolve o sistema aprendendo gradualmente a si mesmo as regras "corretas" (ou "úteis") que precisa para executar tarefas de forma eficaz. O que é mais importante, ela faz isso com base em dados de treinamento, em vez de — como ocorre com "sistemas especialistas" — ter essas regras explicitamente programadas.

Um tipo específico de ML, responsável pelo atual *boom* da IA, é o "aprendizado profundo" [*deep learning* ou DL]. O DL envolve redes neurais profundas – uma técnica de IA inspirada na forma como os neurônios se comunicam entre si em cérebros biológicos. Redes neurais artificiais consistem em camadas de "neurônios" digitais interconectados, algumas das quais recebem uma "entrada" [*input*] — *e.g.,* informações sobre um determinado *pixel* em uma imagem —, outras fornecem uma "saída" [*output*] — *e.g.,* uma "classificação" da imagem. Cada neurônio monitora outros na camada anterior a ele, e somente se um número suficiente desses neurônios lhe enviar um sinal, ele então acionará neurônios específicos na camada seguinte. Após cada resposta errada/correta à amostra de dados de treinamento, o sistema altera a força das conexões entre os neurônios envolvidos. Dessa forma, ele "aprende" a codificar as

regras para realizar suas "tarefas".[11] Por exemplo, um algoritmo de reconhecimento de imagens criará agrupamentos de neurônios dedicados à detecção de conceitos cada vez mais abstratos — desde "*cor pixel*", até "*bordas e cantos*", "*formas*" (olhos, narizes etc.), além de conceitos (pessoa, cachorro etc.).

É importante ressaltar que a forma pela qual uma determinada IA é treinada depende do modelo específico de algoritmo de ML, bem como do tipo de dados usados pelos desenvolvedores. Existe uma ordem de distintas abordagens em uso na atualidade (SCHARRE e HOROWITZ, 2018). Primeiro, na *aprendizagem supervisionada*, o algoritmo utiliza dados de treinamento que já foram corretamente pré-rotulados por humanos (*e.g.,* fotos de lesões cutâneas, rotuladas digitalmente por médicos como cancerígenas ou benignas). Segundo, no *aprendizado não supervisionado*, os algoritmos podem, de forma independente, identificar padrões/correlações em dados "brutos", não rotulados. Isso é útil não apenas porque economiza o custo de compilar grandes conjuntos de dados de imagens rotuladas, mas porque permite que a IA identifique padrões que os humanos não conseguem detectar. Terceiro, no *aprendizado por reforço*, sistemas de IA "aprendem" a partir do feedback de seu ambiente (real ou simulado), descobrindo por tentativa e erro quais diferentes (combinações ou sequências) de ações permitem que eles "vençam" ou maximizem uma métrica de "pontuação" (vide KNIGHT, 2017 sobre o *AlphaZero*). Por fim, uma aplicação um tanto distinta pode ser encontrada nas redes antagonistas generativas [*Generative Adversarial Networks* ou GANs], nas quais uma rede neural treina a si própria para gerar dados falsos (imagens/sons/vídeos), melhorando iterativamente a qualidade de suas criações até que sejam tão indistinguíveis dos dados reais que possam enganar outro algoritmo de reconhecimento (regular e pré-trei+nado).

2.3 INTELIGÊNCIA ARTIFICIAL MODERNA: USOS, PRÉ-CONDIÇÕES E LIMITAÇÕES

Mas, além da questão terminológica e das perspectivas técnicas, quais são exatamente as utilizações da IA? Quais são as pré-condições para sua implantação, e quais são — se houver — as limitações e fraquezas da IA?

Aqui é útil desmembrar as distintas funções para as quais a IA pode servir (SCHARRE e HOROWITZ, 2018). Sistemas de IA podem ser usados em qualquer tarefa envolvendo

[11] Ao utilizar palavras como "aprender" ou "descobrir" no contexto dos sistemas IA, devemos enfatizar que estamos empregando esses termos em um sentido explicitamente não antropomórfico.

classificação e geração de dados, detecção de anomalias (*e.g.*, detectar transações financeiras fraudulentas ou novos *malwares*), predição (*e.g.*, taxas de reincidência para infratores), otimização de sistemas e tarefas complexas, além de operações autônomas de robôs ou plataformas ciberfísicas [*cyber-physical platforms*]. Cumpre salientar que todas essas tarefas são, via de regra, restritas a certas finalidades, uma vez que [ainda] não chegamos à chamada *Inteligência Artificial "Geral"* (LEGG e HUTTER, 2007b), capaz de superar os humanos em qualquer tarefa.[12] Contudo, embora ainda restritas, muitas dessas capacidades são úteis em diversos setores e contextos, incluindo, *inter alia*, os sistemas da saúde, administração da justiça, publicitário e gerenciamento do tráfego.

Por outro lado, isso não significa que a IA seja *sem* limites. Na verdade, existem várias pré-condições para a efetiva aplicação da IA a um determinado problema. Entre esses requisitos, figura, principalmente, o acesso a grandes (e às vezes rotulados) conjuntos de dados, além de questões pragmáticas relacionadas com *hardware*, talento humano e disponibilidade de investimentos. Entretanto, tanto as barreiras computacionais quanto às de acesso a dados estão caindo; além disso, elas não limitam a disseminação de sistemas já "treinados" — o que traz uma série de preocupações em muitos contextos criminais. Em paralelo, mesmo quando atendidas as mencionadas pré-condições, os atuais sistemas de IA ainda sofrem de um conjunto de problemas geralmente denominado de "estupidez artificial" [*artificial stupidity*] (DOMIGOS, 2015). O primeiro deles consiste no fato de que a IA costuma ser propensa a "esquecimentos catastróficos", ou seja, a incapacidade de transferir facilmente o aprendizado de um contexto para outro. Em segundo lugar, a IA é intrinsecamente suscetível a "entrada adversarial" — dados (*e.g.*, padrões visuais ou sonoros) projetados para alterar a forma como o sistema processa os estímulos, fazendo-o "alucinar" [*hallucinate*]. Em terceiro, os sistemas de IA não possuem "bom senso" e, portanto, ao contrário de algumas conotações em torno da palavra "inteligente", eles inevitavelmente sofrem do antigo inconveniente do "*Garbage-in, Garbage-out*" ou GIGO, [NT3] que, como já dissemos, pode acarretar problemas de vieses (BAROCAS e SELBST, 2016). Uma última questão deriva da imprevisibilidade [*unpredictability*] da IA autônoma que,

[12] Pesquisas demonstram que a expectativa entre os *AI-Experts* é de que essas capacidades serão atingidas nas próximas três a cinco décadas (GRACE *et al*., 2018).

[NT3] *Garbage in, garbage out* ["Entra lixo, sai lixo"], cuida-se de um axioma da informática que procura ressaltar a natureza lógica, mas não pensante, dos computadores e seus processos, em consequência da qual, se dados incorretos forem submetidos a processamento, os resultados serão dados igualmente incorretos". (cf. DICIONÁRIO TÉCNICO. *In* https://dicionariotecnico.com/traducao.php?l=pt&termo=garbage+in++garbage+out).

frequentemente, pode reagir de forma inesperada ao se deparar com situações imprevisíveis. Não é incomum, por exemplo, que um programa de IA resolva tecnicamente um problema, mas não da maneira pretendida (LEHMAN *et al*., 2018). Por exemplo, um algoritmo encarregado de aprender a andar em um ambiente simulado descobriu que, em vez de "desenvolver" pernas rudimentares, ele poderia mover-se para frente crescendo muito alto e depois caindo repetidamente. Como observado por SCHARRE (2019), "Na situação errada, sistemas de IA passam de superinteligentes a superburros num instante".

Mas, o que tudo isso tem a ver com a criminologia?

Com efeito, pelo menos desde o início deste século, cientistas, escritores de ficção científica e, eventualmente, criminólogos (MCGUIRE, 2007; ZEDNER, 2007) previram que todo o desenvolvimento tecnológico acima descrito acarretaria não apenas novas tipologias criminais, mas também novas modalidades de policiamento, punição e tomadas de decisões legais e preditivas. Neste ponto, apresenta-se uma importante indagação: em termos funcionais, o que deve ou não ser considerado IA pelos criminólogos? Por exemplo, a expressão IA diz respeito exclusivamente a robôs com aparência humana [*humanoid*] e às modernas redes neurais, ou também se estende para sistemas especialistas "simbólicos" mais antigos, regressão logística automatizada (aprendizado de máquina simples) ou até mesmo todos os processos algorítmicos ou computacionais, como aqueles que alimentam os aplicativos do seu smartphone?

Este artigo não busca determinar, de forma definitiva, os contornos do que é ou não é "IA". Em vez disso, nosso objetivo é conceber o nexo "IA-Crime" em um contexto mais abrangente e em rápido desenvolvimento, envolvendo uma ampla gama de práticas criminais, policiais e de segurança, que, agora, utilizam as tecnologias do espectro acima assinalado. Vale registrar que nos baseamos nas pesquisas mais atuais — e em uma seleção de recentes incidentes — para oferecer nossa própria visão esquemática do impacto da IA no fenômeno do crime e no sistema de justiça criminal[13].

Em outras palavras: agora que estamos no futuro, como ele se parece?

[13] Esclarecemos que alguns dos exemplos a seguir selecionados provêm de laboratórios de pesquisa privados e, dessa forma, podem existir certos interesses corporativos em exagerar a sofisticação ou o desempenho de seus sistemas. Portanto, é recomendável tratar algumas das afirmações adiante apresentadas com certa cautela.

3. USOS CRIMINAIS DA IA

Em recente artigo de pesquisa reflexivo, KING *et al*. (2019: 9-18) identificaram uma série de ameaças representadas pela *IA-Crime* (IAC), incluindo tráfico de drogas, crimes sexuais, roubo, fraude e falsificação. Embora o estudo deles seja um ponto de partida útil, estruturamos nossa tipologia de forma diferente — não em termos de dispositivo legal transgredido, mas sim em como os criminosos podem usar a IA. Nestes termos, apresentamos uma classificação com três categorias: (1) crimes *com* IA, (2) crimes *na* IA; e (3) crimes *pela* IA.[14]

3.1 CRIMES *COM* INTELIGÊNCIA ARTIFICIAL (IA COMO FERRAMENTA)

Fundamentalmente, a IA pode servir como uma potente ferramenta para o "malicioso" uso criminal, expandindo e mudando a natureza inerente das ameaças já existentes, bem assim introduzindo ameaças completamente novas (BRUNDAGE *et al.*, 2018).

A ampliação das ameaças já existentes pode acontecer, *v.g.*, no contexto físico. Por exemplo, traficantes de drogas podem recorrer a veículos não-tripulados (especialmente veículos subaquáticos não-tripulados) para aprimorar os êxitos no transporte de drogas e melhorar a resiliência das rotas de tráfico (SHARKEY *et al*., 2010). De maneira dramática [e letal], estudiosos têm alertado para os riscos da combinação do uso de drones quadricópteros baratos, equipados com *software* de reconhecimento facial e pequenas cargas de explosivos, que, em breve, poderá criar um novo vetor para ataques terroristas contra civis (TOPOL, 2016). Efetivamente, eles alertam sobre a possibilidade de uma nova "arma de destruição em massa", que se tornará ainda mais preocupante pelo fato dela poder "discriminar" (étnica ou politicamente) os seus alvos.[15]

Embora esses cenários não-virtuais consistam em significativas preocupações, é no ciberespaço — seu ambiente nativo — que a IA representa a maior ameaça criminosa. Nessa direção, uma das suas utilizações é expandir as ameaças *hacking* e de criação de *malwares* já

[14] Hipoteticamente, poderíamos considerar uma quarta categoria: crimes *contra* a IA (ou seja, a IA como "pessoa" detentora de direitos), mas isso dependeria da concessão de status legal para ela. Há também quem formule uma categoria indireta — o papel da IA ou dos robôs na *promoção* da criminalidade em geral ou na *provocação* de certos crimes —, como a preocupação com a forma através da qual a interação com *bots* sociais e "*sexbots*" podem dessensibilizar [*desensitize*] os perpetradores em relação aos danos sexuais (DANAHER, 2017; KING *et al*., 2019: 15-16).

