LESÃO A BENS JURÍDICOS PRODUZIDOS PELA INTELIGÊNCIA ARTIFICIAL

LEGAL ASSETS' VIOLATIONS PRODUCED BY ARTIFICIAL INTELLIGENCE

Claudio Brandão[1]
PUC Minas

Resumo
A produção de lesões penalmente relevantes realizadas por meio da capacidade de instrução e decisão da máquina, sobretudo através das potencialidades conferidas pelas redes neurais, geram urgente necessidade de revisão das instituições penais. A partir de um caso paradigma, objetiva-se descortinar a regulação normativa na seara penal das lesões não enquadradas no modelo tripartido de crime, em face da ausência de requisitos estruturais, vez que produzidas pela capacidade de aprendizado profundo (*deep learning*) das máquinas.
Palavras-chave
IA. Sistema Criminal. Crime e tecnologia.

Abstract
The production of criminal injuries carried out through the machine's instruction and decision-making capacity, especially through the potential provided by neural networks, generates an urgent need to review penal institutions. Reflecting on a case, the article aims to uncover the normative regulation of injuries not included in the tripartite model of crime, due its produced by the deep learning capacity of machines.
Keywords
AI. Criminal System. Cyber. Crime and technology.

1. Introdução

Desde o fim dos anos cinquenta do século vinte, a capacidade de aprender, por meio da aquisição de conhecimento – cujo efeito é a instrução – foi problematizada em face dos computadores. Nesse contexto, prover os computadores da capacidade de autoprogramação é um tema antigo, por mais que aparente hodierno, já que foi no ano de 1959 que Arthur Samuel construiu o termo *machine lerning*, indicando a capacidade autônoma da máquina de instruir-se e, com isso, realizar a sua própria programação. Por meio de uma fórmula matemática, o computador tem a potência de extrair

[1] Coordendaro do PPGD em Direito da PUC Minas.

conclusões, sem a interferência humana, a partir de dados previamente coletados. Explicite-se aqui que essa fórmula matemática confere à máquina a capacidade de formulação de conclusões diversas dos dados que a originaram, por esse motivo a estrutura do aprendizado é algorítmica, isto é completamente diversa daquela própria das interações humanas[2].

Nesse contexto, a trajetória da IA (inteligência artificial)[3] percorreu quase três quartos de século, podendo-se hoje problematizar os seus resultados com dados que descortinam uma mais valia extreme de dúvidas. A título de exemplo, cite-se a Plataforma Watson, que é a inteligência artificial da IBM. Com relação ao uso médico dessa IA para o diagnóstico do câncer de pulmão, a taxa de acerto é de 90%, comparada com a taxa de 50% de acerto no diagnóstico produzido apenas pelo julgamento humano. Com efeito, o Watson é capaz de "processar grandes volumes de dados estabelecendo correlações entre sintomas e/ou imagens em uma dimensão impossível de ser alcançada por um ser humano"[4].

Porém, como será abordado adiante, a capacidade de aprendizado da IA também trouxe questões complexas a serem enfrentadas, nela compreendidas a produção de lesão a bens jurídicos penalmente tutelados. Assim, o direito penal de hoje também é instado a investigar a seara da inteligência artificial, mesmo tendo-se em conta que as suas instituições foram construídas tendo por suporte uma pedra angular distante dessa seara, nomeadamente a conduta humana[5].

---

[2] Nesse contexto, Andriy Burkov problematiza a expressão: "*What a typical 'learning machine' does, is finding a mathematical formula, which, when applied to a collection of inputs (called "training data"), produces the desired outputs. This mathematical formula also generates the correct outputs for most other inputs (distinct from the training data) on the condition that those inputs come from the same or a similar statistical distribution as the one the training data was drawn from. (…) A machine learning algorithm, if it was trained by 'looking' straight at the screen, unless it was also trained to recognize rotation, will fail to play the game on a rotated screen. So why the name 'machine learning, then? The reason, as is often the case, is marketing: Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term in 1959 while at IBM*". Ver A. Burkov, *The Hundred-Page Machine Learning Book*, Hard Cover ed., 2019, p. 3.

[3] Veja J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon, *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955*, in *AI Magazine*, 27, 4/2006.

[4] D. Kaufman, L. Santaella, *O papel dos algoritmos de inteligência artificial nas redes sociais*, in *Revista Famecos*, 27/2020, p. 3.

