Arriscado, mas útil: O novo desafio do Direito Penal na Era dos Sistemas de IA Generativa

Risky but Useful: The New Challenge of Criminal Law in the Age of Generative AI Systems

Andrea R. Castaldo
Professor Catedrático de Direito Penal na Universidade de Salerno e advogado, em Nápoles e em Munique, exercendo a sua atividade no Supremo Tribunal. É Presidente do Osservatorio Giuridico sulla Criminalità Economica (O.G.C.E.) (Observatório Jurídico da Criminalidade Económica) que fundou na Universidade de Salerno, em colaboração com o C.E.R.A.D.I. - Luiss "Guido Carli".

Fabio Coppola
Advogado criminalista em Salerno e fundador do Studio Legale Coppola, colabora profissionalmente com o Studio Legale Avv. Prof. Andrea R. Castaldo. Professor adjunto de Anticorrupção na Escola de Estudos Avançados em Anticorrupção e Contratação na Administração Pública e de Direito e Processo Penal e Funções e Técnicas de Polícia Judiciária na Escola Regional de Polícia.

1. "Quem tem medo do lobo mau?"[1]

Todas as inovações são intimidadoras de primeira. Aconteceu no advento da energia elétrica, com a internet, e agora com a ascensão dos sistemas de inteligência artificial; em especial os chamados "fortes", os quais são capazes de forma autônoma e independente emular os processos decisórios humanos através de sistemas de aprendizagem automáticas.[2]

Será o medo do desconhecido o bastante para virar as costas aos avanços tecnológicos?

A pergunta é retórica, pois os sistemas de IA são arriscados, mas certamente úteis à sociedade.

Basta considerar, por exemplo, o uso que fazemos de dispositivos digitais para checar a previsão de temperatura para o dia, ou para achar a melhor rota para chegar ao trabalho; para trabalhar remotamente, para se comunicar com amigos e conhecidos. Em adição podemos considerar os usos mais sofisticados de IA no mercado financeiro[3], em estratégias de marketing, e na assistência/substituição de

---

[1] For a more detailed discussion, please refer to the recent A. R. CASTALDO, "*IA: chi ha paura del lupo cattivo?*", currently in press in the proceedings from the international congress "*Scuola, Università e Ricerca: diritti, doveri e democrazia nello stato di cultura*", Cava dei Tirreni (Salerno), 2 Dec. 2023.

[2] For a detailed analysis and definition, please refer to the canonical S. J. RUSSEL–P. NORVIG, *Artificial Intelligence. A modern approach*, 4th edition, Global Edition, 2022, 1032 ff. On the same subject, see the extensive F. BASILE, *Intelligenza artificiale e diritto penale: qualche aggiornamento e qualche nuova riflessione*, in *Diritto penale e intelligenza artificiale. "Nuovi Scenari"*, G. BALBI-F. DE SIMONE-A. ESPOSITO-S. MANACORDA (eds), Giappichelli, 2022, 4, B. FRAGASSO, *La responsabilità penale del produttore di sistemi di intelligenza artificiale*, in *Sistema penale*, 13 June 2023, 6-8 and V. MANES, *L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia*, in *disCrimen*, 15 May 2020, 1-22.

[3] Consulte as reflexões perspicazes em in F. CONSULICH, *Flash offenders. Le prospettive di accountability penale nel contrasto alle intelligenze artificiali devianti*, in *Riv. it. dir. proc. pen.*, 3/2022, 1015 ff.

humanos em certas profissões[4].

Então, como já é reconhecido, a questão relevante não é sobre aceitar ou rejeitar um fenômeno já estabelecido, mas como regular atividades perigosas/arriscadas.

Por esses motivos, não abordaremos a questão de crimes intencionalmente cometidos através de IA nesta análise[5]. Ao invés disso, falaremos exclusivamente do problema de gestão de erros cometidos por sistemas de IA generativos. Em outras palavras, investigamos se - e em que medida - é possível evitar a penalização de eventos nocivos que ocorrem apesar da teórica ideal gestão de riscos de IA, mas que são defeituosos em prática devido a imprevisível e "diabólica" intervenção da máquina[6].

