ÉTICA DA INTELIGÊNCIA ARTIFICIAL

MÁRIO GASPAR DA SILVA – INESC-ID, Instituto Superior Técnico, Universidade de Lisboa

O conceito de Inteligência Artificial (IA) evoluiu radicalmente desde que foi proposto há quase 70 anos. No presente, é conotado com a tecnologia usada em sistemas informáticos capazes de se auto-programar a partir de dados obtidos anteriormente («aprender automaticamente») de forma a produzir modelos de inferência capazes de responder aos pedidos de informação mais variados.

Nos próximos meses assistiremos ao lançamento no mercado de consumo dos primeiros produtos de IA generativa, capazes de gerar novos conteúdos, ideias ou soluções. Esse tipo de IA inclui ferramentas como o *ChatGPT*, máquinas (ou modelos) de geração de imagens, modelos de escrita de código e outros que produzem trabalho criativo ou cognitivo. Antecipa-se uma rápida disseminação à escala global destas ferramentas e modelos de geração, com redução drástica de custos e aumento de sofisticação no médio prazo.

A transição não será instantânea, mas os efeitos da introdução de sistemas e ferramentas baseados em IA terão impacto comparável ao que observámos no passado com outras tecnologias, desde a máquina a vapor ao computador, ou, só nos últimos 50 anos na área das TIC, o PC, a Internet/Web, *smartphones* e computação em nuvem. Em cada uma destas transformações, se por um lado nascem novas profissões até aí inexistentes, por outro a automação faz desaparecer várias outras até aí desempenhadas por humanos, e altera profundamente a forma de trabalhar e a nossa organização social, trazendo ganhos de produtividade.

No presente, a substituição de empregos pela IA refere-se ao processo

A IA tem o potencial de criar desigualdades em recursos, oportunidades e poder no mundo empresarial e pode perpetuar injustiças históricas.

em que tarefas hoje realizadas por humanos são cada vez mais automatizadas ou executadas por sistemas de IA Generativa. Tarefas de rotina correntes, como responder a *e-mails*, gerar relatórios e resumir dados, que exigem pensamento estruturado, mas não necessariamente criatividade, estão entre as primeiras a ser automatizadas. A novidade da IA Generativa face a processos anteriores de automação reside, por outro lado, em realizar tarefas antes consideradas não automatizáveis, como criação de conteúdo (escrever artigos, gerar arte, criar música), desenvolvimento de *software* e design, pesquisas jurídicas iniciais ou diagnósticos médicos.

Antecipa-se que, tal como com a introdução dos *smartphones*, o impacto desta tecnologia seja transversal a todas as profissões, com efeito particularmente visível em setores como a comunicação social, serviços de apoio ao cliente, saúde, direito ou programação informática. Por outro lado, além da substituição e capacitação de empregos existentes, haverá também lugar à criação de novos empregos, desde os relacionados com a preparação de dados para treino de modelos de IA, aos engenheiros de *prompts* e especialistas em ética de IA.

A IA tem o potencial de criar desigualdades em recursos, oportunidades e poder no mundo empresarial e pode perpetuar injustiças históricas. As empresas em regiões desenvolvidas com uma infraestrutura digital robusta têm mais probabilidades de beneficiar das tecnologias de IA do que aquelas onde o acesso à tecnologia e às competências digitais é limitado (a chamada divisão digital). Ao mesmo tempo, os Estados mais desenvolvidos

e as grandes empresas globais neste domínio dispõem de recursos financeiros, conhecimento técnico e acesso aos dados necessários para desenvolver e implementar sistemas avançados de IA, enquanto os Estados mais desfavorecidos e as PME encontrarão maiores dificuldades em aceder à tecnologia de IA.

Também se antecipam desigualdades no acesso a oportunidades de emprego. Os algoritmos de IA utilizados em processos de recrutamento podem, inadvertidamente, perpetuar preconceitos. Se os dados usados para treinar esses algoritmos refletirem preconceitos históricos (por exemplo, favorecendo candidatos de certos grupos demográficos), a IA pode continuar a discriminar grupos marginalizados, limitando as suas oportunidades. Sistemas de classificação de crédito baseados em IA poderão vir a prejudicar indivíduos de comunidades historicamente marginalizadas, se os dados existentes refletirem preconceitos. Por exemplo, a IA pode correlacionar códigos postais ou percursos educativos com competências procuradas e vir por esse meio reforçar as disparidades existentes.