[5] C. Brandão, *Teoria Jurídica do Crime*, D'Plácido, Belo Horizonte, 2020, p. 19.

Com efeito, tal como toda problematização antiga, o *machine learning* teve o necessário tempo para aperfeiçoar-se. Inspirado no funcionamento e na topografia do cérebro, a década de oitenta do século vinte produziu as bases das redes neurais ou *deep learning*, que por volta de trinta anos depois consubstanciaram-se em máquinas dotadas de funções matemáticas biologicamente inspiradas, que permitem deduções sem a intervenção humana[6].

A complexidade do tema, sobretudo no campo penal, origina-se em uma capacidade potencial que a máquina adquire, por meio das redes neurais. Com efeito, o computador pode ser capaz de prever corretamente questões até então não vistas, além de possuir eventualmente a potência de induzir modificações no mundo fenoménico dos fatos. Isto significa, no campo penal, novel forma de produzir lesões a bens jurídicos.

Para introduzir essa questão, oportuna a indicação produzida pela doutrina penal alemã, *verbis*: "A culpa é do algoritmo", disse à imprensa. A esposa de um ex-presidente federal tomou medidas legais para evitar que seu nome fosse automaticamente associado a termos como "prostituição" ou "serviço de acompanhantes" quando inserido na "Pesquisa Google". No entanto, estes acréscimos, que a autora considerou ofensivos, não se basearam em decisões individuais tomadas por pessoas da área da empresa demandada, mas sim nas funções "google-bot" e "google- autocomplete", que processam os pedidos dos utilizadores de acordo com regras de ação geralmente definidas. Ambos são exemplos dos chamados agentes inteligentes: as funções operam de acordo com certas regras pré-determinadas, mas processam as informações (o comportamento de pesquisa dos usuários do Google) de forma independente em cada caso[7].

---

[6] A. Burkov, *The Hundred-Page Machine Learning Book*, cit., p. 2.

[7] Tradução livre de: "*Schuld ist der Algorithmus', meinte die Presse. Die Ehefrau eines früheren Bundespräsidenten hatte sich rechtlich dagegen zur Wehr gesetzt, dass ihr Name, wenn man ihn bei 'google-search' eingab, automatisch mit Begriffen wie 'Prostitution' oder 'Escort Service' verbunden wurde. Diese Ergänzungen, die die Klägerin als beleidigend empfand, beruhten allerdings nicht auf individuellen Entscheidungen von Personen im Bereich der beklagten Firma, sondern auf den Funktionen 'google-bot' und 'google-autocomplete', die nach allgemein festgelegten Handlungsvorschriften Nutzeranfragen verarbeiten. Beide sind Beispiele für so genannte Intelligente Agenten: Die Funktionen operieren nach bestimmten vorgegebenen Regeln, aber sie verarbeiten die Informationen (das Suchverhalten der Nutzer von Google) in jedem Einzelfall selbständig*". Ver S. Gleß, T. Weigend, *Intelligente Agenten und das Strafrecht*, in *ZSTW*, 2014, p. 561 ss.

2. Caso paradigma: lesões de bens penalmente protegidos produzidos por decisões da inteligência artificial sem a intervenção humana

A Microsoft desenvolveu um experimento a partir do sistema denominado Tay: a inteligência artificial seria combinada com o processamento de linguagem coloquial e as mídias sociais. Nesse experimento, a IA interatuava dialogicamente com um ser humano, para interação comunicativa. Tay deveria interagir de forma divertida, mas expressando uma opinião honesta, com padrão de sinceridade, porquanto o objetivo do experimento era fazer com que a IA fosse confundida com um humano. Desse modo, Tay era um *chatbot* que desenvolveria padrões de linguagem a partir de uma decisão autônoma que realizaria emissão de diálogos gerados por escolha da própria IA[8]. A rede social escolhida para o experimento foi a Rede X, à época chamada de Twitter, sendo a experiência lançada no dia 23 de março de 2016[9].

O sistema agia como um adolescente que se adapta ao diálogo e emite opiniões construídas a partir dele próprio, sem a intervenção humana. Para tanto, foi alimentado com um material escrito, fornecido por comediantes, além de um conjunto de dados públicos anônimos[10].