2. O papel do direito criminal na administração de riscos da IA

Regular a intervenção do direito criminal perante os novos riscos tecnológicos não é uma tarefa fácil[7] já que a imprevisibilidade decisional dos IA generativos é uma razão para preocupação prevalente. É impossível prever todos os possíveis erros que podem ocorrer, porque eles são resultado de uma combinação massiva de dados que vai além da compreensão humana. Na melhor das hipóteses, é razoável assumir que a inteligência artificial comete erros eventualmente. Sobre isso, podemos falar do "previsível do imprevisível"[8].

Nas áreas altamente sensíveis, tais como o trânsito com a introdução de veículos autônomos[9], cirurgias executadas por robôs, o uso de algoritmos em investimentos financeiros, o uso de chatbots na reaquisição de informação ou para a criação de trabalhos intelectuais ou artísticos, sem mencionar robótica industrial, é necessário para decidir se precisamos tolerar ou proibir sistemas de IA imprevisíveis.

Na nossa opinião, os seguintes aspectos precisam ser levados em consideração quando esta decisão for feita.

Por um lado, se os riscos da IA são excessivamente criminalizados, isso pode levar a um efeito desestimulante no desenvolvimento tecnológico. Empresas seriam hesitosas em investir e testar ferramentas de IA avançadas, já que seriam confrontadas com acusações criminais se produtos imprevisíveis e potencialmente perigosos fossem introduzidos ao mercado. Uma decisão desta espécie consequentemente traria outro problema: sacrificar a evolução de sistemas de IA significa abrir mão de uma ferramenta efetiva para promover direitos fundamentais. Tecnologia certamente pode ajudar na redução de acidentes no tráfego, ferimentos em ambiente de trabalho, e na poluição do meio-ambiente. Ela pode também aumentar o acesso aos serviços de saúde através da telemedicina e educação remota

[4] Na Itália, o Governo está intervindo com normas urgentes para regular o uso de sistemas de IA nos setores mais sensíveis. Cf., R. AMORUSO-F. PACIFICO, *Intelligenza artificiale, c'è il decreto: stretta su copyright e tutele agli under 14*, in *Il Mattino*, 10 aprile 2024, 7.
[5] Sobre esse ponto, consulte F. CONSULICH, *op. cit*., 1033-1035.
[6] Para usar o termo eficaz cunhado em F. CONSULICH, op. cit., p. 1053.
[7] Digno de releitura, o seminal U. BECK, *La società del rischio. Verso una seconda modernità*, a versão italiana de W. Privitera, Carocci editore, 2013.
[8] Assim, C. PIERGALLINI, *Intelligenza artificiale: da 'mezzo' ad 'autore' del reato?*, in *Riv. It. Dir. Proc. Pen*., 4/2020., 1750.
[9] Consulte o volume monográfico mais recente e exaustivo, M. LANZI, *Self-driving cars e responsabilità penale. La gestione del "rischio stradale" nell'era dell'intelligenza artificiale*, Giappichelli, 2023.

(e mais recentemente, no Metaverso)[10]. Desistir destes avanços significa perder uma ferramenta positiva para o desenvolvimento das pessoas e da sociedade exclusivamente por razões de segurança precatórias.

Se essa abordagem viesse a prevalecer, é fácil de imaginar como seria a definição de novos crimes para aqueles que introduzem novos sistemas de IA que não se conformam com as "melhores práticas atuais"[11]. Entretanto, a identificação destas melhores práticas requer contributo de trabalhadores do setor que possuem conhecimento profundo do seu produto[12]. Mas, como foi mencionado, em razão da imprevisibilidade dos resultados dos IAs generativos, esse conhecimento pode ser incompleto ou inexistente.

Outrossim, fabricantes e programadores podem ser incentivados a ocultar ou esconder qualquer *red flag* que foi descoberta depois do lançamento do sistema de IA no mercado, já que a transparência completa dessas informações expô-los-ia a ofensas penais, as quais foram antemente discutidas.