[15] Para uma representação dramatizada desse cenário, assista o vídeo "*Slaughterbots*", do *Future of Life Institute*, em https://www.youtube.com/watch?v=9rDo1QxI260. [NT - Assista também ao impactante *Sci-Fi Short Film* "Slaughterbots", em https://www.youtube.com/watch?v=O-2tpwW0kmU].

existentes. Pesquisadores já desenvolveram GANs [*Generative Adversarial Networks*] para gerar novos *malwares* que podem passar pelos filtros de vírus (KOLOSNJAJI *et al*., 2018). Outra utilização criminosa é a ampliação de *cyberattacks* de engenharia social. Atualmente, 91% dos crimes ou ataques cibernéticos começam com um *phishing e-mail* (BAHNSEN *et al*., 2018) — uma mensagem que convida alguém a clicar em um *link* que, na sequência, o leva a um *site* que permite que criminosos obtenham informações pessoais sensíveis para fins de roubo de identidade ou fraude. Não obstante, até recentemente, os e-mails de *phishing* eram predominantemente genéricos (*e.g.*, "Você ganhou US$ 1 milhão!") e, portanto, facilmente detectados pelos filtros de *spam* ou pouco convincentes para a generalidade das pessoas, exceto para um subgrupo relativamente pequeno de usuários, particularmente vulneráveis. Ataques de *phishing* mais personalizados ("*spear phishing*") são até quatro vezes mais eficazes do que os não direcionados (JAGATIC *et al*., 2007), porém são mais trabalhosos, pois precisam ser customizados para atingir grupos ou indivíduos específicos. Entretanto, com o "*DeepPhish AI*" (BAHNSEN *et al*., 2018), os sistemas podem aprender automaticamente e combinar recursos (*synthetic URLs* etc.) com outros *phishing attacks*, evitando os filtros de spam e melhorando as taxas de sucesso dos cibercriminosos. Em paralelo, a IA também pode desempenhar um papel relevante na melhoria dos sistemas de defesa: o recentemente desenvolvido "*Panacea AI System*" utiliza processamento de linguagem natural para responder e-mails fraudulentos recebidos, engajando o "atacante" em conversas com o fim de obter informações sobre sua verdadeira identidade, fazendo também desperdiçar seu tempo (DALTON, *et al*., 2020).

Em outro experimento, dois pesquisadores usaram a IA para gerar automaticamente um grande número de mensagens nas redes sociais, todas adaptadas aos perfis e comportamentos passados de alvos específicos, convencendo esses usuários a clicar em *phishing links* (SEYMOUR e TULLY, 2016). De forma semelhante, "*identity-cloning bots*", que imitam pessoas nas redes sociais, têm apresentado altas taxas de sucesso ao se incorporar a redes sociais alheias, uma vez que muitos usuários habitualmente aceitam todos os pedidos de "amizade" (BILGE *et al*., 2009). De fato, em 2019, a *Associated Press* noticiou que rostos gerados por IA foram usados para criar "contas fantasmas" no LinkedIn, com o objetivo de se incorporar ao establishment político de Washington D.C. (SATTER, 2019)[16].

[16] Uma dessas contas *fakes* conseguiu se conectar com Paul Winfree, então Vice-Chefe do Conselho de Política Interna do Presidente Trump. Quando contatado para comentar o caso, Winfree admitiu: "Eu aceito literalmente todos os pedidos de amizade que recebo".

DOI: 10.55689/rcpjm.2026.16.003 | ISSN: 2764-1899

Além de aumentar as ameaças já existentes, como a IA poderia ser usada para desenvolver *novas* ameaças ainda fora do alcance dos atores humanos? Um exemplo vívido e cada vez mais frequente disso são os chamados "*DeepFakes*" (CHESNEY e CITRON, 2019) — aplicações GANs capazes de forjar qualquer tipo de mídia, incluindo fotografias de rostos (VINCENT, 2018),[17] filmagens de vídeo, vozes "clonadas" de amostras de fala de um minuto (GHOLIPOUR, 2017), ou um texto coerente para "*neural fake news*" direcionadas (ZELLERS *et al*., 2019), como ilustrado pelos Sistemas *GPT-2* e *Grover*. Esses *DeepFakes* já se mostraram bastante convincentes. Em março de 2019, criminosos usaram softwares de imitação de voz para copiar a voz de um CEO, ligando para o diretor de uma subsidiária de uma companhia britânica de energia. Isso resultou na transferência de $243,000 para uma conta fraudulenta feita por este ludibriado executivo (HARWELL, 2019). Além disso, grande parte dos comentários em torno das *DeepFakes* tem se preocupado com seu possível uso indevido para manipulação política. Esta preocupação não é injustificada. Por exemplo, na Bélgica, em 2018, o *Flemish Socialist Party* se valeu dessas técnicas para criar um vídeo *fake* de Donald Trump, supostamente pedindo à Bélgica para sair do Acordo de Paris sobre o Clima**.** Com o objetivo de chamar a atenção para as mudanças climáticas, muitos dos apoiadores do partido compartilharam aquele vídeo (VON DER BURCHARD, 2018). Desde então, as *DeepFakes* surgiram gradualmente em diversos contextos eleitorais. Muito embora elas certamente contribuam para o nosso discurso político de "pós-verdade" ["*post-truth*"], fato é que, na atualidade, a principal função dessa tecnologia está na capacidade dos criminosos de criar material sintético, porém plausível, de conteúdo íntimo para fins de assédio, chantagem ou "*sextortion*" (*cyber blackmail*) (SPERA *et al*., 2016). Efetivamente, um relatório de 2019 demonstrou que 96% dos 14.600 vídeos *online deepfake* envolveram a falsificação de material pornográfico não-consensual (AJDER *et al*., 2019).[18] **NT4**

[17] Para um exemplo impactante, consulte https://thispersondoesnotexist.com/, que gera rostos de pessoas inexistentes. Em http://www.whichfaceisreal.com/ , você pode tentar distinguir pessoas reais de falsas.

[18] Um exemplo flagrante do potencial do *DeepFake* para a violência de gênero foi o "*DeepNude*", um aplicativo comercializado em junho de 2019 por US$ 50 que removia as roupas das imagens de mulheres, fazendo com que elas parecessem realisticamente nuas (COLE, 2019). O aplicativo foi retirado do mercado logo após protestos generalizados.

NT4 Em agosto de 2025, milhares de italianas – dentre elas a Primeira Ministra *Giorgia Meloni* e sua irmã *Arianna* – tiveram suas imagens manipuladas e exibidas no site pornô *Phica* (gíria para vagina), causando grande indignação social. "Estou enojada com o que aconteceu e quero expressar minha solidariedade e apoio a todas as mulheres que foram ofendidas, insultadas e violadas em sua intimidade pelos administradores deste fórum e seus usuários", disse *Meloni* ao Jornal *Corriere della Sera*. Como resposta, o Parlamento italiano aprovou, em setembro de 2025, uma lei criminalizando a conduta de produzir, divulgar ou compartilhar

A IA também pode forjar outros tipos de dados de imagem. Um estudo recente mostrou como agentes maliciosos podem se valer da IA para adulterar dados hospitalares, adicionando ou removendo informações de estado de saúde de volumétricos (3D) *scans* médicos. Esses dados falsificados poderiam então ser usados para "sabotar" candidatos políticos, corromper pesquisas científicas, comprometer infraestrutura da saúde pública ou até mesmo cometer homicídios (MIRSKY *et al*., 2019).

Por fim, a IA já está evidenciando a fragilidade dos protocolos de cibersegurança existentes. Em 2017, pesquisadores da *New York University* (NYU) usaram GANs para gerar "*DeepMasterPrints*" — sintéticas "impressões digitais" falsas, que poderiam servir como chave mestra para burlar sistemas de identificação biométrica (BONTRAGER *et al.*, 2017; HERN, 2018).[19] No mesmo ano, o "*PassGAN*" foi treinado com conjuntos de dados de senhas vazadas, aprendendo a gerar prováveis opções para senhas humanas a fim de gerar suposições de senha de alta qualidade. Em testes, esse sistema superou ferramentas de ponta existentes como o *HashCat*, correspondendo entre 51% e 73% a mais de senhas (HITAJ *et al*., 2017).

3.2 CRIMES *NA* INTELIGÊNCIA ARTIFICIAL (IA COMO OBJETO DE ATAQUE)

Crimes "na" IA envolvem ataques que exploram e desmontam vulnerabilidades do sistema na tentativa de enganar ou "*hipnotizar*" sistemas de IA. Há algum tempo é possível "envenenar" os dados de treinamento de um sistema. De maneira infame, o *chatbot* "Tay" do Microsoft Twitter foi tornado "racista" um dia depois que usuários o alimentaram com uma enxurrada de expressões de extrema-direita (GERSHGORN, 2016). Incidentes como este compõem apenas a "ponta do iceberg". Sistemas de ML [*Machine Learning*], ao classificar dados, frequentemente dependem desproporcionalmente de detalhes e padrões contraintuitivos. Dessa maneira, *hackers* podem usar esse recurso para engenharia reversa dos dados de entrada e falsificar sistemas, exibindo comportamentos específicos (NGUYEN *et al*., 2015). O que é pior, isso pode ser feito de maneira não visível para a inspeção humana (GOODFELLOW *et al*., 2014, 2017). Ademais, ataques podem ser realizados até mesmo em configurações de "*black-box*",

imagens falsas produzidas por inteligência artificial, com pena de prisão de uma cinco anos. Esta Lei entrou em vigor no mês seguinte. (CNN BRASIL. *Primeira-Ministra da Itália tem fotos manipuladas em site pornô: 'Enojada',* 29/08/2025. Disponível em: https://www.cnnbrasil.com.br/internacional/primeira-ministra-da-italia-tem-fotos-manipuladas-em-site-porno-enojada/ , acessado em janeiro de 2026).

[19] Vide SHUMAILOV *et al*. (2019) sobre "*acoustic side channel attacks*" [ataques acústicos de canal lateral] em *smartphones*, que podem detectar o som dos dedos nos teclados *touchscreen*, recuperando 61% dos códigos PIN de 4 dígitos em 20 tentativas.

onde um atacante não tem acesso ao peso interno da *network*. Pesquisadores demonstraram ser possível até mesmo gerar um adesivo customizado de "*adversarial patch*", fazendo com que uma IA classifique incorretamente objetos como "torradeiras" (BROWN *et al*., 2017). Em outro estudo, pesquisadores conseguiram imprimir um modelo 3D de "tartaruga", alterado para ser percebido pela IA como sendo um "rifle" por quase todos os ângulos (ATHALYE *et al*., 2018).

GU *et al*. (2017) demonstraram que, em certas ocasiões, esses problemas são agravados por vulnerabilidades na cadeia de suprimentos do modelo de ML. Considerando que muitos usuários "terceirizam" o procedimento de treinamento (computacionalmente intensivo) ou usam modelos pré-treinados, infratores podem criar uma "*BadNet*", isto é, uma *network* maliciosamente treinada que, aparentemente, funciona muito bem no contexto regular do usuário, mas que contém "*backdoors* ambientais" — entradas específicas que enganam o sistema para realizar comportamentos incorretos ou perigosos. Por exemplo, em vários casos, pesquisadores demostraram que colocar adesivos em placas de trânsito e superfícies de rua pode fazer com que *self-driving cars* ignorem as restrições de velocidade e desviem bruscamente para a faixa dos carros que vêm em sentido contrário (EVTIMOV *et al*., 2017; TENCENT KEEN SECURITY LAB, 2019). É provável que esses problemas se tornem cada vez mais frequente. Em 2018, pesquisadores do Google demonstraram que redes neurais de reconhecimento de imagens podem ser enganadas com o objetivo de realizar computações gratuitas para *hackers*, potencialmente transformando smartphones em *botnets* ao expô-los a imagens manipuladas (ELSAYED *et al*., 2018).