A procura por competências relacionadas com IA está a crescer na área da educação e formação profissional, mas as condições de acesso nestas áreas são desiguais. Aqueles com acesso a uma educação avançada e formação técnica terão acesso a melhores oportunidades, enquanto os que não têm ficam para trás, acentuando desigualdades já hoje acentuadas nas competências digitais. A automação impulsionada pela IA pode afetar desproporcionalmente os empregos de baixa remuneração e baixa qualificação, que são mais frequentemente ocupados por indivíduos de grupos historicamente marginalizados. Isto pode agravar as disparidades salariais existentes e limitar a mobilidade económica desses trabalhadores. A IA abrirá novas oportunidades económicas, mas se o acesso a estas for desigual, pode acentuar desigualdades existentes.

Haverá também um acentuar de desigualdades entre empresas. As que possuem ou controlam tecnologias de IA avançadas poderão vir a dominar mercados, estabelecendo as condições da concorrência e, potencialmente, sufocando a inovação. Esta concentração de poder pode levar a práticas monopolistas e reduzir a competitividade das pequenas empresas. O valor da IA é impulsionado largamente pela abundância dos dados. As empresas que controlam grandes conjuntos de dados disporão de uma vantagem significativa, o que nos pode levar a uma situação em que um pequeno grupo de gigantes tecnológicos detém um poder desproporcional sobre a economia e a sociedade.

Assistiremos também a mudanças nos processos e paradigmas de governação das empresas e organismos públicos, onde os sistemas de IA passarão a ser cada vez mais utilizados nos processos de tomada de decisão. Estes sistemas permitem centralizar o poder de decisão em diretores executivos e cientistas de dados que disporão de meios para influenciar políticas e regulamentações a seu favor, reforçando ainda mais o seu poder e erguendo barreiras aos concorrentes. Podem assim vir a dominar mercados, estabelecendo as condições da concorrência e sufocando a inovação através de práticas monopolistas. O valor dos sistemas de IA é largamente determinado pelos dados, pelo que as empresas que controlam grandes conjuntos de dados disporão de uma vantagem que reforçará o seu poder e influência. Tal poderá levar a que um pequeno grupo de gigantes tecnológicos venha a deter um poder ainda mais desproporcional sobre a sociedade.

O desenvolvimento e implementação de tecnologias de IA são frequentemente controlados por um grupo relativamente homogéneo de indivíduos, tipicamente provenientes de contextos mais privilegiados. Isto pode resultar em sistemas de IA que não consideram as necessidades e perspetivas de comunidades marginalizadas, reforçando ainda mais as desigualdades sociais e económicas. A falta de diversidade significa que o desenvolvimento de sistemas de IA poderá não refletir a diversidade das populações que se destinam a servir. Isto pode levar a aplicações de IA que, inadvertidamente, excluem ou prejudicam certos grupos. No Instituto Superior Técnico, por exemplo, a preferência por mulheres de cursos com ensino mais aprofundado em inteligência artificial e ciência de dados continua reduzida (aproximadamente 20%). Uma tal assimetria poderá levar a que este grupo, apesar de maioritário, possa, tal como no passado, ser

A falta de diversidade significa que o desenvolvimento de sistemas de IA poderá não refletir a diversidade das populações que se destinam a servir.