Por outro lado, renunciar a intervenção do direito criminal em uma área tão delicada traria a mesma quantidade de problemas. Seria irrazoável, implausível, e contrário a direção tomada pela Europa através do AI Act imaginar não utilizar a força simbólica do direito criminal para proteger "interesses públicos" e "direitos fundamentais" contra decisões imprudentes e ousadas, tomadas como resultado da lógica whatever it takes do progresso[13].

Na nossa perspectiva, precisamos ser pragmáticos e achar uma maneira de gerenciar riscos vindos de IA, que leva em consideração os dois lados mencionados.

3. Qual instrumento penal pode ser usado para manejar riscos resultantes da inteligência artificial?

Tentemos elaborar uma proposta que traça o papel do direito criminal no manuseio dos riscos postos por IA. Com isto, não abordaremos a perspectiva de acusar sistemas de IA pelos crimes que cometem, pois acreditamos que essa abordagem traz problemas inultrapassáveis, os quais já foram discutidos extensivamente[14].

---

[10] Em 1º de dezembro de 2023, na Universidade de Salerno, organizamos o "MetaCourt", a primeira simulação processual no Metaverso, envolvendo estudantes do Departamento de Ciências Jurídicas em um "desafio" altamente educacional e interativo. Usando fones de ouvido de RV, eles simularam os papéis do promotor público, da defesa e do júri em um tribunal virtual e discutiram um caso jurídico hipotético.

[11] Veja L. PICOTTI, *Intelligenza artificiale e diritto penale: le sfide ad alcune categorie tradizionali*, in *Diritto penale e processo*, 3/2024, 299.

[12] Sobre esses tópicos, veja especialmente G. FORTI, *"Accesso" alle informazioni sul rischio e responsabilità: una lettura del principio di precauzione*, in *Criminalia*, 2006, 155-225 e o trabalho monográfico C. PIERGALLINI, *Danno da prodotto e responsabilità penale. Profili dommatici e politico-criminali*, Giuffrè, 2004.

[13] O Regulamento Europeu sobre Inteligência Artificial atualizado para a Resolução Legislativa do Parlamento Europeu de 13 de março de 2024, sobre a proposta de regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas sobre inteligência artificial (Lei de Inteligência Artificial) e e que altera determinados atos legislativos da União (COM(2021)0206 - C9-0146/2021 - 2021/0106(COD)) pode ser consultado no seguinte link: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_IT.pdf; Para uma análise inicial das atualizações aprovadas, consulte G. M. RICCIO-P. GENTILI, *AI Act, ok dal Parlamento UE: ecco cosa prevede il testo approvato*, in *Il Quotidiano Giuridico*, 14 March 2024; for an anlysis of the *AI Act* from the criminal law perspective, see. M. COLACURCI, *Quale diritto penale dell'IA*? In *Jus*, 3/2023, 357-378.

[14] On criminal liability concerning AI, see. G. HALLEVY, *Liability for Crimes Involving Artificial Intelligence Systems*, Springer, 2015, 1-229. We echo the critical considerations of A. CAPPELLINI, Machina delinquere non potest? *Brevi appunti*

Primeiramente, precisamos considerar a imprevisibilidade dos erros da inteligência artificial.

É plausível considerar que uma possível morte de um acidente no tráfego causada por um veículo autónomo não pode ser atribuída ao fabricador/programador, porque seria considerado um erro imprevisível e, consequentemente, inevitável?

Nessa linha de raciocínio, o fabricador/programador poderia ser responsável somente pelos riscos previsíveis, enquanto os imprevisíveis seriam "tolerados" pela sociedade.

Mesmo essa solução sendo promissora, acreditamos que ela tem diversos defeitos. Primeiramente, ela não seria válida em todos os sistemas de direito penal, incluindo os europeus. De fato, alguns países criminalizam a mera introdução de produtos imprevisivelmente perigosos no mercado, mesmo se nenhum dano foi causado por eles.