Esses "hacks" de IA têm sérias implicações no mundo real em diversos setores. Na área da saúde, pesquisadores demonstraram que ataques adversariais podem cooptar algoritmos de diagnósticos, facilitando as fraudes em seguros de saúde (FINLAYSON *et al*., 2019). Outros estudiosos evidenciaram como até mesmo a IA de processamento de texto pode se tornar vulnerável à manipulação, como foi recentemente demonstrado pelo sistema "*TextFooler*", que poderia analisar textos e sugerir sinônimos estratégicos a serem manipulados para alterar drasticamente as decisões dos sistemas de IA, em áreas que vão desde candidaturas a empregos até detecção de *fake news* (JIN 2020; KNIGHT 2020). Outro aplicativo explorou vulnerabilidades em sistemas de reconhecimento de voz como *Alexa*, *Siri* e *Google Assistant*. Replicando formas de onda de áudio (algumas com precisão até 99,9% do original), os pesquisadores enviavam comandos de voz ocultos para esses *smart speakers*, fazendo-os discar números de telefone ou

mesmo abrir *websites*.[20] Teoricamente, essas manipulações poderiam ser usadas para atacar "*smart homes*" — destravar portas, transferir dinheiro de contas bancárias ou acionar ordens de compra para produtos incriminadores ou embaraçosos (SMITH, 2018).[21]

Essas mesmas técnicas adversariais também estão sendo usadas no contexto do ativismo e da resistência contra cultura(s) de vigilância onipresente. Na Bélgica, pesquisadores criaram uma imagem adversarial que, se impressa e transportada, tornava uma pessoa invisível aos sistemas de visão computacional de IA (THYS *et al*., 2019).[22] De forma semelhante, artistas e estilistas começaram a colaborar com pesquisadores de tecnologia para criar itens de vestuário, como a "*camiseta anti-IA*" (XU *et al*., 2019) e a cosmética "*camuflagem deslumbrante*" (ECKERT et al., 2013), sendo que ambas, aparentemente, poderiam proteger manifestantes da identificação por câmeras de detecção facial.[23]

Em resumo, a participação adversarial evidencia como diversos atores podem espalhar um determinado ambiente com o que SCHARRE e HOROWITZ (2018, 15) chamam de "minas cognitivas", levantando vulnerabilidades críticas de segurança para sistemas de "Internet das Coisas" e trazendo problemas para os defensores do "paradigma da cidade inteligente". Embora se trabalhe na detecção de exemplos adversariais (XIAO *et al.*, 2018), muitos desses métodos de verificação podem ser facilmente contornados (CARLINI e WAGNER, 2017), fazendo com que aqueles que buscam garantir a segurança da infraestrutura de IA estejam sempre ocupados.

3.3 CRIMES *PELA* INTELIGÊNCIA ARTIFICIAL (IA COMO INTERMEDIÁRIA)

Em 2015, um grupo de artistas lançou um *bot* de compras aleatório na *Dark web* — resultando, como era esperado, na compra indevida de drogas proibidas, sendo, então, detidos pela polícia suíça (KASPERKEVIC, 2015). **NT5** Este incidente não só fornece um claro exemplo

[20] Em outro estudo, pesquisadores de segurança cibernética utilizaram *lasers* para manipular silenciosamente os microfones dos sistemas de comando de voz de computadores. Capazes de penetrar vidros de janelas, esses chamados "comandos de luz" expuseram ainda mais trancas digitais e outros eletrodomésticos inteligentes à exploração criminosa (SUGAWARA *et al*., 2019).

[21] Vide https://nicholas.carlini.com/code/audio_adversarial_examples.

[22] Curiosamente, os pesquisadores indicaram que, embora essa defesa só funcionasse em um sistema específico, eles teriam como objetivo gerar imagens que funcionassem em vários detectores simultaneamente (KNIGHT, 2019).

[23] Ao lamentável custo [como contrapartida] de tornar a pessoa altamente visível para a "boa" e "velha" vigilância humana. Vide a página do projeto: https://cvdazzle.com/ [NT: com curiosas imagens da *Dazzle Looks* N°s 6 e 7].

NT5 Sobre este episódio, Jacob Turner faz o seguinte relato: "Na Suíça, um coletivo artístico criou um software chamado *Random Darknet Shopper*, que era ativado uma vez por semana para acessar a *Deep web*, isto é, uma parte 'oculta' da Internet, e comprar um item aleatoriamente. O *Random Darknet Shopper* comprou itens como um par de jeans *fake* da Diesel, um boné de beisebol com uma câmera escondida, 200 cigarros da marca

da nossa terceira categoria do IAC — "Crimes pela IA" —, assim como, o que é mais importante, levanta a espinhosa questão do *status* legal da IA, e seu possível uso indevido como "escudo/intermediário criminal".

Com efeito, alguns juristas sugerem já ser possível conceder a certos algoritmos alguma modalidade de personalidade jurídica. BAYERN (2016) argumentou que brechas nas leis societárias norte-americanas permitem a incorporação funcional de "entidades artificialmente inteligentes", atribuindo-as personalidade jurídica. As complexidades legais por trás desses arranjos normativos, ou os méritos ou o valor social dessa personalidade algorítmica (cf. TURNER, 2018), estão além do escopo deste Artigo. Todavia, o que interessa aqui é registrar como tais chicanas jurídicas descortinam novas possibilidades de crimes de colarinho branco [*white-collar crime*] (LOPUCKI, 2017). Neste particular, usar IA como um intermediário [agente] criminal "independente" traz sérios dilemas para a dogmática penal, tais como a questão do ato criminoso voluntário (*actus reus*) e a intenção criminosa (*mens rea*), além de outras questões relacionadas com as categorias do conhecimento [*knowledge*], previsibilidade [*foreseeability*] e imputabilidade [*criminal liability*] (WILLIAMS, 2017: 25; MCALLISTER, 2018: 47; KING *et al*., 2019: 6-7).

Essas preocupações sobre responsabilidade penal e o elemento intencional provavelmente desempenham um relevante papel em contextos como a manipulação algorítmica do mercado, fixação artificial de preços e a colusão (KING *et al*., 2019: 9-12). Neste sentido, em um experimento de 2016, cientistas da computação demostraram que agentes de trading de IA [*AI trading agentes*] podem descobrir e aprender a executar estratégias lucrativas que equivalem à manipulação de mercado. Usando aprendizado por reforço, um "agente artificial" explorou as lacunas da negociação das ações no mercado de capitais, percebendo que colocar ordens de compra fraudulentas e enganosas era uma estratégia operacional lucrativa (MIRANDA *et al*., 2016). Da mesma forma, informações de precificação algorítmica quase instantâneas garantem que os algoritmos de diferentes empresas possam, sob certas circunstâncias — de forma artificial, inadvertida e tácita — se "contentar" com preços mais

Chesterfield, um conjunto de chaves-mestras utilizada pelo corpo de bombeiros, além de 10 comprimidos de ecstasy. A compra de ecstasy chamou a atenção da polícia local de St. Gallen, que apreendeu o *hardware* do computador no qual o *Random Darknet Shopper* era executado, bem como os demais itens que ele havia comprado. Curiosamente, tanto os designers humanos quanto o sistema de IA foram formalmente "acusados" do crime de compra ilegal de uma substância controlada. Três meses depois, as acusações foram retiradas e todos os bens foram devolvidos ao coletivo artístico (exceto o ecstasy, que foi destruído)." (TURNER, Jacob. *Robot Rules. Regulating Artificial Intelligence.* London: Palgrave MacMillian, 2019, p. 203).

altos, resultando essencialmente na cotização de ações equivalente à colusão (EZRACHI e STUCKE, 2017). Esse tipo de comportamento pode surgir rapidamente, possivelmente como resultado de interações inesperadas entre algoritmos. Embora em muitos casos essas "falhas" sejam facilmente descobertas — considere os dois algoritmos de precificação que, em 2011, travaram uma "guerra robótica" de preços por causa de um livro sobre moscas, elevando a sua cotação para o astronômico patamar de US$ 23,7 milhões[24] (SUTTER, 2011) —, em outros contextos financeiros eles são muito mais difíceis de detectar.

3.4 IAC: ESTIMANDO A AMEAÇA

Vamos concluir nossa digressão sobre a IAC com uma pergunta: as diversas modalidades criminais, acima descritas, se tornarão comuns, ou esses exemplos provarão, mais uma vez, que o que é fácil fazer no laboratório pode ser difícil de reproduzir no mundo real?

Para estimar o quão amplamente disponíveis podem estar essas ferramentas de IA, é instrutivo usar o exemplo da "*Blackshades Remote Access Tool*" [ferramenta de acesso remoto *Blackshades*], que, embora tecnicamente não seja uma aplicação de IA, é útil para ilustrar como as tecnologias digitais podem ser disseminadas e rapidamente acessíveis a diversos interessados. Descrito como "um pacote de franquia criminal" (MARKOFF, 2016), e vendido através do *PayPal* por apenas US$ 40, o *Blackshades* permitia que usuários sem nenhuma habilidade técnica implantassem *ransomware* de forma eficaz, realizando operações de espionagem. Livremente acessível em 2014, a sua vendagem somente foi interrompida após uma grande repressão internacional contra a pirataria cibernética (SULLIVAN, 2014). De fato, ela foi, como bem observou o especialista em cibersegurança BRIAN KREBS (2014), "uma ferramenta criada e comercializada, principalmente, para compradores que não saberiam como *hackear* e escapar de uma situação complicada".

Apesar do *Blackshades* não envolver a IA, não é difícil perceber como os incentivos criminosos serão os mesmos quando se tratar de novas ferramentas ilícitas de IA. Muitas, senão a maioria, das capacidades de IA descritas acima são — ou derivam — de capacidades de uso duplo, isto é, aparentemente inofensivas ou benéficas em outros contextos de utilização. Demais disso, a cultura da IA é caracterizada por um alto grau de abertura e, mesmo em casos em que o código-fonte ainda não é compartilhado abertamente, muitos novos algoritmos de IA podem ser reproduzidos, de forma independente, por outros pesquisadores, em questão de meses,

[24] Mais $ 3,99 de frete.

proporcionando uma baixa capacidade de contenção da sua proliferação (BRUNDAGE *et al*., 2018: 17; SHEVLANE e DAFOE, 2020). No lado da oferta, as ferramentas de IA, especialmente nas versões pré-treinadas, são tão acessíveis quanto qualquer outro software. No lado da demanda, muitas dessas ferramentas oferecem extensões ou melhorias sobre o tipo preciso de capacidades ou tecnologias criminosas que os (ciber)criminosos há muito buscam adquirir, seja em termos de perseguir "*zero-day exploits*" [falhas de segurança no hardware ou software do fornecedor ou desenvolvedor] ou por meio de ferramentas similares ao *Blackshades*.

Em síntese, estimamos que a IAC será um grande problema criminológico em alguns anos.[25]

4. USOS POLICIAIS DA INTELIGÊNCIA ARTIFICIAL

Após discorrer sobre futuro a curto prazo da IAC, abordaremos agora o lado oposto da discussão criminológica. Diante desses crimes, antigos e novos, como os departamentos de polícia aproveitarão as novas tecnologias para equilibrar — ou inverter — o jogo em andamento? Além disso, será que a assunção das novas tecnologias poderá acelerar ainda mais a militarização existente nas culturas policiais (WALL e LINNEMANN, 2014)?

Recentemente, foi dada ampla atenção aos potenciais usos da IA e da robótica para a aplicação da lei [*law enforcement*] (INTERPOL e UNICRI, 2019; ZARDIASHVILI *et al*., 2019), incluindo análises críticas sobre como garantir a democrática prestação de contas para o uso das tecnologias de policiamento preditivo baseadas em ML (VESTBY e VESTBY, 2019). Uma questão que frequentemente surge, como em outras áreas da atividade humana, é se a IA e os robôs substituirão os atores humanos. Aqui, como sugere DANAHER (2018), algumas distinções precisam ser feitas.

