1

PUBLICUM

Direitos Humanos e Inteligência Artificial: Uma Agenda Urgentemente Necessária[†]

Matthias Risse

Professor de Filosofia e Políticas Públicas. Seu trabalho aborda, principalmente, questões de justiça global que vão desde direitos humanos, desigualdade, tributação, comércio e imigração à mudança climática, obrigações para as futuras gerações e o futuro da tecnologia. Ele também trabalhou em questões de ética, teoria da decisão e filosofia alemã do século XIX, especialmente Nietzsche (em cujo trabalho ele ensina regularmente um seminário de calouros em Harvard). Além disso, leciona na Faculdade de Harvard e na *Harvard Extension School*, e é afiliado ao departamento de filosofia de Harvard. Ele também esteve envolvido com educação executiva em Harvard e em outros lugares do mundo. Risse é o autor de *On Global Justice* e *Global Political Philosophy*, ambos publicados em 2012. E-mail: mathias_risse@harvard.edu

Resumo

A inteligência artificial gera desafios para os direitos humanos. A inviolabilidade da vida humana é a ideia central por trás dos direitos humanos, uma suposição implícita subjacente sendo a superioridade hierárquica da humanidade para outras formas de vida que merecem menos proteção. Essas suposições básicas são questionadas pela chegada antecipada de entidades que não estão vivas de maneiras usuais, mas que, no entanto, são sencientes e intelectual e, talvez, eventualmente, moralmente superiores aos humanos. Certamente, esse cenário pode nunca acontecer e, de qualquer forma, está em uma parte do futuro além do alcance atual. Contudo, é urgente colocar esse assunto na agenda. As ameaças colocadas pela tecnologia a outras áreas dos direitos humanos já são uma realidade entre nós. Meu

[†] [N.T.] Traduzido para o português, com a autorização do autor, por Carina de Castro Quirino e Renan Medeiros de Oliveira e revisado por Eduarda Oliveira Rodrigues, do artigo "Human Rights and Artificial Intelligence: An Urgently Needed Agenda", de Mathias Risse, originalmente produzido em língua inglesa. Carina de Castro Quirino é doutoranda em Direito Público na Universidade do Estado do Rio de Janeiro, Mestre em Direito pela Universidade Federal do Rio de Janeiro. Professora substituta de Direito Constitucional/Administrativo da Faculdade Nacional de Direito – FND/UFRJ e Pesquisadora do Laboratório de Regulação Econômica da Universidade do Estado do Rio de Janeiro - UERJ Reg. E-mail: carinacastrodir@gmail.com. Renan Medeiros de Oliveira é Mestrando em Direito Público e Bacharel em Direito pela Universidade do Estado do Rio de Janeiro (UERJ). Pós-graduando em Direito Público pela Pontifícia Universidade Católica de Minas Gerais (PUC Minas). Pesquisador no Centro de Justiça e Sociedade da Fundação Getulio Vargas (CJUS/FGV) e na Clínica de Direitos Fundamentais da Faculdade de Direito da UERJ - Clínica UERJ Direitos. Pesquisador Permanente do Laboratório de Regulação Econômica da UERJ - UERJ Reg. E-mail: renanmedeirosdeoliveira@gmail.com. Eduarda Oliveira Rodrigues é graduanda em economia da Universidade Federal Fluminense. E-mail: oliveiraeduarda@id.uff.br.

Revista Publicum
Rio de Janeiro, v.4, n.1, 2018, p. 17-33
http://www.e-publicacoes.uerj.br/index.php/publicum
DOI: https://doi.org/10.12957/publicum.2018.35098

objetivo aqui é analisar esses desafios de uma forma que distingue as perspectivas de curto, médio e longo prazo.
Palavras-chave
Direitos Humanos; Inteligência Artificial; Tecnologia

Human Rights and Artificial Intelligence: An Urgently Needed Agend

Abstract
Artificial intelligence generates challenges for human rights. Inviolability of human life is the central idea behind human rights, an underlying implicit assumption being the hierarchical superiority of humankind to other forms of life meriting less protection. These basic assumptions are questioned through the anticipated arrival of entities that are not alive in familiar ways but nonetheless are sentient and intellectually and perhaps eventually morally superior to humans. To be sure, this scenario may never come to pass and in any event lies in a part of the future beyond current grasp. But it is urgent to get this matter on the agenda. Threats posed by technology to other areas of human rights are already with us. My goal here is to survey these challenges in a way that distinguishes short-, medium term and long-term perspectives.
Keywords
Human Rights, Artificial Intelligence, Technology

Sumário

Introdução

A inteligência artificial gera desafios para os direitos humanos. A inviolabilidade da vida humana é a ideia central por trás dos direitos humanos, uma suposição implícita subjacente sendo a superioridade hierárquica da humanidade para outras formas de vida que merecem menos proteção. Essas suposições básicas são questionadas pela chegada antecipada de entidades que não estão vivas de maneiras conhecidas, mas que, no entanto, são sencientes e intelectual e, talvez, eventualmente, moralmente superiores aos humanos. Certamente, esse cenário pode nunca acontecer e, de qualquer forma, está em uma parte do futuro além do alcance atual. Mas é urgente colocar esse assunto na agenda. As ameaças colocadas pela tecnologia a outras áreas dos

direitos humanos já são uma realidade entre nós. Meu objetivo aqui é examinar esses desafios de uma forma que distingue as perspectivas de curto, médio e longo prazos[1].