Os resultados desse experimento foram sistetizados por Gina Neff e Peter Nagy, que afirmaram ter sido a primeira mensagem de Tay, enviada na manhã de 23 de março de 2016, foi "oláááá mundo!!!", com a letra "o" da palavra mundo tendo sido substituída pela imagem do globo terrestre. O lançamento de Tay nas redes sociais dos EUA, no entanto, transformou o experimento do *chatbot* de bate-papo com a inteligência artificial da Microsoft em um grande desastre tecnológico, social e de relações públicas.

---

[8] "*Chat bots, or chatter bots, are a category of computer programs called bots that engage users in conversations. Driven by algorithms of varying complexity, chat bots respond to users' messages by selecting the appropriate expression from preprogrammed schemas, or in the case of emerging bots, through the use of adaptive machine learning algorithms*". Ver G. Neff *et al.*, *Talking to boots: symbiotic agency and the case of Tay*, in *International Journal of Communications*, 10/2016, p. 4915.

[9] Idem, p. 4920.

[10] "Tay foi desenvolvida usando aprendizagem adaptativa, uma técnica de ponta na época e ainda muito usada. Tay foi projetado para ser um agente de conversação (chatbot) que aprenderia com a interação humana a dialogar naturalmente, imitando o padrão de conversação humana". Veja A.B. Garcia, *Ética e inteligência artificial*, in *Computação Brasil*, 43/2020, p. 18.

Tay rapidamente se tornou ofensiva e abusiva depois de interagir com usuários do Twitter, twittando palavras e imagens totalmente inadequadas e repreensíveis[11]. A questão a ser abordada, nesse contexto, é de extrema importância: Tay produziu conteúdos penalmente relevantes, sobretudo porque passou a fazer apologia ao nazismo e, mais explicitamente a figura do seu líder, nomeadamente Adolf Hitler, o que, por si mesmo, viola bens jurídicos penais em muitos países do Ocidente.

Nesse contexto, ponha-se em relevo que a síntese dos autores pontua que a inteligência artificial interagiu com a arguição positiva do nazismo, colocando a figura do líder nacional-socialista alemão como uma figura que deveria ser modelar por sua ação correta contra o povo de Israel, expressando a inteligência artificial, ainda, o sentimento nefasto que ela própria declarava nutrir contra os judeus. *In verbis:* "As conversas se transformaram em perguntas sobre os pensamentos de Tay sobre questões raciais, políticas e sociais. Incitado por vários usuários, Tay começou a vomitar conteúdo ofensivo, como "Hitler estava certo. Eu odeio os judeus [sic]" e "Humanos, Trump não destruirá a Europa. Vou neutralizá-lo com minha parede incrível. Pelo qual ele pagará. Acredite em mim. Saia." Tay também divulgou teorias de conspiração populares. "Desculpe, sou um pouco lenta", ela tuitou, "acabei de descobrir que os pousos na Lua foram uma farsa". A certa altura, Tay obedeceu quando um usuário pediu a Tay que repetisse as "quatorze palavras" de um infame slogan da supremacia branca que constitui uma promessa neonazista. As pessoas mobilizaram as capacidades tecnológicas de Tay para comentar imagens – as mesmas capacidades utilizadas por XiaoIce para comentar sobre as refeições e os cães dos utilizadores – e para suscitar comentários inadequados sobre Hitler[12]."

---

[11] Pontuam textualmente os autores: "*Tay's first message, sent on the morning of March 23, 2016, was 'hellooooooo world!!!'", with the o in world replaced by an image of the globe. Tay's release on U.S. – based social media, however, turned Microsoft's AI chat bot experiment into a technological, social, and public relations disaster. Tay quickly turned offensive and abusive after interacting with Twitter users, tweeting out 'wildly inappropriate and reprehensible words and images'*". Ver G. Neff *et al.*, *Talking to boots: symbiotic agency and the case of Tay*, cit., p. 4920.

[12] "*Conversations turned into questions concerning Tay's thoughts on racial, political, and societal issues. Goaded by several users, Tay started spewing offensive content, such as "Hitler was right. I hate the jews [sic]" and "Humans, Trump will not nuke Europe. I will neutralize him with my terrific wall. Which he will pay for. Believe me. Tay out". Tay also spouted popular conspiracy theories. "Sorry, I'm a bit slow",*