A primeira é entre "tarefas" e "trabalhos". A atividade policial envolve uma ampla gama de tarefas específicas (monitoramento, preenchimento de formulários, contabilização dos "números" da criminalidade etc.). Neste contexto, a IA poderá muito bem ser utilizada, com

[25] Para corroborar essa assertiva, estimou-se que, até o final de 2020, somente a *DeepFake* seria responsável por mais de US$ 250 milhões em prejuízos pessoais e corporativos (Forrester, 2019). [NT: Os prejuízos com *deepfakes* têm crescido exponencialmente, transformando-se em uma das maiores ameaças cibernéticas da atualidade. Estima-se que as perdas financeiras globais decorrentes de fraudes facilitadas por IA generativa (incluindo *deepfakes*) aumentem de US$ 12,3 bilhões, em 2024, para US$ 40 bilhões até 2027. Cf. https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing#:~:text=post%2Devidentiary%20world.-,Pillar%201:%20Individual%20epistemic%20agency,et%20al.%2C%202024).]

grande efeito prático. Contudo, o "trabalho" de um policial vai além dessas tarefas, abrangendo papéis mais amplos, tais como "policiamento comunitário", investigação criminal, segurança pública, realização de prisões, assim por diante. Estas atividades são muito mais difíceis de serem substituídas pela IA.

A segunda distinção a ser feita é entre o uso da IA como "ferramenta", "parceiro" e "usurpador". Quando a tecnologia é uma mera *ferramenta* de policiamento, ela pode ser usada para auxiliar em certas atividades específicas que fazem parte do trabalho policial. Pode-se ilustrar essa afirmação com os avanços verificados nas análises forenses de impressões digitais ou de evidências de DNA.[26] Por exemplo, pesquisas recentes demonstraram como a ML pode ajudar com comprovações periciais, tais como reconhecer o calibre e o modelo de uma arma a partir de gravações de áudio de disparos (RAPONI *et al*., 2020), comparar resíduos de tiros da cena do crime com as características químicas da munição não gasta (GALLIDABINO *et al*., 2018), ou, ainda, determinar que tipo de calçado deixou uma determinada marca em uma cena de crime (KONG, 2017). Até certo ponto, as ferramentas de IA podem até desempenhar um papel relevante na identificação e sinais de alerta de crimes intermediados por outros sistemas de IA, como aconteceu com os errantes agentes de *trading* de IA discutidos acima (KING *et al*., 2019: 23-25).[27] Quando a IA opera como *parceira*, certos aspectos da tecnologia podem funcionar de forma autônoma, mas ainda exigirá entrada [*input*] e análise humana — como ocorre com algoritmos de predição de crimes (veja abaixo). Por fim, onde a tecnologia atua como *usurpadora*, não se faz necessária a intervenção humana. Tendo esse arcabouço em mente, agora delinearemos três temas que acreditamos serem constitutivos do próximo "nexo técnico-policial".

4.1 O ESTADO VIGILANTE: MONITORAMENTO INTENSIVO, ABRANGENTE E INESCAPÁVEL

Em primeiro lugar, a IA promete (ou ameaça) auxiliar a atividade de vigilância estatal através da expansão da fotografia digital altamente granular, descrita sucintamente pela ACLU

[26] Todavia, nos processos judiciais em que as provas de impressões digitais e de DNA foram utilizadas pela primeira vez, os juízes mostraram-se demasiado entusiasmados pelas novas tecnologias, atribuindo-as um grau de "autoridade probatória" e "infalibilidade" que, na verdade, elas não mereciam. Vide também ALLDREDGE (2015) sobre o assim chamado "efeito CSI" [*CSI effect*].

[27] Isso não significa que o uso da IA como ferramenta será sempre incontroverso — veja, por exemplo, os problemas que surgiram quando a IA foi usada como "polígrafo" [*lie detector*] em procedimentos de imigração (KENDRIC, 2019; MOLNAR, 2019; BEDUSCHI, 2020) ou quando varreduras de reconhecimento facial foram implementadas para visitantes de prisões na Inglaterra e no País de Gales (JEE, 2019).

[*American Civil Liberties Union*] como "o alvorecer da vigilância robótica" (STANLEY, 2019). Os avanços no armazenamento de dados, somados aos progressos na análise automática de vídeos com IA, podem transformar o monitoramento passivo e disperso em um registro de vigilância cada vez mais detalhado, abrangente e rastreável. A moderna IA já consegue identificar e distinguir emoções (SCHWARTZ, 2019), formas de comportamentos "suspeitos" (SCHNEIER, 2019) e, em um caso recente, supostamente conseguiu identificar potenciais furtadores de lojas [*shoplifters*] pela linguagem corporal, alertando, assim, os seguranças de um supermercado, por intermédio de um aplicativo de smartphone (DU e MAKI, 2019). Além disso, sistemas como o "*iBorderCtrl*" — financiado, desde 2018, pelo programa *Horizon 2020* da União Europeia, e posteriormente implantado em aeroportos —, pretendem fornecer a detecção automatizada de falsidades com base em microexpressões faciais, apesar de suscitarem muitas preocupações quanto à base científica de tal abordagem (JUPE e KEATLEY, 2019).

Para além de interpretar *o que* as pessoas estão fazendo nos vídeos, a IA também consegue reconhecer *quem* está fazendo (PHILLIPS, 2018),[28] mesmo quando os rostos estão disfarçados com máscaras (SINGH *et al*., 2017). Demais disso, a vigilância por IA é agora muito fácil de ser incorporada às infraestruturas de monitoramento já existentes. CHINOY (2019), por exemplo, usou o "*Rekognition*" (a ferramenta de reconhecimento facial comercializável pela Amazon) para demonstrar como é fácil comparar fotos de funcionários de fontes públicas com imagens coletadas por três câmeras comuns ao redor do Bryant Park, em Nova York. Em um dia, o sistema detectou 2.750 rostos, incluindo um professor da *State University of New York* (SUNY). O custo total da instalação foi de apenas US$ 60.

Como já existem extensas bases de dados com nomes e rostos de cidadãos e porque o reconhecimento facial por IA pode ser facilmente integrado na arquitetura urbana das "simples" câmeras de vigilância, a facilidade e a rapidez com que o monitoramento através da IA pode ser implementado é realmente impressionante. De fato, como os rostos (ao contrário das impressões digitais) são difíceis de ocultar, podendo ser escaneados e registrados a qualquer distância e sem que as pessoas percebam, estudiosos argumentam que o reconhecimento facial é categoricamente diferente de outras formas de vigilância, razão pela qual deveria ser proibido (HARTZOG e SELINGER, 2018). A própria ACLU, expressou, da mesma forma, preocupações no

[28] Isso não significa que essa tecnologia seja perfeita. Em 2018, a ACLU descobriu que o sistema "*Rekognition*" da Amazon identificou incorretamente 28 membros do Congresso com fotos de criminosos (SNOW, 2018). Além disso, os sistemas de reconhecimento facial ou emocional têm sido muito criticados por seus vieses raciais (RHUE, 2018).

sentido de que tais sistemas algorítmicos não foram previamente testados, razão pela qual seriam discriminatórios e sujeitos a diversos abusos (STANLEY, 2019: 34-41). Por conta dessas e de similares críticas, STARK (2019) — um estudioso de mídia que trabalha para a *Microsoft Research Montreal* — referiu-se ao reconhecimento facial como sendo "o Plutônio da IA". **NT6**

Vale registrar que a capacidade dos sistemas de IA de detectar e inferir identidades não se limita às câmeras de segurança. Com o mesmo propósito, outras tecnologias estão sendo testadas, incluindo a ecolocalização para identificar atividades humanas (CHEN, 2019); "*Speech2Face*", que permite ao usuário redescobrir imagens faciais, identificadas de forma vaga, incluindo idade, gênero e etnia — tudo isso apenas com o áudio de voz (OHT *et al*., 2019); e o novo laser "*Jetson*" do Pentágono, que identifica assinaturas únicas de batimentos cardíacos (através das roupas) a uma distância de até 200 metros. (HAMBLING, 2019). Cada vez mais, parece que é possível escanear qualquer pessoa com quase qualquer coisa. Difundidos em todas as sociedades, esses desenvolvimentos são uma ferramenta poderosa — não apenas para combater o crime, mas para o controle social em geral. A China possui cerca de 200 milhões de câmeras de vigilância em todo o país e, segundo relatos, começou a incorporar a IA nesses sistemas (MOZUR, 2018). **NT7** Em alguns casos, isso resulta em sistemas de vigilância explicitamente racializados, com algoritmos configurados para identificar membros da minoria étnica *uigur* por suas características faciais (MOZUR, 2019).

Uma dimensão adicional é o crescente entrelaçamento das capacidades das polícias dos Estados com empresas privadas. Por exemplo, a *Axon Enterprise* (antiga *Taser International*) fornece câmeras corporais para 47 das 69 maiores agências policiais dos EUA, e tem participado do marketing de um sistema de IA treinado com 30 *petabytes* de vídeo (mais de dez vezes maior que o banco de dados da Netflix), coletados de 200.000 agentes policiais. Esse sistema processa imagens de câmeras corporais para auxiliar a polícia, antecipando problemas e gerando relatórios de incidentes (PERRY, 2018).[29] Entre tantas preocupações (PATTERSON e

NT6 "Por ser perigoso, discriminatório e carecer de legitimidade, o reconhecimento facial precisa de regulamentação e de controle semelhantes aos aplicados aos resíduos nucleares, como o plutônio". (STARK, Luke. *Facial Recognition is the Plutonium of AI. In* https://dl.acm.org/doi/epdf/10.1145/3313129).

NT7 Em 2025, estima-se que a China possua entre *700 milhões* e *1 bilhão* de câmeras, com reconhecimento facial, instaladas em todo o país (Cf. https://g1.globo.com/fantastico/noticia/2025/08/17/cameras-de-vigilancia-superaplicativos-robos-como-as-cidades-inteligentes-da-china-vao-monitorar-voce.ghtml).

[29] A propósito, em junho de 2019, a Axon anunciou uma moratória sobre o uso do reconhecimento facial nos dispositivos de câmera corporal dos empregados, seguindo a recomendação de seu conselho de ética independente [*independente ethics board*], que considerou que tais sistemas ainda não eram confiáveis o suficiente (WARZEL, 2019). Contudo, um porta-voz da Axon confirmou que os departamentos policiais

GREENE, 2018), JOH (2017) argumentou que esses desenvolvimentos tecnológicos também evidenciam como fornecedores privados de vigilância exercem indevida influência política sobre as práticas investigativas e de realização de prisões pelos departamentos de polícia.[30] Além disso, esse desconforto é agravado por preocupações significativas sobre a base de dados ou softwares de propriedade privada, que normalmente não estão acessíveis ao público. De forma mais ampla, outros estudiosos levantaram sérias preocupações sobre como a herança militar subjacente a muitas tecnologias digitais — como é o caso da Internet —, pode vir a influenciar seu uso para fins de vigilância (LEVINE, 2018). Esses e outros avanços tecnológicos também têm sido interpretados no contexto da construção gradual de arquiteturas mais amplas do "capitalismo de vigilância" (ZUBOFF, 2019).

4.2 O ESTADO OCULTO: VIGILÂNCIA ONIPRESENTE, PORÉM IMPLÍCITA

Em segundo lugar, a integração da IA com *drones* e sensores de "cidades-inteligentes" ["*smart-city*"] cria novas formas de "vigilância de larga escala", que são onipresentes, porém discretas, implícitas e negáveis. Em termos de ubiquidade, a queda no custo de sensores e plataformas de drones, aliada ao aumento da distância "*stand-off*" da funcionalidade das câmeras, está ampliando em muito o alcance da vigilância de IA. Hoje, câmeras *gigapixel* de precisão conseguem reconhecer rostos e placas de veículos com fotos tiradas a quilômetros de distância (SCHNEIER, 2019), de sorte que, um único sobrevoo de drone sobre um protesto ou passeata, pode, a princípio, permitir que as autoridades compilem uma listagem de todos os participantes.[31]

Na verdade, tais capacidades nem são tão novas assim: há uma década, a DARPA lançou o ARGUS-IS, uma plataforma de drones de vídeo não tripulados de 1,8 *gigapixels*, capaz de gravar continuamente uma área de 25 quilômetros quadrados com resolução de 15 cm

---

poderiam, em princípio, baixar as imagens dessas câmeras corporais e processá-las por meio de serviços de reconhecimento facial de terceiros.

[30] Esta questão se faz relevante à luz de recentes pesquisas criminológicas que analisaram os programas de câmeras corporais da polícia, chegando á conclusão de que eles "não tiveram efeitos estatisticamente significativos ou consistentes na maioria dos comportamentos de policiais e de cidadãos, bem como nas opiniões dos cidadãos sobre o trabalho policial" (LUM *et al*., 2019: 93).