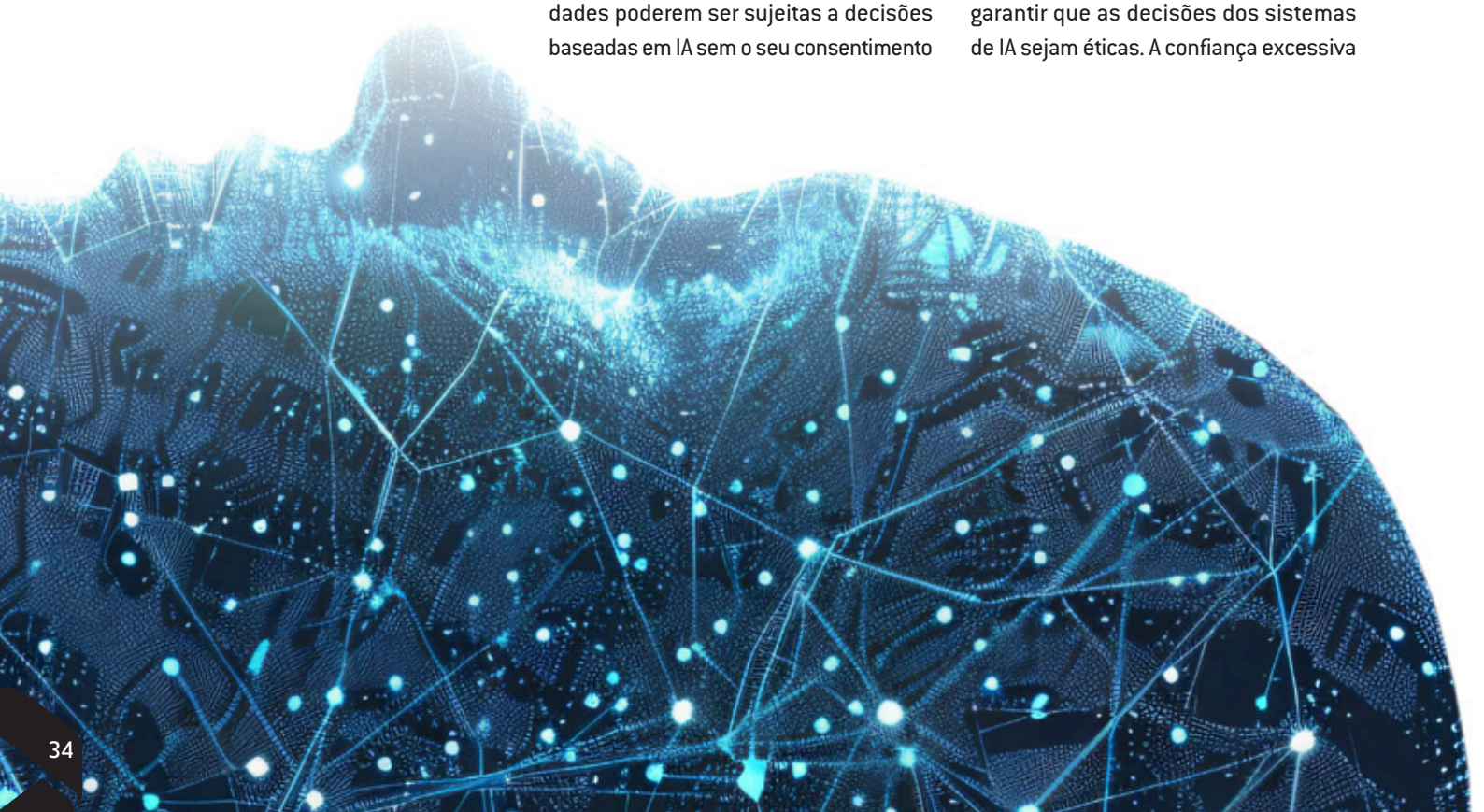
desproporcionalmente afetado. Da mesma forma, como a conceção dos sistemas de IA reflete necessariamente valores e normas daqueles que os desenvolvem, poderemos assistir à imposição de tecnologias culturalmente insensíveis noutras sociedades, exacerbando injustiças globais.