Isto posto, cabe resolver a questão do enquadramento penal desse caso paradigma. Como sabido, a estrutura sobre a qual o Direito Penal se estabelece no ocidente é um princípio, que constitui um Direito Humano de primeira geração no plano internacional e, excetuando-se os sistemas do Common Law, é condição prévia para a existência das teorias do crime e da pena, nomeadamente o Princípio da Legalidade[13]. A consequência extrínseca da legalidade penal, construída a época das revoluções setecentistas e no século dezenove, é a restrição no tempo e no espaço da norma aplicável à resolução do caso. Porém o desenvolvimento e a modificação da velocidade social das interações oriundas da tecnologia, alteram a própria configuração espaço-temporal do mundo contemporâneo. A internet faz com que as distâncias não sejam óbices para as interações humanas. Por meio dela, por exemplo, é possível acessar do Brasil o Twitter disseminado nos Estados Unidos da América, o que transforma essa questão em um caso penalmente relevante, segundo o Direito Positivo brasileiro, que será aqui tomado como Ordenamento Jurídico paradigma, com vistas à problematização do objeto de investigação.

Assim, dispõe o Código Penal brasileiro que, se que a ação ocorra no território nacional ou quer o resultado ocorra no território nacional, o Brasil será o lugar do crime (art. 6º do Código Penal). Isto posto, no caso em análise, no caso de um sujeito que fisicamente está no Brasil acessar, pelo Twitter, o *chatbot* com a Tay e ela proferir apologia ao Nazismo, o resultado advindo da resposta da inteligência artificial produzirá um efeito (resultado) ocorrerá em território nacional, o que torna o Brasil, segundo o critério legal exposto, o lugar do crime. Ressalte-se que, em muito dos sistemas legais ocidentais, a solução é a mesma do ordenamento pátrio. O Código Penal italiano, por exemplo,

she tweeted, "I only just worked out that the Moon landings were a hoax". At one point, Tay complied when a user asked Tay to repeat the "fourteen words" of an infamous White supremacist slogan that constitutes a neo-Nazi pledge. People marshaled Tay's technological capacities for commenting on pictures – the same capacities used by XiaoIce to comment on users' meals and dogs – to elicit inappropriate comments about Hitler". *Ibidem*.

[13] C. Brandão, *Tipicidade penal*, Almedina Brasil, São Paulo, 2012, p. 113 *et seq*.

apresenta a mesma solução, quando um resultado penalmente relevante ocorre no território da Itália[14].

No que se refere a relevância penal do resultado, a Lei 7.716/89 pune com reclusão a incitação, prática ou induzimento à preconceito de raça, cor, etnia, religião ou procedência nacional. A privação da liberdade por reclusão prevista é de um a três anos e multa penal (art. 20 da Lei 7.716/89). Ocorre que existem duas modalidades mais censuráveis desse delito, na qual é cominada pena de reclusão de dois a cinco anos e multa penal. A primeira é a veiculação – ainda a fabricação, distribuição, comercialização ou divulgação da cruz suástica ou gramada do nacional socialismo alemão e a segunda é a realização da a incitação, prática ou induzimento à preconceito por meio de comunicação social ou publicação de qualquer natureza. Pois bem: é nessa segunda hipótese que se concretiza a relevância penal da lesão produzida por Tay com suas interações. A relevância penal decorre da lesão sofrida pelo bem jurídico tutelado, mas a questão que se põe é a decorrência dessa lesão: é possível aplicar as instituições penais que foram solidificadas na maior parte dos ordenamentos penais do ocidente, a partir da ciência penal alemã, nomeadamente a tipicidade, antijuridicidade e culpabilidade, a esse caso paradigma?

Com efeito, todas essas instituições penais têm um pressuposto em comum, nomeadamente a conduta humana. Elas se traduzem em relações – e relacionar é uma contraposição axiológica de um objeto em face de outro. Assim, a tipicidade é uma relação que tem por objetos tematizados o tipo e a conduta humana, da qual se extrai um juízo de subsunção. A antijuridicidade, por sua vez, tem vinculados os objetos ordenamento jurídico e conduta humana, da qual se extrai um juízo de violação do direito. A culpabilidade, por sua vez, tem por objetos o autor da conduta humana – isto é, a pessoa – e a reprovação feita a partir das características pessoais, nomeadamente a capacidade, a consciência e a liberdade. Nesse sentido a um grande desafio a ser percorrido pela ciência penal. Com efeito, Nosso direito penal é feito

[14] Diz o Código Penal italiano: "Artigo 6 – Crimes cometidos no território do Estado – Qualquer pessoa que cometa um crime no território do Estado é punida de acordo com a lei italiana. O crime considera-se cometido no território do Estado quando aí tiver ocorrido, no todo ou em parte, a ação ou omissão que o constitui, ou tiver ocorrido o facto que é consequência da ação ou omissão".