Segundamente, quem decide se o erro foi previsível ou não? Teria de ser decidido pelos tribunais, caso a caso. Entretanto, nas áreas em que o conhecimento científico é insuficiente, essa avaliação se torna *probratio diabolica*, e só pode ser solucionada com a atribuição da culpa na parte responsável por "não ter feito todo o necessário para prevenir os danos"[15]. Mais uma vez, o único jeito de evitar responsabilidade criminal é prevenir o dano de acontecer. Isso é impossível depois do fato ter acontecido, pois não temos como voltar no passado… pelo menos, por enquanto!

Uma situação relacionada a essas circunstâncias aconteceu durante a pandemia do COVID-19. Em resumo: na Itália, leis tiveram que ser introduzidas para clarificar as obrigações de segurança que os empregadores deveriam seguir para a prevenção da transmissão do vírus; estas serviram para tranquilizar as empresas de que elas não sofreram acusações por "não terem feito tudo possível" se algum dano ocorresse[16].

Por essas razões, a avaliação da previsibilidade/imprevisibilidade caso a caso não parece abranger os riscos dos sistemas generativos de IA uniformemente em todos os sistemas de direito criminal.

4. Como podemos determinar o limiar de risco aceitável para IA a priori?

Comecemos pelas poucas certezas que temos: sabemos que sistemas avançados de IA irão mais e mais fazer decisões por eles mesmos, e nem sempre estes sistemas nos darão as razões por trás da black box de redes computacionais que levaram ao resultado decisório[17].

O espaço entre a informação dada para a máquina e o seu resultado autônomo só nos leva a

su intelligenza artificiale e responsabilità penale, in *disCrimen*, 27 March 2019, 1-23 and C. PIERGALLINI, *op. cit.*, 1763-1771.

[15] Kindly refer to some earlier reflections found in A. R. CASTALDO, *La teoria dell'aumento del rischio e l'illecito colposo*, in *Studi Urbinati, A - Scienze Giuridiche, Politiche Ed Economiche, 40 (1988)*, 2021, esp. 127-131.

[16] On the flexibility of preventive rules in the *corpus* of Legislative Decree No 81/2008 and the application of living law, please refer to the exhaustive C. CUPELLI, *Obblighi datoriali di tutela contro il rischio di contagio da Covid-19: un reale ridimensionamento della colpa penale?*, in *Sistema penale*, 15 June 2020. For a rigorous application of the precautions imposed on employers to prevent contagion, please refer to *Cass. Pen., Sez. III, 1° dicembre 2023, 47904*. On the liability of healthcare workers during the pandemic emergency, please refer to M. CAPUTO, *Logiche e modi dell'esenzione da responsabilità penale per chi decide e opera in contesti di emergenza sanitaria*, in *La legislazione penale*, 22 June 2020, esp. 9 ff.

[17] Advocating for full transparency, although he considers it "far from being achieved and perhaps (...) an impossible result", see F. CONSULICH, *op. cit.*, 1027-1028.

conclusão de que haverá algum erro, mas não nos ajuda a estabelecer *como* ou *por quê*.

Outra certeza que temos é da utilidade social dos sistemas generativos de IA. Isso pressupõe que existe um rácio positivo de custo/benefício - esperamos em sociedade que as máquinas tenham um desempenho maior que o nosso e reduzam a nossa margem de erro. Queremos que os veículos autônomos causem menos acidentes do que motoristas humanos, e cirurgias realizadas por robôs resultem em menos mortes ou sequelas do que as realizadas por seres humanos.

O desafio, entretanto, é identificar um limiar que é absolutamente intolerante a máquinas que são mais perigosas do que humanos. Ou seja, não podemos deixar que IA proponha mais riscos do que humanos executando a mesma ação. A tarefa de uma possível reforma da lei penal que pode ser exportada mundialmente e oferecer uma avaliação uniforme e previsível do risco de IA, é estabelecer, antecipadamente, o percentual de eventos prejudiciais que são considerados socialmente aceitáveis em relação à IA. Esse percentual deve ser geralmente inferior ao número de eventos negativos causados por humanos realizando a mesma atividade 18. A fórmula que propomos para guiar legisladores nacionais regulando o risco concedido para máquinas é: % Eμ < Eα, where "Eμ" significa Erro Imprevisível da máquina e "Eα" significa erro humano. A tolerância de um risco desconhecido vindo de sistemas de IA "Ɵ" sempre pressupõe um rácio menor de eventos que causam danos por máquinas do que por humanos.