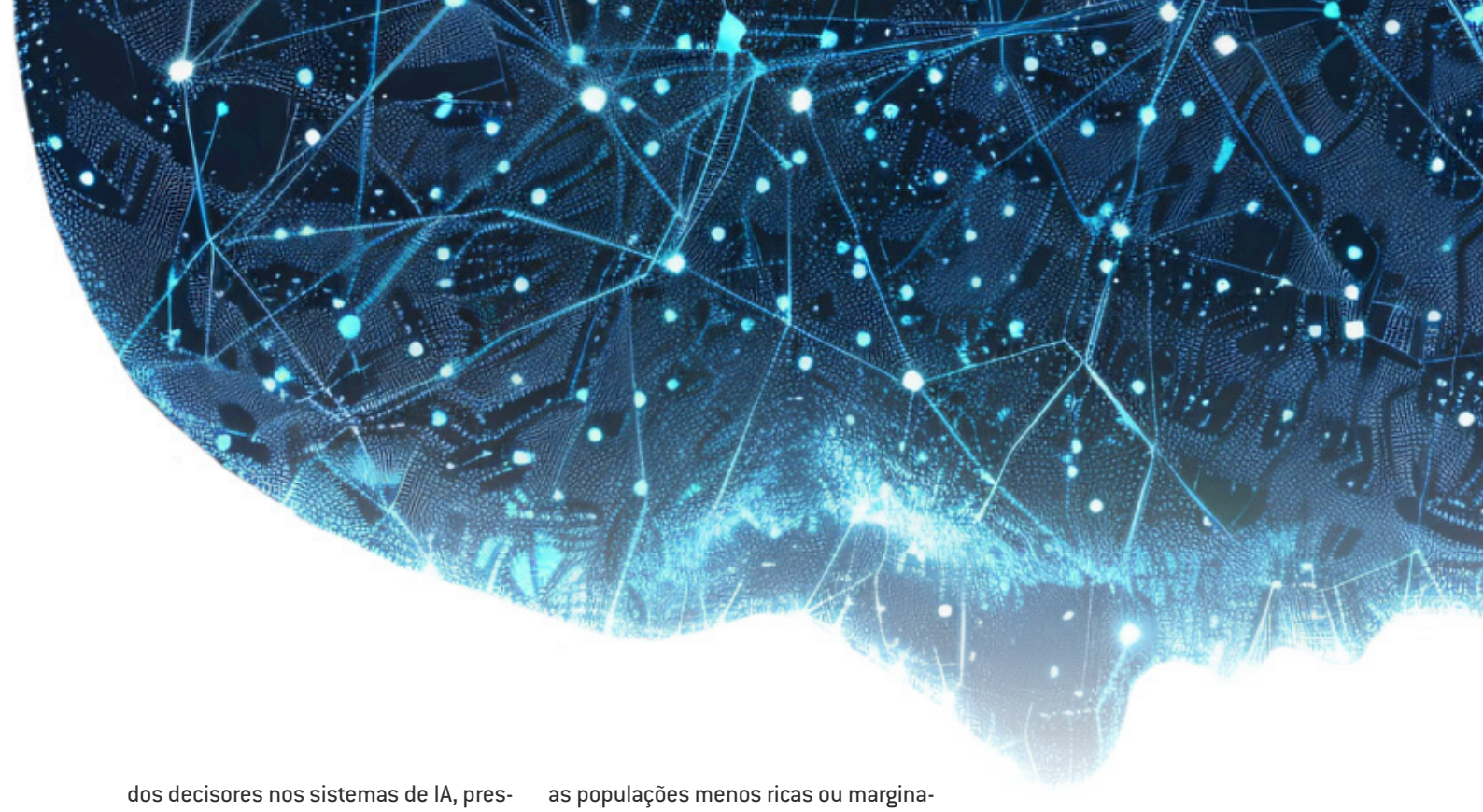
No que respeita à governação global, os países têm leis e regulamentos diversos sobre a IA, o que pode levar a inconsistências nas normas éticas. A União Europeia aprovou recentemente o *AI Act*, a primeira lei compreensiva que estabelece regras em função do risco dos sistemas de IA e obrigações quanto à transparência das decisões baseadas em sistemas de IA. Porém, noutros grandes espaços económicos, enquadramentos semelhantes são apresentados apenas como recomendações e boas práticas. O desafio de aplicar práticas de IA ética a nível global reside em soberanias nacionais e diferentes níveis de capacidade regulatória poderem levar a uma definição e aplicação desigual das normas de IA. Colocam-se então questões éticas relacionadas com o acesso equitativo, podendo as nações mais pobres sofrer consequências negativas, como a perda de empregos ou abusos de vigilância, sem usufruir dos benefícios correspondentes.

As empresas, sendo orientadas para maximizar o lucro, podem entrar em conflito com práticas de IA ética. Muitos sistemas de IA são opacos, no sentido em que o entendimento de como as decisões são tomadas é difícil. Os sistemas de IA dependem frequentemente de grandes volumes de dados, que podem ser obtidos de forma pouco ética, como através de vigilância ou da exploração de mão-de-obra mal paga em países em desenvolvimento para a etiquetagem de dados. Existe o risco de estes processos reforçarem as desigualdades existentes, aumentando o fosso entre ricos e pobres a nível global. Esta falta de transparência pode dificultar a responsabilização dos intervenientes, pelo que garantir práticas éticas ao longo da cadeia de tratamento de dados da IA é um grande desafio. Existe ainda a questão de as comunidades poderem ser sujeitas a decisões baseadas em IA sem o seu consentimento informado, particularmente em contextos onde os sistemas de IA podem ser implementados por governos ou empresas sem consulta pública adequada ou transparência.