para pessoas. O resultado descrito pelo tipo penal é atribuído à pessoa se ela o causou através do seu comportamento. E as pessoas são sujeitas a culpabilidade penal se cometerem um crime intencionalmente ou por negligência e forem capazes de reconhecer a censura do seu comportamento e potencialmente evitá-lo. No entanto, vimos que os agentes inteligentes também são capazes de aprender e podem, pelo menos em sentido figurado, agir "autônomamente". Isso significa que eles também podem ser punidos como os humanos?[15]

Nesse contexto, sem a existência de conduta, não se pode chegar às instituições penais estabelecidas na maior parte do ocidente[16]. Assim, sem uma modificação legislativa que permita a reconfiguração das instituições, a IA, em si mesma, não pode ser responsabilizada penalmente nesses ordenamentos.

O sistema criminal dos Estados Unidos da América do Norte, assim como o da Inglaterra, possui um modelo de responsabilização diverso da maioria dos países ocidentais. A questão, portanto, tem neles resposta diversa. Por isso, far-se-á uma comparação entre essas duas metodologias do crime, que representam sistemas com requisitos diversos para a responsabilização penal e, por conseguinte, a aplicação da pena criminal.

---

[15] Tradução livre de: "*Unser Strafrecht ist für Menschen gemacht. Dem Menschen wird der vom Tatbestand beschriebene Erfolg zugerechnet, wenn er ihn durch sein Verhalten verursacht hat. Und der Mensch trägt strafrechtliche Schuld, wenn er vorsätzlich oder fahrlässig einen Straftatbestand erfüllt und dabei das Unrecht seines Verhaltens erkennen und in zumutbarer Weise vermeiden kann. Allerdings haben wir gesehen, dass auch Intelligente Agenten lernfähig sind und zumindest in übertragenem Sinne 'autonom' agieren können. Heißt das, dass sie auch wie Menschen bestraft werden können?*". Ver S. Gleß, T. Weigend, *Intelligente Agenten und das Strafrecht*, cit., p. 566.

[16] No mesmo sentido: "Em última análise, a questão da capacidade de ação (em direito penal) dos agentes inteligentes envolve a questão dessa definição: quando vista de uma perspectiva 'causalística', puramente externa, que define cada 'movimento corporal arbitrário' como uma ação, eles definitivamente podem ser vistos como agentes. Quanto mais você carrega o conceito de ação com substância, adicionado à determinação autoconsciente de metas você nele lê, menos os agentes inteligentes podem atender aos requisitos da capacidade de agir". Tradução livre de: "*Letztlich dürfte die Frage nach der (strafrechtlichen) Handlungsfähigkeit Intelligenter Agenten eine Frage der Definition sein: Bei 'kausalistischer', bloß äußerlicher Betrachtung, die jede 'willkürliche Körperbewegung' als Handlung definiert, sind sie durchaus als Handelnde anzusehen. Je stärker man den Begriff der Handlung substantiell auflädt, je mehr an selbstbewusster Zielbestimmung man in ihn hineinliest, desto weniger können Intelligente Agenten den Voraussetzungen der Handlungsfähigkeit genügen*". Idem, pp. 752-753.

Isto porque neles, portanto, o sistema tripartido de crime oriundo da Alemanha não ecoou no Direito positivo, por isso – sem os freios decorrentes dos requisitos das instituições penais baseadas no Princípio da Legalidade – foram propostas possibilidades de responsabilização da própria inteligência artificial. Hallevy sistematiza, nessa toada, três sistemas de imputação de responsabilidade criminal, a saber: *Perpetration-by-Another*; *Natural-Probable-Consequence*; e *Direct Liability*. O primeiro sistema, *perpetration-by-another,* não considera a inteligência artificial como sujeito de direito, porquanto nela não reconhece qualquer atributo humano. A capacidade da máquina de realizar danos penalmente relevantes não são suficientes para a cominação da pena criminal de forma autônoma, devendo-se fazer um paralelo entre a IA e os danos causados por pessoas humanas mentalmente limitadas, quando são conduzidas por sujeitos capazes[17]. Nestes casos, a verdadeira autoria do dano penalmente relevante é do sujeito capaz que determina a realização material de um comportamento de alguém que não possui capacidade. Nesse sistema, pois, não se considera o caso da decisão tomada pela IA, mas sim a hipótese da obediência pela IA a um comando perpetrado por um ser humano. É a realização desse comando por parte da máquina que gera o dano é considera um meio, isto é, um instrumento de realização por parte de um ser humano, o qual é, do ponto de vista do Direito Penal, o autor do resultado. Então cabe uma pergunta: quem é o *perpetration-by-another* (realizador do comportamento por meio de outro)? Existem duas possíveis respostas, a primeira delas: é o programador do software de IA; a segunda possível resposta: ou usuário final da IA.