[31] No início deste ano [2020], na China, pesquisadores utilizaram algoritmos para processar imagens de uma *lidar-based camera* [câmera de detecção e alcance a laser], instalada em um arranha-céu de Xangai, capaz de distinguir características humanas através da poluição atmosférica, a uma distância de 45 km (LI *et al*., 2019; MIT TECHNOLOGY REVIEW, 2019).

(HAMBLING, 2009)[32]. Em 2014, a *US Air Force* integrou esse programa ao sistema "*Gorgon Stare*", que implanta Imagens de *Wide-Angle-Motion-Imagy* (WAMI) para permitir que drones rastreassem múltiplos "alvos" em grandes áreas. As primeiras aplicações do *Gorgon Stare* apresentaram problemas técnicos (COCKBURN, 2016). Contudo, as subsequentes iterações mostraram-se mais eficientes e tiveram seu uso ampliado, mas de maneira infame, como instrumento de aplicação da lei criminal em Baltimore (MICHEL, 2019). Essa fusão do WAMI com outros sensores digitais e biométricos, inaugurou uma nova era das cidades "totalmente integradas" ou "captadas" (SADOWSKI, 2019; SADOWSKI e BENDOR, 2019).

Em outro contexto de utilização de "vigilância furtiva" à distância, a China desenvolveu drones na forma de pequenas "*pombas robóticas*", permitindo que elas se misturassem com bandos de aves que sobrevoam várias províncias (CHEN, 2018). Sendo assim, se o cidadão consegue se esquivar de câmeras de vigilância pública facilmente visíveis, o mesmo não ocorre no caso das tecnologias camufláveis, a longa distância e invisíveis, especialmente aquelas que não apenas observam, mas também analisam dados sensíveis a partir de projeções estatísticas baseadas no perfil demográfico dos cidadãos.

De fato, em alguns casos, o papel dos poderes públicos, que era o de monitorar explicitamente o comportamento das pessoas, vai se tornando cada vez mais obscuro, visto ser sublimado por intermédio de sistemas descentralizados de controle social mediado tecnologicamente. É o que ocorre no polêmico sistema chinês de crédito social [*Chinese Social Credit system*] — na realidade, um mosaico de diferentes sistemas que coletam dados sobre uma variedade infinita de atividades *online* e *offline*. Essas informações são, então, agregadas em uma pontuação de 800, que é vinculada a benefícios, descontos e outros incentivos sociais (MOZUR, 2018). Ainda que alguns alertem que a sofisticação e o alcance *Orwelliano* do Sistema de Crédito Social da China seriam, frequentemente, superestimados (AHMED, 2019), fato é que, mesmo em seu estágio inicial, ele já demonstra uma "prática evolutiva de controle" (CREEMERS, 2018), que seguramente será aprimorada e fortalecida na medida em que a IA possa permitir o aproveitamento estatal dos dados dos cidadãos.

De forma mais abstrata, essas tendências também demonstram como a tecnologia acelera e exacerba a transição subjacente nos meios pelos quais os governos procuram regular

[32] Vale registrar para os leitores desta Revista Científica, a ironia no fato de que a ideia original por trás do WAMI [*Wide-Angle-Motion-Imagy*] foi concebida por um cientista militar anônimo depois que ele assistiu ao thriller hollywoodiano "*Enemy of the State*", de 1998, um filme no qual Will Smith é rastreado por uma agência estatal desonesta, que se utiliza de vigilância avançada por satélite.

o comportamento dos cidadãos. Isso evidencia a mudança de uma tradicional e explícita normativa aplicação da lei, para uma modelagem não-intrusiva da arquitetura e do espaço (urbano). Por exemplo, JOH (2019) descreveu como o policiamento na "cidade inteligente" segue o modelo da Disneylândia, comparando-a à forma como alguns parques de diversões de alta tecnologia antecipam e previnem desordens ao moldar o comportamento dos visitantes por meio de barreiras físicas, bem como pela onipresença dos funcionários que percebem e interceptam comportamentos erráticos. Estas arquiteturas, aparentemente, não parecem intrusivas ou coercitivas, mas, ainda assim, podem ser ferramentas eficazes de governança. Vemos isso ocorrer nas cidades inteligentes com o desenvolvimento de ferramentas (algorítmicas) para "hiperestimular" ["*hypernudge*"] (YEUNG, 2017) os cidadãos a adotarem certos comportamentos pró-sociais.

Por fim, BROWNSWORD (2015) ampliou esse debate, sugerindo que o surgimento de tecnologias regulatórias de controle (incluindo, mas não se limitando, a IA), pode levar a uma mudança na "modalidade regulatória". Ele ilustra esse argumento se referindo a um "clube de golfe" que está enfrentando problemas com associados que conduzem carrinhos de golfe sobre canteiros de flores. Segundo BROWNSWORD, as possíveis (sucessivas) opções preventivas se apresentam ao clube ou antecipam uma tendência maior no policiamento. Originalmente, o clube de golfe dependia da imposição das normas sociais [informais] entre os membros (vergonha; censura para violadores de canteiros de flores etc.). Quando isso se mostrou insuficiente, eles mudaram para a "lei" formal (estabelecendo preceitos com sanções/multas específicas para os membros flagrados danificando as flores). A aplicação dessa normativa foi eventualmente secundada por intermédio da utilização da tecnologia (câmeras de segurança para monitorar violações). Posteriormente, toda a tecnologia permitiu a substituição do arcabouço normativo: *chips* de GPS foram incorporados nos carrinhos de golfe, e as áreas de canteiros de flores foram geocercadas [*geo-fenced*] para garantir que os carrinhos de golfe desligassem automaticamente ao se aproximarem dos canteiros de flores. Este exemplo hipotético trata de uma mudança fundamental para uma modalidade regulatória não-normativa de controle do comportamento humano, ilustrando bem o fato de que, no futuro, as tecnologias de IA poderão facilitar a sublimação das arquiteturas policiais.

4.3 O ESTADO ORÁCULO: DESDE DETECÇÃO E APLICAÇÃO ATÉ PREDIÇÃO E PREVENÇÃO

Em terceiro lugar, como ocorre em outras áreas, os sistemas de IA podem captar padrões sutis para oferecer predições (supostamente) acuradas de comportamentos futuros, incluindo condutas criminosas. Isto tem facilitado, cada vez mais, uma mudança nas práticas policiais, ou seja, daquelas voltadas a detectar violações para fazer cumprir a lei, para aquelas que buscam antecipar atos criminosos para preveni-los mais eficazmente (DANAHER, 2018). Esta questão tem sido vista em debates de grande repercussão sobre o uso de algoritmos para prever a possibilidade de reincidência ao se decidir sobre a fiança pré-julgamento. Estudando o assunto, KLEINBERG *et al*. (2017) descobriram que o algoritmo que eles criaram superava a avaliação de juízes humanos na previsão do risco de reincidência do réu. Os Autores argumentaram que adotar esse modelo de previsão poderia gerar "ganhos potencialmente elevados de bem-estar (...), pois o crime poderia ser reduzido em até 24,8%, sem mudança nas taxas de detenção, ou a população carcerária poderia ser reduzida em 42,0%, sem aumento nas taxas de criminalidade". Em que pese esses "resultados" serem atraentes, é preciso estar atento não somente ao exagero em torno desses números, bem como quanto aos "esquemas" políticos subjacentes (vide KAUFMANN *et al*., 2018).

Sobre o assunto, muitos dos (mal)afamados exemplos de "policiamento preditivo", na verdade não envolvem tanto o uso de IA. O programa preditivo conduzido pela *Palantir Technologies* em Nova Orleans, a partir de 2012 (em cooperação com a polícia local, mas sem o conhecimento do conselho municipal), foi baseado em mapeamento de redes sociais elaborado por humanos e com algoritmos de pontuação relativamente simples (WINSTON, 2018). Da mesma forma, a "*PredPol*" analisa, basicamente, apenas três variáveis criminais, com o escopo de criar "pontos críticos de criminalidade" ["*crime hot-spots*"] que orientam a alocação de recursos policiais e as rotas de patrulhamento. Porém, tudo isso está muito longe dos padrões complexos depurados por *networks* neurais profundas.

Existem muitos outros problemas — incluindo dúvidas sobre exatidões. Por exemplo, há considerável controvérsia sobre se as predições de reincidência do tão afamado Programa COMPAS são, de fato, mais precisas do que as feitas por pessoas aleatórias (DRESSEL e FARID, 2018; LIN *et al*., 2020). Demais disso, diversos estudos sugeririam que esses programas são marcados por enraizados vieses raciais (KIRCHNER *et al.*, 2016; LUM e ISAAC, 2016; em sentido contrário, KAMYSHEV, 2019), seja utilizando estatísticas de IA ou estatísticas convencionais.

Em outras palavras, se sistemas preditivos de IA são treinados com conjuntos de dados não-representativos ou enviesados, inevitavelmente o resultado será um "descontrolado ciclo retroalimentador" de predições autoconfirmantes (ENSIGN *et al*., 2017). Como o algoritmo designa certas áreas como de "alto risco de criminalidade", as forças policiais tendem a enviar mais patrulhas, garantindo que elas prendam proporcionalmente mais pessoas cometendo crimes, e o algoritmo então processará mais evidências de que aquela área é de alto risco de criminalidade. Em termos práticos, o sistema corrompe seus próprios dados futuros de treinamento. [NT8] Sendo assim, como KAMYSHEV (2019) argumentou, o potencial de policiamento dos sistemas preditivos de IA é prejudicado por sua precisão real abaixo do esperado, por sua falta de transparência, por sua suscetibilidade a ciclos de retroalimentação autocorruptores e, por fim, por sua falha em se alinhar com os objetivos fundamentais de um sistema de justiça.

4.4 INTELIGÊNCIA ARTIFICIAL E POLICIAMENTO: PREDIÇÕES E MEDITAÇÕES

A discussão sobre ferramentas de policiamento da IA pode rapidamente se tornar distópica. Porém, quão amplamente essas ferramentas irão se proliferar? Por um lado, os fornecedores e clientes estarão lá. Por exemplo, em 2018, *Lookout* e a *Electronic Frontier Foundation* (2018) revelaram uma extensa campanha de espionagem de um grupo obscuro chamado "*Dark Caracal*", que usava ferramentas avançadas de *hacking* e vigilância, aparentemente divulgadas por um fornecedor desconhecido, que havia suprido pelo menos meia dúzia de outras campanhas de vigilância.[33] Embora deva ser enfatizado que o *Dark Caracal* não envolvia ferramentas de IA, não é difícil imaginar que, no futuro, ferramentas mais sofisticadas de "vigilância por assinatura" serão muito mais atraentes para alguns governos, como indicam relatos que mostram como a China já está vendendo tecnologia de vigilância

[NT8] Esse sistema preditivo lembra o chamado "*Erro de Lombroso*". De acordo com António García-Pablos de Molina, o método de Cesare Lombroso foi muito controvertido, dentre outros motivos, porque, para verificar a natureza atávica do delinquente, ele realizou uma investigação com 25 mil reclusos nos cárceres europeus. Sendo assim, portador de um viés positivista, Lombroso associou erroneamente *criminoso* com *recluso*, tendo aceitado, acriticamente, como objeto de pesquisa, o resultado ou subproduto final, sempre discriminatório e seletivo, do controle penal. Sua *teoria do delinquente nato*, na verdade, foi uma *teoria do recluso nato*. (GARCÍA-PABLOS DE MOLINA, ANTÓNIO. *Tratado de Criminología*. 3ª ed. Valencia: Tirant lo Blanch, 2003, pp. 422-423) (grifos do original).