Retornemos ao exemplo de operações cirúrgicas: se operações do mesmo tipo executadas por humanos têm uma taxa fatal de 20% (Eα = 20%), a zona de risco permissível para as máquinas necessariamente precisa cair abaixo desse limiar, pois esperamos que as máquinas devem entregar melhores e mais seguras execuções do que a humana, se formos aceitar os seus riscos.

Vamos supor que, hipoteticamente, fixamos um limiar de tolerância de "Ɵ Eμ = 10%" (tolerância de erro das máquinas é de 10%).

Na ausência de dados suficientes para determinar se um evento danoso foi causado por defeito na construção ou programação do sistema de IA, essa estratégia nos permite preventivamente abordar essa questão de responsabilidade criminal para erros imprevisíveis que vamos discutir brevemente.

Antes disso, vamos abordar a pergunta: é justificável que o Estado renuncie o seu direito de punir eventos danosos que caem dentro do determinado limiar de tolerância?

Na nossa perspectiva, qualquer abordagem objetivando uma maior previsibilidade na aplicação deveria hipotéticamente envolver uma renúncia total à intervenção do direito criminal em eventos danosos ou fatais que ocorrem dentro do limiar estabelecido de tolerância. As categorias tradicionais de responsabilidade penal retomariam o seu funcionamento para além da decisão política de tolerar os acontecimentos prejudiciais resultantes de um erro imprevisível por parte de um homem dentro do limite pré-determinado.

Esta abordagem garantirá imediatamente uma maior segurança no que respeita ao risco penal admissível.

[18] The need to establish a threshold of "acceptable risk" is also discussed in B. FRAGASSO, *op. cit*., 17, referencing G. Forti on “the marginal rick (…) not to be defused”.

5. Como distribuir a percentagem de tolerância social dos riscos de IA pelos diferentes operadores?

Nossa proposta está incompleta no momento. Nós ainda temos que debater e explorar o problema de distribuir o limiar de tolerância de risco "Θ Eμ" dentro das dinâmicas e variedades de casos específicos, para que assim fabricadores ou programadores individuais possam prever a margem de erro tolerada para o desenvolvimento de IA generativo nos seus respectivos campos de trabalho.

Para alcançar isso, podemos usar provisões do European AI Act, pois impõem requisitos de transparência e comunicação aos produtores de sistemas de IA generativo que operam dentro da Europa[19]. Com isto, poderemos criar um registro para operadores do campo de IA e o seu setor de desenvolvimento, o que consequentemente facilitará o seguinte cálculo.

Uma advertência: por fins ilustrativos, propomos um cálculo padronizado para todos os fabricantes/programadores do setor. Nada nos impede de revisitar este ponto para melhor definir os detalhes, depois de uma elaboração minuciosa.

Uma vez que é estabelecido o nível de risco que o público está disposto a aceitar para um risco imprevisível de IA "Θ Eμ" para um determinado setor de desenvolvimento "Ση", este deve ser traduzido em termos numéricos e o resultado deve ser distribuído pelo número de produtores reconhecidos em conformidade com as obrigações previstas no European AI Act.

Vamos assumir o limiar hipotetizado previamente de 10%, e 500 produtores de IAs generativos no setor. O raciocínio é o seguinte:

i. Os dados estatísticos conhecidos relativos ao risco humano para a mesma atividade, denotados por "**%Eα**" de "**Ση**", são de 20%. Suponhamos que isto corresponde estatisticamente a 2000 incidentes individuais danosos "**X**". Z. Por conseguinte, **Eα deΣη = 2000 X**.

ii. Para IA, o percentual de *risco desconhecido* a ser distribuído é de 10%, que seria metade de **2000 X**: **Θ Eμ de Ση = 1000 X**.

iii. O resultado alcançado deve ser dividido pelo número de produtores reconhecidos no setor. "Πs", que assumimos ser 500. Assim, teremos **1000X: 500Πs** = iv. O valor "**Θ Eμ**" é igual a **2X** para cada **Πs** individual.