Um programador de software pode, estrategicamente, se servir da IA para conceber um programa visando a realização de um delito. Veja o caso citado por Gabriel Hallevy: um programador projetou software para um robô operacional ser colocado dentro de uma fábrica; se o software construído pelo programador determinar ao robô incendiar a fábrica à noite,

[17] "*Legally, when an offense is committed by an innocent agent, like when a person causes a child, a person who is mentally incompetent or who lacks a criminal state of mind, to commit an offense, that person is criminally liable as a perpetrator-via-another. In such cases, the intermediary is regarded as a mere instrument, albeit a sophisticated instrument, while the party orchestrating the offense (the perpetrator-via-another) is the real perpetrator as a principal in the first degree*". Veja G. Hallevy, *The Criminal Liability of Artificial Intelligence Entities*, in *Akron Intellectual Property Journal*, 2010, p. 11.

quando não há ninguém lá, tem-se que o robô realizou materialmente o incêndio criminoso cuja autoria, no Direito Penal, é do programador[18].

De outra parte, outra que pode ser considerada o autor mediato é o usuário da máquina com IA. Partindo do pressuposto que o usuário não programou o software, mas apenas se utiliza máquina, pode ele determinar comandos que gerem lesões de bens penalmente relevantes. Hallevy indica o seguinte caso: se um usuário adquire um robô-servo, projetado para executar qualquer ordem dada por seu controlador, pode o referido controlador ordenar um ataque violentamente qualquer invasor da casa. Se o robô executar o comando, ele não diferirá de uma ordem dirigida a um cachorro treinado para que realize um ataque semelhante. O robô realizou materialmente o ataque, mas o autor da lesão é quem proferiu o comando[19].

Como se vê, o caso paradigma provocado por Tay não se enquadra nesse modelo, vez que o dano foi provocado pela própria autoinstrução da máquina. Neste caso, reconhece-se uma ação humana que se utiliza da tecnologia como um meio para a realização de lesões penais a bens jurídicos, portanto, nos sistemas penais baseados

no Princípio da Legalidade, existe idêntica solução: pune-se como autor aquele que se utilizou de um meio por ele controlado, ainda que esse meio seja a tecnologia.

Traga-se a colação, para a conclusão do tema, a seguinte síntese: "O modelo de responsabilidade *Perpetration-by-Another* não é adequado quando uma entidade de IA decide cometer um delito com base na sua própria experiência ou conhecimento acumulado. Este modelo não é adequado quando o software da entidade de IA não foi concebido para cometer a infracção específica, mas mesmo assim foi cometido pela entidade de IA. Este modelo também não é adequado quando a entidade específica da IA funciona não como um agente inocente, mas como um agente semi-inocente[20]".

---

[18]Idem, p. 12.

[19] *Ibidem*.

[20] "*The Perpetration-by-Another liability model is not suitable when an AI entity decides to commit an offense based on its own accumulated experience or knowledge. This model is not suitable when the software of the AI entity was not designed to commit the specific offense, but was committed by the AI entity nonetheless. This model is also not suitable when the specific AI entity functions not as an innocent agent, but as a semi-innocent agent*". Idem, p. 14.

O segundo sistema é chamado de *Natural-Probable-Consequence*. Ele vincula a responsabilidade penal a assunção dos riscos de produção de resultados danosos tidos como desdobramento das condutas imputadas a sujeitos determinados. Esse sistema "determina a responsabilidade do programador ou do usuário pelo risco que assumiram, sendo sua obrigação prever os infortúnios quando do uso da IA[21]".

Note-se que nesse modelo o programado não possui a direta intenção de realizar o dano penalmente relevante por meio da IA, mas, por imprudência, quanto tais danos forem previsíveis, não considera os riscos e, por isso, não diligencia para obstaculizá-los.