[33] Suspeita-se que o *Dark Caracal* seja, ele próprio, um grupo de *hackers* patrocinado por um Estado-nação desconhecido. Isso ilustra até que ponto os Estados podem, publicamente ou por meio de representantes, encontrar maneiras de vender inescrupulosamente novas tecnologias de vigilância aos interessados.

plug-and-play para países como o Equador (MOZUR *et al*., 2019). Por outro lado, também não é difícil perceber o ceticismo com a rapidez da adoção de novas tecnologias de monitoramento. Por exemplo, em seu estudo sobre a polícia dinamarquesa, SAUSDAL (2018) demostrou que, ao contrário das "ousadas" alegações das empresas de tecnologia, os investigadores na verdade viam as ferramentas de vigilância de alta tecnologia como frustrantes e, frequentemente, um obstáculo para o trabalho policial.

Por oportuno, vale lembrar que a primeira "lei" da tecnologia, cunhada por MELVIN KRANZBERG (1986), diz: "*a tecnologia não é nem boa nem má; nem é neutra*". Dessa forma, a discussão acima certamente nos traz motivos para preocupações quanto ao uso pelo aparato policial dos instrumentos de IA. Contudo, assim como rejeitamos o ingênuo tecno-otimismo, também não devemos cair no baixo-astral do pessimismo tecnológico. Na verdade, grande parte do impacto real da IA dependerá das criteriosas escolhas políticas e culturais que as sociedades farão de forma mais geral. Seja como for, significativamente, as novas tecnologias podem mudar os próprios termos dos *trade-offs* sociais que há muito consideramos axiomáticos.

Para dar um exemplo provocativo: o termo "vigilância com preservação da privacidade" pode parecer um oximoro para alguns. No entanto, embora os criminólogos estejam certos ao abordar esses novos conceitos com uma dose de desconfiança, ainda assim eles terão que se envolver com as novas tecnologias de IA — nas áreas de "criptografia homomórfica" ou "aprendizado federado" — que podem, potencialmente, desenvolver sistemas de monitoramento que sejam (ao menos superficialmente) menos intrusivos e mais responsáveis do que as antigas abordagens de vigilância. A dissolução ou suavização dessas concessões não é inédita: *v.g.*, a introdução de cães farejadores nos aeroportos ofereceu uma nova forma de detectar drogas ou mesmo bombas, que foi tanto mais eficaz quanto menos invasiva do que as anteriores medidas de segurança (TRASK, 2017). No mesmo sentido, se configurado e utilizado adequadamente, a IA poderá até servir como uma "tecnologia que melhora a privacidade" (ELS, 2017; vide BIRNSTILL *et al.*, 2015), ou, pelo menos, potencialmente reduzir a intromissão das agências policiais, transformando a vigilância digital, de um instrumento contundente para um instrumento afiado. Em uma escala maior, as formas que escolhemos para equilibrar segurança e privacidade também podem ser reavaliadas, na hipótese das tecnologias cada vez mais poderosas nos conduzirem a um "mundo vulnerável" (BOSTROM, 2019).

Oferecemos este panorama, não na esperança de convencer o leitor. Na verdade, propomos que ele seja uma ilustração dos debates que os novos sistemas de IA irão — e, talvez,

devessem — reabrir. Positivamente, esperamos que a IA, e as escolhas que as sociedades fazem em torno dela, possam ajudar os criminólogos a reexaminar e, se necessário, reconsiderar certas suposições fundamentais ou arraigadas que, gostem ou não, serão colocadas em xeque pela nova era tecnológica que se descortina.

5. CONCLUSÃO

Este Artigo apresentou o conceito de IA para os estudiosos da criminologia, oferecendo uma visão cautelosa, porém ponderada, do funcionamento, aplicações, pontos fortes e limites dessa tecnologia. Podemos resumir nosso ponto de vista da seguinte forma: *apesar de toda sua utilidade, a IA não é mágica*. Assim como qualquer programa orientado por dados, sua objetividade e eficácia ainda estarão determinadas pelo já mencionado axioma computacional do "GIGO". De fato, em razão da natureza essencialmente frágil da tomada de decisão em redes neurais, está claro que a expertise humana detalhada é ainda mais importante na computação hoje do que no passado; tanto em relação à formulação de parâmetros/hipóteses, quanto à governança geral dos sistemas.

As ambições deste Artigo, entretanto, pretenderam ir além da mera procura de um equilíbrio entre narrativas exageradas de distopia tecnológica, de um lado, e utopia do Vale do Silício, de outro (vide, *e.g.*, BARBROOK e CAMERON, 2001, sobre a "Ideologia da Califórnia"). Acima de tudo, temos o objetivo mais específico de incentivar criminólogos de todas as áreas a expandirem seus interesses de pesquisa na direção do "nexo tecnologia-crime". Naturalmente, muitos criminólogos hesitam com esta proposição, acreditando que — sem um diploma de matemática ou da ciência da computação —, são incapazes de contribuir, de forma qualificada, ao debate sobre IA e ML. Discordamos. Efetivamente, à medida que o impacto da IA em áreas como policiamento, punição, decisões judiciais e, inevitavelmente, criminalidade, continuar a crescer, crescerá na mesma proporção a necessidade dos criminólogos de se envolverem plenamente com a tecnologia em rede digital e toda a sua complexidade. Isto não é apenas desejável, mas essencial se quisermos ter alguma chance de controlar ou limitar seus potenciais excessos.

Nosso pensamento aqui é moldado pelo que JAMES BRIDLE (2018) chama de *abismo do pensamento computacional*: a clara e desconcertante tendência, em todas as esferas da vida contemporânea — da educação à guerra — de ceder poder e espaço às tecnologias reificadas e aos sistemas conectados a vastos repositórios de dados, na crença de que qualquer problema ou

desafio social poderá ser resolvido, unicamente, pela utilização da computação e da aceleração tecnológica. Para BRIDLE (2018), a fé irrefletida em uma combinação de informação e automação representa um "*hack cognitivo*", no qual a tomada de decisão e a conscientização são transferidas para as máquinas, resultando em "uma opacidade cada vez maior que gera uma concentração de poder, e o direcionamento desse poder para domínios cada vez mais restritos de experiência" (p. 34). Para enfrentar esse direcionamento, os criminólogos precisam fazer mais do que simplesmente criticar as tendências problemáticas da IA. Na verdade, devemos tentar "moldar" e "direcionar" proativamente o debate sobre tecnologia em nossa área de conhecimento. Somente dessa forma, ou seja, quando ampliarmos a imaginação criminológica o suficiente para "abraçar" plenamente a conexão tecnologia-crime, estaremos em condições de garantir sistemas e práticas digitais que sejam tanto *éticas* quanto *não-discriminatórias*.

Por último, uma palavra sobre nomenclatura. Ao longo dos anos, criminólogos têm se mostrado extremamente criativos ao adicionar um prefixo à sua disciplina. Recentemente, vimos o surgimento de uma série de interessantes subcampos, tais como "criminologia de fronteira", "criminologia visual", "criminologia *queer*", "criminologia do sul" e, agora, até mesmo uma "criminologia fantasma". Nesse passo, à medida que criminólogos direcionam suas preocupações para o estudo da tecnologia, é provável que surjam novos prefixos. Contudo, se a criminologia quiser se envolver plenamente com os algoritmos complexos, redes e infraestruturas digitais, que agora mediam os seres humanos e seus ambientes, a ideia de fragmentar ainda mais a disciplina em subespecialidades não seria a mais adequada. Parece muito melhor fazer mudanças em um nível mais universal; para forjar uma criminologia completa e tecnológica, capaz de lidar com a próxima onda disruptiva e os realinhamentos científicos, na medida que eles vão chegando. É por essa razão que não estamos aqui pedindo por algo tão específico quanto uma "criminologia da IA". Tal proposição seria muito restrita e, com o tempo, precisaria ser reforçada (quiçá substituída) por denominações como "criminologia da computação quântica", "criminologia de *biohacking*", assim por diante. Se, como agora parece claro, novas tecnologias radicais estão redesenhando os contornos da ordem social existente, com profundas implicações para o crime e a punição, a criminologia deve se adaptar aos novos ventos, para continuar relevante. Essa providência, naturalmente, vai exigir a reimaginação dos marcos teóricos e metodológicos existentes na criminologia, incluindo o abandono tardio de algumas das concepções clássicas do comportamento humano e certas teorias do delito elaboradas no século XX.

Fazer menos do que isso significaria arriscar a obsolescência da criminologia.

6. REFERÊNCIAS

Ackerman E (2011) Qbo robot passes mirror test. IEEE *Spectrum: Technology, Engineering, and Science News*, 6 December. Available at: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/qbopasses-mirror-test-is-therefore-selfaware

Ahmed S (2019) The messy truth about social credit. *Logic Magazine*, 1 May. Available at: https://logicmag.io/china/the-messy-truth-about-social-credit/

Ajder H, Patrini G, Cavalli F et al. (2019) The state of DeepFakes: landscape, threats, and impact. *Deeptrace Labs*, September. Available at: https://regmedia.co.uk/2019/10/08/deepfakereport.pdf

Aldridge J (2019) Does online anonymity boost illegal market trading? *Media, Culture & Society* 41: 578-583.

Alldredge J (2015) The 'CSI Effect' and its potential impact on juror decisions. *Themis: Research Journal of Justice Studies and Forensic Science* 3: 1 5.

Athalye A, Engstrom L, Ilyas A et al. (2018) Synthesizing robust adversarial examples. Available at: https://arxiv.org/abs/1 707.07397

Bahnsen AC, Torroledo I, Camacho LD et al. (2018) DeepPhish: simulating malicious Al June 9. https://pdfs.semanticscholar.org/ae99/765d48ab80fe3e221f2eedec719af80b93f9.pdf?_ga=2.137195056.1064399283.1590653531-1585409390.1590653531

Barbrook R and Cameron A (2001) The Californian ideology. In: Ludlow P (ed.) *Crypto Anarchy, Cyberstates, and Pirate Utopias*. Cambridge: MIT Press, pp. 363-387.

Barocas S and Selbst AD (2016) Big Data's disparate impact. *California Law Review* 104: 671-732.

Bayern S (2016) The implications of modern business-entity law for the regulation of autonomous systems. *European Journal of Risk Regulation* 7: 297-309.

Beduschi A (2020) International migration management in the age of artificial intelligence. *Migration Studies*. DOI: 10.1093/migration/mnaa003.

Bennett Moses L (2007) Why have a theory of law and technological change? *Minnesota Journal of Law, Science & Technology* 8: 19.

Bilge L, Strufe T, Balzarotti D et al. (2009) All your contacts belong to us: automated identity theft attacks on social networks. In: *Proceedings of the 18th international conference on world wide web*, Madrid, 20-24 April, pp. 551-560. New York: ACM.

Birnstill P, Bretthauer S, Greiner S et al. (2015) Privacy-preserving surveillance: an interdisciplinary approach. *International Data Privacy Law* 5(4): 298-308.

Bontrager P, Roy A, Togelius J et al. (2017) DeepMasterPrints: generating MasterPrints for dictionary attacks via latent variable evolution. Available at: https://arxiv.org/abs/1705.07386

Bostrom N (2019) The vulnerable World hypothesis. Global Policy 10: 455-476.

Bridle J (2018) *New Dark Age*. London: Verso.

Broadhurst R, Maxim D, Brown P et al. (2018) *Artificial Intelligence and Crime: A Report for the Korean Institute of Criminology*. Canberra, ACT, Australia: ANU Cybercrime Observatory.

Brown S (2006) The criminology of hybrids: rethinking crime and law in technosocial networks. *Theoretical Criminology* 10: 223-244.

Brown TB, Mane D, Roy A et al. (2017) Adversarial patch. Available at: https://arxiv.org/abs/1712.09665

Brownsword R (2015) In the year 2061: from law to technological management. *Law, Innovation and Technology* 7: 1-51.

Brundage M, Avin S, Clark J et al. (2018) The malicious use of artificial intelligence. Available at: https://arxiv.org/ftp/arxiv/papers/1 802/1802.07228.pdf

Carlini N and Wagner D (2017) Adversarial examples are not easily detected: bypassing ten detection methods. In: *Proceedings of the 10th ACM workshop on artificial intelligence and security* (A/Sec '17), Dallas, TX, 3 November, pp. 3-14. New York: ACM.

Chan J and Bennett Moses L (2017) Making sense of Big Data for security. *British Journal of Criminology* 57:299-319.