O cálculo, como já mencionado, é aproximado e simplificado. Ele apenas destaca algumas lacunas das quais estamos cientes, e é necessário estabelecer se o percentual tolerado deve ser entendido como relativo a cada ano de referência estatística ou a toda vida de produção/programação. Ao distribuirmos a margem de tolerância, deve-se considerar o impacto variável dos operadores dentro do mercado. Deste ponto de vista, diferentes *clusters* de operadores do setor devem ser identificados e divididos em classes de referência uniformes. Consequentemente, o limiar de tolerância de risco deve ser distribuído uniformemente dentro do cluster mas diferentemente entre os grupos, priorizando aqueles que mostram maior produtividade.

O que deve ser feito com o limiar individual de tolerância obtido dessa maneira?

[19] Cf. A. LONGO, *AI Act, ecco il testo: così l'Europa regola l'intelligenza artificiale generativa*, in *Il Sole 24 Ore*, 8 Feb. 2024.

Apesar das aproximações destacadas, o cálculo proposto pode permitir que cada produtor/programador tivesse acesso a duas margens de erros imprevisíveis para testar e desenvolver seus sistemas generativos de IA sem incorrer em responsabilidade criminal pelos eventos resultantes. A decisão de não processar um certo número de casos relativos a um "evento adverso" permite, em linha com a premissa de **%Eμ < Eα** e com a imprevisibilidade de erros, um equilíbrio entre as necessidades relacionadas a seguranças de interesses fundamentais e as de avanço tecnológico. De fato, o produtor ou programador poderá contar com esse benefício, ou "bônus" para desenvolver IA com maior tranquilidade. Ao mesmo tempo, poderá compartilhar informações com a comunidade sobre os erros encontrados, o que ajudará outros produtores e programadores para os quais os erros se tornarão previsíveis. Ao adotar a lógica de "só se aprende fazendo" , serão os próprios operadores dos sistemas de IA que realmente reduzirão o espectro de erros que não são puníveis por serem imprevisíveis.

O Estado poderia centralmente administrar o controle e registrar os erros gerados por IA, medidas para serem "ensinadas" às máquinas, e a disseminação dentro da comunidade dos novos erros "conhecidos" que devem ser prevenidos, junto de medidas de segurança atualizadas, através de uma Agência para o Controle de Correção de Sistemas de IA. Esta entidade deve ter o poder de impor penalidades adicionais no produtor ou programador de algum sistema de IA, como por exemplo, condicionando a possibilidade de um benefício para o produtor ou programador ao seu compromisso com corrigir erros imprevistos, transformando-os em erros previsíveis. Essa lógica é muito parecida com a conhecida abordagem de carrot and stick[20] Como podemos adaptar nossa fórmula aos diferentes sistemas de direito penal?

A primeira opção é estabelecer uma área de irrelevância criminal que engloba qualquer evento danoso resultante de um erro imprevisível de sistemas de IA que se encontra dentro do limiar de tolerância de risco predeterminado. O espaço na proteção então criado pode ser tampado com medidas de compensação monetária para danos, possivelmente através da criação de um fundo "distribuído *pro rata* por todos os produtores que projetam, programam, e usam IA, baseado no tamanho da sua parcela no mercado"[21]. A distinção entre erros previsíveis e imprevisíveis cometidos por IA iria desencorajar condutas excessivamente relaxadas dentre os agentes do mercado. Eles então seriam responsáveis por todos os erros que poderiam ter sido previstos de início, assim como por aqueles que se tornam previsíveis basicamente na experiência coletiva de outros produtores e programadores.

A fórmula descrita acima tem outra possível aplicação. Ela se tornaria o padrão para garantir a adequação e segurança dos sistemas de IA, e a conformidade seria um requisito para "certificar" a confiabilidade dos sistemas de IA em circulação. Em outras palavras, antes de lançar o sistema de IA generativa no mercado, o produtor ou programador precisaria realizar uma série de testes para identificar a margem de risco que se enquadra dentro do limite de tolerância.