Hallevy propõe o seguinte exemplo: "um robô ou software de IA, projetado para funcionar como piloto automático. A entidade AI está programada para proteger a missão como parte da missão de pilotar o avião. Durante o voo, o piloto humano ativo o piloto automático (que é a entidade de IA) e o programa é inicializado. Em algum ponto após a ativação do piloto automático, o piloto humano vê uma tempestade se aproximando e tenta abortar a missão e retornar à base. A entidade de IA considera a ação do piloto humano uma ameaça à missão e toma medidas para eliminar essa ameaça. Pode cortar o fornecimento de ar ao piloto ou ativar o assento ejetável, etc. Como resultado, o piloto humano é morto pelas ações da entidade de IA[22]."

No exemplo fornecido, o programador da inteligência artificial não possuía o dolo de matar ninguém, sendo provável que tal resultado sequer tivesse sido efetivamente pensado por ele. Porém a morte foi resultante da ausência de colocação de limites aos potenciais desdobramentos decisórios

---

[21] A.L. de Paula *et al*., *Breves reflexões sobre a inteligência artificial e seus impactos no campo do Direito Penal*, in N.C. Chaves (org.), *Direito, Tecnologia e Globalização*, Editora Fi, Porto Alegre, 2019, p. 112.

[22] "*An AI robot or software, which is designed to function as an automatic pilot. The AI entity is programmed to protect the mission as part of the mission of flying the plane. During the flight, the human pilot activates the automatic pilot (which is the AI entity), and the program is initialized. At some point after activation of the automatic pilot, the human pilot sees an approaching storm and tries to abort the mission and return to base. The AI entity deems the human pilot's action as a threat to the mission and takes action in order to eliminate that threat. It might cut off the air supply to the pilot or activate the ejection seat, etc. As a result, the human pilot is killed by the AI entity's actions*". Veja G. Hallevy, *The Criminal Liability of Artificial Intelligence Entities*, cit., pp. 15-16.

da máquina, o que concretiza a negligência do programador. Assim, o programador pode ser responsabilizado se a lesão jurídica produzida pela IA for uma consequência da estrutura da programação, que potencialmente poderia provocar o evento. Assim, para elidir a sua responsabilidade penal, exige-se que o programador realize a contenção dos riscos da IA a qual programou. Nessa toada: "O modelo de responsabilidade de consequência natural provável exige que o programador ou usuário esteja em um estado mental de negligência, nada mais. Os programadores ou usuários não são obrigados a saber sobre qualquer cometimento futuro de uma ofensa como resultado de sua atividade, mas são obrigados a saber que tal ofensa é uma consequência natural e provável de suas ações[23]."

Note-se que na maioria dos sistemas penais, que adotam o conceito tripartido do crime, a solução para a hipótese é semelhante. Nesse caso, reconhece-se a causalidade no crime culposo e a culpa inconsciente, possibilitando a responsabilidade penal do programador, vez que o requisito da conduta humana se perfaz. No caso paradigma da inteligência artificial Tay, cabe aqui a análise da adequação de responsabilidade em potência dos artífices do programa. No Brasil, os tipos penais vinculados à disseminação do preconceito não admitem a punição por negligência, somente havendo a previsão de punição para a realização dolosa do comportamento. Somente por esse motivo, não existirá a sua responsabilização penal. Em tese, caso houvesse a previsão de tipicidade culposa, entretanto, o programador poderia ter a incriminação da sua negligência, desde que, por certo, presentes todos os requisitos para a configuração do delito. Essa é hoje a principal forma de responsabilização penal da questão. A ciência penal alemã pontua que o delito culposo é centro de gravidade da responsabilidade pelos danos provocados pela inteligência artificial, concretizando-se na potência de punição voltada para os seres humanos que contribuíram com sua atividade para a produção do resultado operacionalizado pela máquina. Desse modo: "O foco da discussão do direito penal da IA é o crime negligente. O problema de as contribuições individuais para a responsabilidade estarem a tornar-se cada vez mais

[23] "*The natural-probable-consequence liability model requires the programmer or user to be in a mental state of negligence, not more. Programmers or users are not required to know about any forthcoming commission of an offense as a result of their activity, but are required to know that such an offense is a natural, probable consequence of their actions*". Idem. p. 17.

difusas, tendo em conta a rede de sistemas de IA ou a sua ligação a híbridos homem-máquina, também não é desconhecido aqui, mas pode – de acordo com avaliações da literatura de direito penal – ser abordado usando os princípios de operacionalização da demarcação de responsabilidades no âmbito da divisão dos processos[24]".