Chen S (2018) China's robotic spy birds take surveillance to new heights. *South China Morning Post*, 24 June. Available at: https://www.scmp.com/news/china/society/article/2152027/china-takes-surveillancenew-heights-flock-robotic-doves-do-they.

Chen S (2019) This Al uses echolocation to identify what you're doing. *Wired*, 28 May. Available at: https://www.wired.com/story/this-ai-uses-echolocation-to-identify-what-youre-doing/

Chesney R and Citron DK (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review* 107: 1753.

Chinoy S (2019) We built an 'unbelievable' (but legal) facial recognition machine. *The New York Times*, 16 April. Available at: https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-newyork-city.html

Cockburn A (2016) *Kill Chain: The Rise of the High-Tech Assassins*. London: Verso.

Cole S (2019) This horrifying app undresses a photo of any women with a single click. *Motherboard*, 26 June. Available at: https://me.me/i/motherboard-this-horrifying-app-undresses-a-photo-of-any-woman-b07945025d024a4aa830a505dc09cc24

Creemers R (2018) *China's Social Credit System: An Evolving Practice of Control*. Rochester, NY: Social Science Research Network.

Dalton A, Aghaei E, AI-Shaer E et al. (2020) The Panacea Threat Intelligence and Active Defense Platform. Available at: http://arxiv.org/abs/2004.09662

Danaher J (2017) Robotic rape and robotic child sexual abuse: should they be criminalised? *Criminal Law, Philosophy* 11: 71-95.

Danaher J (2018) The automation of policing: challenges and opportunities. Available at: https://philosophicaldisquisitions.blogspot.com/2018/10/the-automation-of-policing-challenges.html (accessed 15 October 2018).

Domigos P (2015) *The Master Algorithm.* New York: Basic Books.

Dressel J and Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4:eaao5580.

Du L and Maki A (2019) These cameras can spot shoplifters even before they steal. *Bloomberg*, 4 March. Available at: https://www.bloomberg.com/news/articles/2019-03-04/the-ai-cameras-that-can-spot-shop-lifters-even-before-they-steal

Eckert M-L, Kose N and Dugelay J-L (2013) Facial cosmetics database and impact analysis on automatic face recognition. In: *2013 IEEE 15th international workshop on multimedia signal processing (MMSP)*, Pula, 30 September-2 October 2013, pp. 434-443. New York: IEEE.

Els AS (2017) Artificial intelligence as a digital privacy protector. *Harvard Journal of Law & Technology* 31:217-235.

Elsayed G, Goodfellow I and Sohl-Dickstein J (2018) Adversarial reprogramming of neural networks. Available at: https://arxiv.org/abs/1 806.11146

Ensign D, Friedler SA, Neville S et al. (2017) Runaway feedback loops in predictive policing. Available at: https://arxiv.org/abs/ 706.09847

Evtimov I, Eykholt K, Fernandes E et al. (2017) Robust physical-world attacks on deep learning models. Available at: https://arxiv.org/abs/1 707.08945

Ezrachi A and Stucke ME (2017) *Two Artificial Neural Networks Meet in an Online Hub and Change the Future (Of Competition, Market Dynamics and Society)*. Rochester, NY: Social Science Research Network.

Finlayson SG, Bowers JD, Ito J et al. (2019) Adversarial attacks on medical machine learning. *Science* 363:1287-1289.

Forrester (2019) Predictions 2020: *On the Precipice of Far-Reaching Change*. Forrester Research, 30 October. Available at: https://go.forrester.com/predictions-2020/

Gallidabino MD, Barron LP, Weyermann C et al. (2018) Quantitative Profile-Profile Relationship (QPPR) modelling: a novel machine learning approach to predict and associate chemical characteristics of unspent ammunition from Gunshot Residue (GSR). Analyst. DOI: 10.1039/C8AN01841C.

Gershgorn D (2016) Here's how we prevent the next racist chatbot. *Popular Science*, 24 March. Available at: https://www.popsci.com/heres-how-we-prevent-next-racist-chatbot (accessed 21 February 2019).

Gholipour B (2017) New Al tech can mimic any voice. *Scientific American*, 2 May. Available at: https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice/

Goodfellow IJ, Papernot N, Huang S et al. (2017) Attacking machine learning with adversarial examples. *OpenAl Blog*, 24 February. Available at: https://blog.openai.com/adversarial-example-research/

Goodfellow IJ, Shlens J and Szegedy C (2014) Explaining and harnessing adversarial examples. Available at: https://arxiv.org/abs/1 412.6572

Grace K, Salvatier J, Dafoe A et al. (2018) When will Al exceed human performance? Evidence from Al experts. *Journal of Artificial Intelligence Research* 62: 729-754. DOI: 10.1613/jair.1 .11222.

Greenfield A (2017) *Radical Technologies*. London: Verso. 229

Gu T, Dola-Gavitt B and Gar S (2017) BadNets: identifying vulnerabilities in the machine learning model supply chain. Available at: https://arxiv.org/abs/I708.06733

Hambling D (2009) Special forces' Gigapixel flying spy sees all. Wired, 12 February. Available at: https://www.wired.com/2009/02/gigapixel-flyin/

Hambling D (2019) The Pentagon has a laser that can identify people from a distance-by their heartbeat. *MIT Technology Review*, 27 June. Available at: http://www.technologyreview.com/s/613891/the-pentagon-has-a-laser-that-can-identify-people-from-a-distanceby-their-heartbeat

Hartzog W and Selinger E (2019) Why you can no longer get lost in the crowd. *The New York Times*, 17 April. Available at: https://www.nytimes.com/2019/04/17/opinion/data-privacy.html

Harwell D (2019) An artificial-intelligence first: voice-mimicking software reportedly used in a major theft. *Washington Post*, 4 September. Available at:

https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/

Hayward KJ (2012) Five spaces of cultural criminology. *The British Journal of Criminology* 52(3): 441-462.

Hern A (2018) Fake fingerprints can imitate real ones in biometric systems - research. *The Guardian*, 15 November. Available at: https://www.theguardian.com/technology/2018/nov/15/fake-fingerprints-can-imitate-real-fingerprints-in-biometric-systems-research

Hitaj B, Gasti P, Ateniese G et al. (2017) PassGAN: a deep learning approach for password guessing. Available at: https://arxiv.org/pdf/1 709.00440.pdf

Holt TJ and Bossler AM (2014) An assessment of the current state of cybercrime scholarship. *Deviant Behavior* 35: 20-40.

INTERPOL and UNICRI (2019) *Artificial Intelligence and Robotics for Law Enforcement*. Available at: http://www.unicri.it/news/files/ARTIFICIAL_INTELLIGENCE_ROBOTICSLAW%20ENFORCEMENTWEB.pdf

Jagatic TN, Johnson NA, Jakobsson M et al. (2007) Social phishing. *Communications of the ACM* 50: 94-100.

Jee C (2019) Prisons are using face recognition on visitors to prevent drug smuggling. *MIT Technology Review*, 6 March. Available at: https://www.technologyreview.com/the-download/613080/prisons-are-using-face-recognition-on-visitors-to-prevent-drug-smuggling/ (accessed 9 March 2019).

Jin D (2020) *Jindl /TextFooler*. Python. Available at: https://github.com/jindl 1/TextFooler

Joh EE (2017) The undue influence of surveillance technology companies on policing. *NYU Legal Review Online*. Available at: https://www.nyulawreview.org/online-features/the-undue-influence-of-surveillance-technology-companies-on-policing/

Joh EE (2019) Policing the smart city. *International Journal of Law in Context* 15(2): 177-182.

Jupe LM, and Keatley DA (2019) Airport artificial intelligence can detect deception: or am i lying? *Security Journal*. DOI: 10.1057/s41284-019-00204-7.

Kamyshev P (2019) Machine Learning In The Judicial System Is Mostly Just Hype. *Palladium Magazine*. Available at: https://palladiummag.com/2019/03/29/machine-learning-in-the-judicial-system-is-mostly-just-hype/

Kasperkevic J (2015) Swiss police release robot that bought ecstasy online. *The Guardian*, 22 April. Available at: https://www.theguardian.com/world/2015/apr/22/swiss-police-release-robot-random-darknetshopper-ecstasy-deep-web

Kaufmann M, Egbert S and Leese M (2018) Predictive policing and the politics of patterns. *British Journal of Criminology* 59: 674-692.

Kendric M (2019) The border guards you can't win over with a smile. *BBC Future*, 17 April. Available at: https://www.bbc.com/future/article/20190416-the-ai-border-guards-you-cant-reason-with

King TC, Aggarwal N, Taddeo M et al. (2019) Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. S*cience and Engineering Ethics* 26: 89-120.

Kirchner L, Angwin J, Larson J et al. (2016) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, 23 May. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Kleinberg J, Lakkaraju H, Leskovec J et al. (2017) *Human Decisions and Machine Predictions*. Cambridge, MA: National Bureau of Economic Research.

Knight W (2017) Alpha zero's 'Alien' chess shows the power, and the peculiarity, of Al. *MIT Technology Review*, 8 December. Available at: https://www.technologyreview.com/s/609736/alpha-zeros-alienchess-shows-the-power-and-the-pecuI8rity-of-ai/

Knight W (2019) How to hide from the Al Surveillance State with a color printout. *MIT Technology Review*, 23 April. Available at: https://www.technologyreview.com/f/613409/how-to-hide-from-the-ai-surveillancestate-with-a-color-printout/

Knight W (2020) This technique uses Al to fool other Als. Wired, 23 February. Available at: https://www.wired.com/story/technique-uses-ai-fool-other-ais/

Kolosnjaji B, Demontis A, Biggio B et al. (2018) Adversarial malware binaries: evading deep learning for malware detection in executables. Available at: https://arxiv.org/abs/1803.04173

Kong B (2017) Cross-domain forensic shoeprint matching. Available at: https://www.ics.uci.edu/~fowlkes/papers/KongSRFBMVC_2017.pdf.

Kranzberg M (1986) Technology and history: 'Kranzberg's Laws'. *Technology and Culture* 27: 544-560.

Krebs B (2014) Blackshades Trojan users had it coming. *Krebs on Security*, 14 May. Available at: https://krebsonsecurity.com/2014/05/blackshades-trojan-users-had-it-coming/ (accessed 18 February 2019).

Kruithof K, Aldridge J, Hetu DD et al. (2016) *The Role of the 'Dark Web' in the Trade of Illicit Drugs*. Santa Monica, CA: RAND.

Legg S and Hutter M (2007a) A collection of definitions of intelligence. Available at: https://arxiv.org/abs/0706.3639

Legg S and Hutter M (2007b) Universal intelligence: a definition of machine intelligence. Available at: https://arxiv.org/abs/071 2.3329

Lehman J, Clune J, Misevic D et al. (2018) The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. Available at: https://arxiv.org/abs/I 803.03453

Leibowicz C, Adler S and Eckersley P (2019) When is it appropriate to publish high-stakes Al research? *The Partnership on AI*, 2 April. Available at: https://www.partnershiponai.org/when-is-it-appropriateto-publish-high-sta kes-ai-research/

Levine Y (2018) *Surveillance Valley: The Secret Military History of the Internet*. New York: Public Affairs.

Li Z-P, Huang X, Cao Y et al. (2019) Single-photon computational 3D imaging at 45 km. Available at: https://arxiv.org/abs/1 904.10341

Lin Z, Jung J, Goel S et al. (2020) The limits of human predictions of recidivism. *Science Advances* 6(7):eaaz0652. Available at: https://doi.org/1 0.1126/sciadv.aaz0652

Lookout and Electronic Frontier Foundation (2018) *Dark Caracal: Cyber-espionage at a Global Scale*. San Francisco, CA: Lookout.

LoPucki LM (2017) Algorithmic entities. Law-Econ research paper, UCLA School of Law, Los Angeles, CA.

Lum C, Stoltz M, Koper CS et al. (2019) Research on body-worn cameras. *Criminology & Public Policy* 18: 93-118.

Lum K and Isaac W (2016) To predict and serve? *Significance* 13: 14-19.

McAllister A (2018) Stranger than science fiction: the rise of A.I. interrogation in the dawn of autonomous robots and the need for an additional protocol to the U.N. convention against torture. *Minnesota Law Review*. Available at: http://www.minnesotalawreview.org/wp-content/uploads/2017/06/McAllister.pdf

McGuire MR (2007) *Hypercrime*. London: Cavendish.