A conformidade dos produtores e programadores com o padrão requerido deve ser considerado como o "melhor comportamento" para conter riscos, o que deve implicar a não-criminalização de quaisquer eventos que possam ocorrer apesar da conduta impecável. A nossa abordagem oferece a vantagem de possibilitar uma estimativa preditiva e objetivamente mensurável do esforço necessário, em contraste com a avaliação de seu comportamento legal caso a caso.

---

[20] Cf. J. C. JR. COFFEE, *"Carrot and Stick" Sentencing: Structuring Incentives for Organizational Defendants*, in *3 Fed. Sent'g Rep*., 1990, 126 ff.
[21] This point of view is examined in F. Consulich, *op. cit.*, 1022.

6. Os prós e contras da solução proposta

A proposta está sujeita a algumas objeções.

Primeiramente, como podemos justificar a renúncia de responsabilidade criminal para eventos sérios, potencialmente danosos ou até fatais?

Esta objeção é válida, mas não é incomum identificar áreas de impunidade. Na Itália isso já aconteceu em outros setores, como por exemplo o medicinal, onde a proteção ao paciente teve que ser balanceada com a necessidade de permitir o bom desempenho dos profissionais de saúde. Como resultado, a responsabilidade médica foi limitada apenas a erros graves ou violações das diretrizes. Da mesma forma, no campo tributário, a legislação penal não pune alguns casos de evasão fiscal que estejam abaixo de um certo limite, devido a complexidade e obscuridade das leis fiscais.

Em relação aos produtores e programadores de sistemas de IA, a impunidade de erros imprevisíveis individuais seria justificada pelo benefício maior para a comunidade, dado que os riscos de usar IA são menores em comparação com humanos realizando as mesmas tarefas. Essencialmente, envolveria "deixar de punir um erro" para "evitar cem". Outra possível objeção diz respeito à potencial distinção entre a responsabilidade criminal de um ser humano e a responsabilidade criminal de uma máquina.

Pode-se argumentar que é irracional esperar um tratamento diferente para erros cometidos por humanos e por IA programadas por humanos. No primeiro caso, a responsabilidade é avaliada individualmente. No segundo, pode-se prever um benefício uniforme e predeterminado.

Em nossa opinião, o comportamento das máquinas tende a aderir a um padrão, ao contrário da inclinação individual e original da ação humana. Isso justifica o tratamento diferenciado de agentes virtuais e a sua categorização em *clusters* com base em seu comportamento. Quanto à responsabilidade de entidades por delitos criminais, apesar das abordagens diferentes entre os sistemas jurídicos, acreditamos que, se o indivíduo que programou ou criou a IA prejudicial não for responsabilizado por um delito graças ao "sistema de benefícios", o mesmo deve se aplicar à entidade. Em tais casos, podem ser consideradas formas de compensação por danos fora da esfera do direito penal.

Por fim, passemos a um possível ponto positivo.

Quanto à atribuição de responsabilidade criminal dentro de organizações complexas que precisam produzir, desenvolver e programar sistemas de IA[22], acreditamos que, uma vez que o mundo da IA seja tranquilizado pelo sistema de benefícios, será mais fácil defender a criação da figura do supervisor de IA. Essa figura seria responsável por monitorar, gerenciar e garantir a segurança dos sistemas de IA generativos, identificando assim um ponto focal para formas residuais de responsabilidade.

Ainda há muito a discutir, e não afirmamos que as sugestões anteriores sejam exaustivas. No entanto, sentimos que era necessário explorar abordagens inovadoras para gerenciar os riscos imprevisíveis associados à IA através de algo melhor do que a penalidade usual[23].

---

[22] Commonly referred to as the 'many hands problem'; for some thoughtful observations on the topic, see F. CONSULICH, *op. cit*., 1040.

[23] On this subject, we refer to the profound insights of L. Eusebi, *Qualcosa di meglio della pena retributiva in margine a C. E. Paliero, Il Mercato della Penalità*, in *Studi in onore di Carlo Enrico Paliero*, I, C. Piergallini-G. Mannozzi- -C. Sotis-C. Perini-M. Scoletta-F. Consulich (eds), Giuffrè, 2022, 388 ff.