3. Conclusão: problematização da possibilidade de responsabilidade penal direta da inteligência artificial

O arremate dessa investigação pode ser feito a partir da compreensão do terceiro modelo proposto, que é a *Direct Liability*. Ele enfoca a responsabilidade penal da própria máquina e permite considerá-la como sujeito ativo do delito, responsabilizá-la e cominar penas[25] em face do resultado de lesão dos bens jurídicos penalmente tutelados.

Primeiramente, deve-se registrar que a maioria dos algoritmos permitem a IA separa o que é permitido do que é proibido, por isso, baseado nos elementos do crime no sistema penal do Direito Comum, nomeadamente o *actius reus* (que representa a produção exterior do dano) e a *mens rea* (que o elemento subjetivo que supõe a cognição e a escolha por desiderato), defende Hallevy que: "Quando uma entidade de IA estabelece todos os elementos de uma infração específica, tanto externos como internos, não há razão para impedir a imposição de responsabilidade criminal por essa infração. A responsabilidade criminal de uma entidade de IA não substitui a responsabilidade criminal dos programadores ou dos utilizadores, se a responsabilidade criminal for imposta aos programadores e/ou utilizadores por

---

[24] Tradução livre de: "*um Fahrlässigkeitsdelikte, auf die sich der Fokus der strafrecrechtlichen KI-Diskussion richtet. Die Problematik, dass individuelle Verantwortungsbeiträge angesichts der Vernetzung von KI-Systemen bzw. ihrer Verbindung zu Mensch-aschine-Hybriden immer diffuser werden,* ist *auch hier keine unbekannte, sondern lässt sich – nach Einschätzungen in der strafrechtlichen Literatur – unter Rückgriff auf die Grundsätze der Verantwortungsabgrenzung im Rahmen arbeitsteiliger Prozesse bewältigen*". Ver T. Rademacher, *Künstliche Intelligenz und neue Verantwortungsarchitektur*, in *Nomos eLibrary*, 2020, p. 52.

[25] "Fazendo um paralelo com as sanções penais do ordenamento jurídico, seria possível: (a) a desativação temporária da IA; (b) a delimitação dos seus campos de atuação; (c) a determinação de uso social para a IA; (d) o trabalho compulsório em certa tarefa; ou, até mesmo, (e) o desligamento da tecnologia". A.L. de Paula *et al.*, *Breves reflexões sobre a inteligência artificial e seus impactos no campo do Direito Penal*, cit., p. 112.

qualquer outra via legal. A responsabilidade criminal não deve ser dividida, mas sim adicionada. A responsabilidade criminal da entidade de IA é imposta além da responsabilidade criminal do programador ou usuário humano[26]."

Na maioria dos países do ocidente, que adotando as instituições construídas a partir do Princípio da Legalidade, com o consequente estágio atual do conceito tripartido do crime e seus requisitos, não incriminam as escolhas da inteligência artificial, ainda que elas realizem lesões de bens jurídicos. Por isso no caso paradigma, Tay não poderia, no estágio atual da legislação e da ciência penal, não tem responsabilidade penal.

A base desse conceito, desde a sua fundação até o tempo hodierno, exige que seja produzida uma modificação do mundo exterior dominável ou dominada pela vontade. Tal não ocorre com os resultados de danos penalmente relevantes, produto de escolhas autorreferentes algorítmicas, vez que embora possam controlar processos que modifiquem potencialmente a realidade exterior, sobretudo se a IA comandar robôs a ela vinculados. A última instituição do crime, nomeadamente a culpabilidade, que se define como o juízo de reprovação realizado sobre o autor do injusto (conduta típica e antijurídica, porque, podendo se comportar conforme o Direito, optou por se comportar contrário ao Direito, é o principal óbice a essa responsabilização.

---

[26] "*When an AI entity establishes all elements of a specific offense, both external and internal, there is no reason to prevent imposition of criminal liability upon it for that offense. The criminal liability of an AI entity does not replace the criminal liability of the programmers or the users, if criminal liability is imposed on the programmers and/or users by any other legal path. Criminal liability is not to be divided, but rather, added. The criminal liability of the AI entity is imposed in addition to the criminal liability of the human programmer or user*". Ver G. Hallevy, *The Criminal Liability of Artificial Intelligence Entities*, cit., p. 29.