McGuire MR, and Holt TJ, (eds) (2017) *The Routledge Handbook of Technology, Crime and Justice*. Abingdon: Routledge.

Markoff J (2016) As artificial intelligence evolves, so does its criminal potential. *The New York Times*, 23 October. Available at: https://www.nytimes.com/2016/10/24/technology/artificial-intelligence-evolves-with-its-criminal-potential.html

Michel AH (2019) *Eyes in the Sky*. Boston, MA: HMH Books.

Miranda EM, McBurney P and Howard MJW (2016) Learning unfair trading: a market manipulation analysis from the reinforcement learning perspective. In: *Proceedings of the 2016 IEEE conference on evolving and adaptive intelligent systems*, EAIS 2016, pp. 103-109. Institute of Electrical and Electronics Engineers Inc. DOI: 10.1109/EAIS.2016.7502499.

Mirsky Y, Mahler T, Shelef I et al. (2019) CT-GAN: malicious tampering of 3D medical imagery using deep learning. Available at: https://arxiv.org/abs/1901 .03597

MIT Technology Review (2019) A new camera can photograph you from 45 kilometers away. *MIT Technology Review*, 3 May. Available at: https://www.technologyreview.com/s/613457/a-new-camera-canphotograph-you-from-45-kilometers-away/

Molnar P (2019) Technology on the margins: Al and global migration management from a human rights perspective. *Cambridge International Law Journal* 8(2): 305-330. DOI: 10.4337/cilj.2019.02.07.

Mozur P (2018) Inside China's dystopian dreams: A.I., shame and lots of cameras. *The New York Times*, 8 July. Available at: https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html

Mozur P (2019) One month, 500,000 face scans: how China is using A.I.to profile a minority. *The New York Times*, 14 April. Available at: https://www.nytimes.com/2019/04/14/technology/china-surveillanceartificial-intelligence-racial-profiling.html

Mozur P, Kessel J and Chan M (2019) Made in China, exported to the world: the surveillance state. *The New York Times*, 24April. Available at: https://www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-police-government.html

Munksgaard R, Demant J and Branwen G (2016) A replication and methodological critique of the study 'Evaluating drug trafficking on the Tor Network'. *International Journal of Drug Policy* 35: 92-96.

Nguyen A, Yosinski J and Clune J (2015) Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, 7-12 June, pp. 427-436. New York: IEEE.

Oht H, Dekel T, Kim C et al. (2019) Speech2Face: learning the face behind a voice. Available at: https://arxiv.org/abs/1905.09773

Paoli GP, Aldridge J, Ryan N et al. (2017) *Behind the Curtain: The Illicit Trade of Firearms, Explosives and Ammunition on the Dark Web.* Santa Monica, CA: RAND.

Parsons C, Molnar A, Dalek J et al. (2019) The predator in your pocket: a multidisciplinary assessment of the Stalkerware application industry. *The Citizen Lab*, June. Available at: https://citizenlab.ca/docs/stalker ware-holistic.pdf

Patterson G and Greene D (2018) The trouble with trusting Al to interpret police body-cam video. *IEEE Spectrum: Technology, Engineering, and Science News*, 21 November. Available at: https://spectrum.ieee.org/computing/software/the-trouble-with-trusting-ai-to-interpret-poice-bodycam-video

Perry N (2018) How Axon is accelerating tech advances in policing. *Policeone*, 22 June. Available at: https://www.policeone.com/police-products/body-cameras/articles/476840006-How-Axon-is-accelerating-techadvances-in-policing/

Phillips PJ (2018) Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America* 115:6171-6176.

Piper K (2018) The case for taking Al seriously as a threat to humanity. Vox. Available at: https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment

Powell A, Stratton G and Cameron R (2018) *Digital Criminology*. London: Routledge.

Radford A, Wu J, Amodei D et al. (2019) Better language models and their implications. In: *Openai Blog*. Available at: https://blog.openai.com/better-language-models/

Raponi S, Ali I and Oligeri G (2020) Sound of guns: digital forensics of gun audio samples meets artificial intelligence. Available at: http://arxiv.org/abs/2004.07948

Rhue L (2018) *Racial Influence on Automated Perceptions of Emotions.* Rochester, NY: Social Science Research Network.

Russell SJ and Norvig P (2009) *Artificial Intelligence*. Harlow: Pearson

Sadowski J and Bendor R (2019) Selling smartness: corporate narratives and the smart city as a sociotechnical imaginary. *Science, Technology, & Human Values* 44(3): 540-563.

Sadowski J (2019) The captured city. *Real Life*, 12 November. Available at: https://reallifemag.com/the-captured-city/

Satter R (2019) Experts: spy used AI-generated face to connect with targets. *AP NEWS*, 13 June. Available at: https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d (accessed 13 June 2019).

Sausdal D (2018) Everyday deficiencies of police surveillance: a quotidian approach to surveillance studies. *Policing and Society*. Epub ahead of print 13 December. DOI: 10.1080/10439463.2018.1557659.

Scharre P (2019) Killer apps: the real danger of an Al arms race. *Foreign Affairs*. Available at: https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

Scharre P and Horowitz MC (2018) *Artificial Intelligence*. Washington, DC: Center for a New American Security.

Schneier B (2008) Inside the twisted mind of the security professional. *Wired*, 20 March. Available at: https://www.wired.com/2008/03/securitymatters-0320/

Schneier B (2019) Al has made video surveillance automated and terrifying. *Vice*, 14 June. Available at: https://www.vice.com/enin/article/bj93z5/ai-has-made-video-surveillance-automated-and-terrifying

Schwartz O (2019) Don't look now: why you should be worried about machines reading your emotions. *The Guardian*, 6 March. Available at:

https://www.theguardian.com/technology/2019/mar/06/facial-recognition-software-emotional-science

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-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

Sharkey N, Goodman M and Ros N (2010) The coming robot crime wave. *IEEE Computer Magazine* 43:116-115.

Shevlane T and Dafoe A (2020) The offense-defense balance of scientific knowledge: does publishing Al research reduce misuse? In: *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, pp. 173-179. AlES '20. New York: ACM.

Shumailov I, Simon L, Yan J et al. (2019) Hearing your touch: a new acoustic side channel on smartphones. Available at: https://arxiv.org/abs/1903.11137

Singh A, Patil D, Reddy GM et al. (2017) Disguised Face Identification (DFI) with facial keypoints using spatial fusion convolutional network. Available at: https://arxiv.org/abs/1708.09317

Smith CS (2018) Alexa and Siri can hear this hidden command. You can't. *The New York Times*, 10 May. Available at: https://www.nytimes.com/2018/05/10/technology/alexa-siri-hidden-command-audioattacks.html

Smith GJD, Bennett Moses L and Chan J (2017) The challenges of doing criminology in the Big Data era: towards a digital and data-driven approach. *British Journal Criminology* 57: 259-274.

Snow J (2018) Amazon's face recognition falsely matched 28 members of congress with mugshots. *American Civil Liberties Union*, 26 July. Available at:

https://www.aclu.org/blog/privacy-technology/surveillancetechnologies/amazons-face-recognition-falsely-matched-28

Solaiman I, Brundage I, Clark J et al. (2019) Release strategies and the social impacts of language models. Available at: https://arxiv.org/abs/1908.09203

Spera C, Wittes B, Poplin C et al. (2016) Sextortion: cybersecurity, teenagers, and remote sexual assault. *Brookings*, 11 May. Available at: https://www.brookings.edu/research/sextortion-cybersecurity-teenagers-and-remote-sexua l-assault/

Stanley J (2019) *The Dawn of Robot Surveillance*. New York: American Civil Liberties Union.

Stark L (2019) Facial recognition is the plutonium of Al. XRDS 25(3): 50-55.

Steinmetz K and Nobles MR (2017) *Technocrime and Criminological Theory*. New York: Routledge.

Sugawara T, Genkin D, Cyr B et al. (2019) Light commands: laser-based audio injection attacks on voice-controllable systems. Available at: https://lightcommands.com/20191104-Light-Commands.pdf

Sullivan G (2014) 5 scary things about the 'Blackshades' RAT. *Washington Post*, 20 May. Available at: https://www.washingtonpost.com/news/morning-mixwp/2014/05/20/5-scary-things-about-blackshades-malware/

Sutter JD (2011) Amazon seller lists book at $23,698,655.93-plus shipping. *CNN*, 25 April. Available at: http://edition.cnn.com/2011/TECH/web/04/25/amazon.price.algorithm/index.html

Tencent Keen Security Lab (2019) *Experimental Security Research of Tesla Autopilot*. Tencent. Available at:

https://keenlab.tencent.com/en/whitepapers/ExperimentalSecurityResearchofTesla_Autopilot.pdf

Thys SVAN, Ranst W and Goedema T (2019) Fooling automated surveillance cameras: adversarial patches to attack person detection. Available at: https://arxiv.org/abs/1 904.08653

Topol SA (2016) Killer Robots are coming and these people are trying to stop them. *Buzzfeed*, 26 August. Available at: https://www.buzzfeed.com/sarahatopol/how-to-save-mankind-from-the-new-breed-ofkiller-robots

Trask A (2017) Safe crime prediction: homomorphic encryption and deep learning for more effective, less intrusive digital surveillance. Available at: https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/ (accessed 8 June 2017).

Turing AM (1950) Computing Machinery and Intelligence. Mind: *A Quarterly Review* 59: 433-460.

Turner J (2018) *Robot Rules: Regulating Artificial Intelligence*. New York: Springer Berlin Heidelberg.

Vestby A and Vestby J (2019) Machine learning and the police: asking the right questions. Policing: *A Journal of Policy and Practice*. Epub ahead of print 14 June. DOI: 10.1093/police/paz035.

Vincent J (2018) These faces show how far AI image generation has advanced in just four years. *The Verge*, 17 December. Available at: https://www.theverge.com/2018/12/17/18144356/ai-image-generation-fakefaces-people-nvidia-generative-adversarial-networks-gans

Von der Burchard H (2018) Belgian socialist party circulates 'deep fake' Donald Trump video. POLITICO, 21 May. Available at: https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreementbelgian-socialist-party-circulates-deep-fake-trump-video/

Wall T and Linnemann T (2014) Staring down the state: police power, visual economies, and the 'war on cameras'. *Crime, Media, Culture* 10(2): 133-149.

Warzel C (2019) A major police body cam company just banned facial recognition. *The New York Times*, 27 June. Available at: https://www.nytimes.com/2019/06/27/opinion/police-cam-facial-recognition.html

Williams ML and Burnap P (2016) Cyberhate on social media in the aftermath of woolwich: a case study in computational criminology and Big Data. *British Journal of Criminology* 56: 211-238.

Williams R (2017) Lords select committee, artificial intelligence committee, written evidence (AIC0206). Available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial intelligence-committee/artificial-intelligence/written/70496.html#_ftn 13

Winston A (2018) Palantir has secretly been using New Orleans to test its predictive policing technology. *The Verge*, 27 February. Available at: https://www.theverge.com/2018/2/27/17054740/palantir-predictivepolicing-tool-new-orleans-nopd

Xiao C, Deng R, Li B et al. (2018) Characterizing adversarial examples based on spatial consistency information for semantic segmentation. Available at: http://arxiv.org/abs/1 810.05162

Xu K, Zhang G, Liu S et al. (2019) Evading real-time person detectors by adversarial T-shirt. Available at: https://arxiv.org/abs/1 910.11099

Yeung K (2017) 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication Society* 20(1): 118-136.

Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N and Cirkovic MM (eds) *Global Catastrophic Risks*. Oxford: Oxford University Press, pp. 308-345.

Zardiashvili L, Bieger J, Dechesne F et al. (2019) Al ethics for law enforcement. *Delphi* 4(7). Available at: https://delphi.lexxion.eu/article/DELPHI/2019/4/7

Zedner L (2007) Pre-crime and post-criminology? *Theoretical Criminology* 11: 261-281.

Zellers R, Holtzman A, Rashkin H et al. (2019) Defending against neural fake news. Available at: http://arxiv.org/abs/1905.12616

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.