No que respeita à tomada de decisões, os sistemas de IA baseiam-se em padrões de dados históricos que podem refletir preconceitos e discriminação de longa data na sociedade, o que pode perpetuar vieses. Em áreas de elevado risco, como a justiça criminal, a saúde e as finanças, o emprego de dados enviesados pode levar a desfechos injustos, especialmente para grupos marginalizados. Por exemplo, os algoritmos de policiamento preditivo que dependem de dados criminais historicamente tendenciosos podem visar desproporcionalmente comunidades marginalizadas. Se os vieses não forem corrigidos, a IA pode também perpetuar e até amplificar injustiças.

Os sistemas de IA baseados em aprendizagem automática podem comportar-se, por vezes, de forma imprevisível em ambientes complexos. Esta imprevisibilidade introduz riscos, tornando difícil garantir que as decisões dos sistemas de IA sejam éticas. A confiança excessiva





dos decisores nos sistemas de IA, pressupondo erradamente a sua infalibilidade ou que a culpa por quaisquer danos será atribuída à tecnologia, poderá levá-los a incorrer em riscos morais e a procurar esquivar-se à responsabilidade.

Hoje, um pequeno grupo de gigantes tecnológicos, nenhum deles localizados na Europa, domina a indústria da IA e concentra o poder económico. Isto pode causar falhas de mercado, onde os benefícios da IA não são amplamente partilhados. Existe uma assimetria significativa de conhecimento entre os detentores das plataformas de IA e os seus utilizadores. Tal poderá causar danos resultantes da incompreensão, pelos últimos, do funcionamento dos sistemas de IA e dos riscos em que incorrem. As falhas de mercado na indústria da IA podem, portanto, vir a exacerbar as desigualdades globais, à medida que a riqueza e o poder se concentram nas mãos de poucas empresas e nações, deixando outras para trás. As tecnologias de IA têm enorme potencial para vir a contribuir com grande impacto para o bem comum em setores como a saúde, educação ou proteção ambiental. Porém, se não forem lucrativas, levantarão questões de justiça global entre as populações menos ricas ou marginalizadas, cujas necessidades não serão atendidas.

CONCLUSÃO

As questões éticas resultantes da massificação das tecnologias de IA colocam-se a propósito da forma como será gerida a transição social, ie, torna-se necessário acautelar efeitos transitórios negativos no processo de reorganização socioeconómica profunda que resultará da introdução da IA. A substituição de empregos por tecnologia de IA levanta muitas preocupações e prenuncia um acentuar dos desafios relacionados com o aumento de desigualdades económicas e éticas na tomada de decisões por máquinas ou com a assistência de máquinas.

Os trabalhadores necessitarão de adquirir novas competências, especializando-se em tarefas que exijam empatia humana, pensamento crítico e criatividade. A substituição de empregos refletirá mudanças nos ambientes de trabalho. Muitas tarefas cognitivas e criativas serão automatizadas, criando tanto oportunidades quanto desafios em diversos setores.

Tendo a IA o potencial de impactar significativamente o mundo empresarial, apresenta também riscos de exacerbar desigualdades existentes em recursos, oportunidades e poder. Se não for cuidadosamente gerida, a IA pode perpetuar injustiças históricas ao reforçar preconceitos sistémicos, aprofundar disparidades económicas e excluir grupos marginalizados dos processos de tomada de decisão. Para mitigar esses riscos, é essencial dar prioridade no desenvolvimento dos sistemas de IA a fatores de justiça, inclusão e transparência.

A implantação de uma nova geração de sistemas informáticos baseados nas tecnologias da IA terá impacto nos domínios da governação e da tomada de decisões sob risco. Estes desafios levantam questões críticas de justiça global, pois podem agravar desigualdades, reforçar vieses e afetar populações vulneráveis em resultado da concentração de poder.

Enfrentar os desafios acima irá requerer esforços concertados para desenvolver e aplicar normas éticas para a IA e promover um desenvolvimento de IA inclusivo e equitativo, permitindo que os seus benefícios venham a ser concedidos a todos.