1. Inteligência Artificial (IA) e Direitos Humanos

A IA está cada vez mais presente em nossas vidas, refletindo uma tendência crescente de pedir conselhos - ou de tomar decisões completas - para algoritmos. Por "inteligência", me refiro à capacidade de fazer previsões sobre o futuro e resolver tarefas complexas. Inteligência "artificial", IA, é essa capacidade demonstrada por máquinas, *smart phones*, tablets, laptops, drones, veículos autônomos ou robôs que podem assumir tarefas que variam de apoio domiciliar e companheirismo até mesmo a companheirismo sexual e policiamento e guerra.

Os algoritmos podem fazer qualquer coisa que possa ser codificada, desde que tenham acesso aos dados de que precisam, na velocidade exigida, e sejam colocados em um quadro de projeto que permita a execução das tarefas assim determinadas. Em todos esses domínios, o progresso tem sido enorme. A eficiência dos algoritmos é cada vez mais aprimorada através do *Big Data*: a disponibilidade de uma enorme quantidade de dados sobre toda a atividade humana e outros processos no mundo que permitem que um tipo particular de IA conhecido como "aprendizado de máquina" (*machine learning*) faça inferências sobre o que acontecerá com base na detecção de padrões. Algoritmos são melhores do que humanos onde quer que sejam testados, embora vieses humanos sejam perpetuados naqueles: qualquer sistema projetado por humanos reflete vieses humanos e algoritmos contam com dados que capturam o passado, automatizando, assim, o *status quo* se falharmos em evitá-los[2]. Mas os algoritmos são livres de ruído: ao contrário dos seres humanos, eles chegam à mesma decisão sobre o mesmo problema quando o apresentam duas vezes[3].

[1] Para discussões introdutórias sobre IA, veja Frankish and Ramsey, The Cambridge Handbook of Artificial Intelligence; Kaplan, Artificial Intelligence; Boden, AI. Para um panorama sobre filosofia da tecnologia além do que é discutido nesse estudo, cf. Kaplan, Readings in the Philosophy of Technology; Scharff and Dusek, Philosophy of Technology; Ihde, Philosophy of Technology; Verbeek, What Things Do. Confira-se, também, Jasanoff, The Ethics of Invention. Especialmente sobre filosofia e inteligência artificial, cf. Carter, Minds and Computers. Para uma discussão inicial de como a relação entre humanos e máquinas pode evoluir, veja Wiener, The Human Use Of Human Beings. Esse livro foi publicado originalmente em 1950.

[2] Confira-se a conferência de 2017 de Daniel Kahneman: https://www.youtube.com/watch?v=z1N96In7GUc On this subject, see also Julia Angwin, "Machine Bias." On fairness in machine learning, also see Binns, "Fairness in Machine Learning: Lessons from Political Philosophy"; Mittelstadt et al., "The Ethics of Algorithms"; Osoba and Welser, An Intelligence in Our Image.

[3] Sobre *Big Data*, cf. Mayer-Schönberger and Cukier, Big Data. On machine learning, see Domingos, The Master Algorithm. Sobre como algoritmos podem ser usados de forma injusta, gananciosa e de outras maneiras perversas, cf. O'Neil, Weapons of Math Destruction. Que os algoritmos podem fazer muito bem está claro também por trás de grande parte do potencial que a ciência social tem para melhorar a vida dos indivíduos e das sociedades, cf. Trout, The Empathy Gap.

Para os filósofos, o que chama a atenção é como, no contexto da IA, muitos debates filosóficos reafirmam que tais debates pareciam desconectados da realidade. Tomemos o problema do bonde, que provoca intuições sobre a moralidade deontológica versus consequencialista ao confrontar os indivíduos com escolhas que envolvem um bonde desgovernado que pode matar várias pessoas, dependendo do que esses indivíduos fizerem. Essas decisões não apenas determinam quem morre, mas também se algumas pessoas que não seriam afetadas são instrumentalizadas para salvar outras pessoas. Muitos professores universitários implantaram esses casos apenas para encontrar estudantes questionando sua relevância, já que, na vida real, as escolhas nunca seriam tão estilizadas. Mas uma vez que precisamos programar veículos autônomos (que acabaram de criar sua primeira fatalidade na estrada), há uma nova relevância e urgência pública para essas questões.

Além disso, há muito tempo os filósofos se intrigam com a natureza da mente. Uma questão é se há mais na mente do que o cérebro. Qualquer que seja a natureza, o cérebro também é um algoritmo complexo. Mas o cérebro é totalmente descrito dessa forma ou isso omite o que nos torna distintos, a saber, a consciência? Consciência é a experiência qualitativa de ser alguém ou algo, isto é, "como-é-ser-assim", como se poderia dizer. Se não há nada mais na mente do que o cérebro, então os algoritmos na era do *Big Data* nos superarão em breve em quase tudo o que fazemos: eles preveem cada vez mais precisamente os livros que gostamos ou onde gostaríamos de ir nas próximas férias; dirigem carros com mais segurança do que nós; fazem previsões sobre a nossa saúde antes que nossos cérebros soem alarmes; oferecem sólidos conselhos sobre quais trabalhos aceitar, onde morar, que tipo de animal de estimação adotar, se é sensato sermos pais ou ficar com a pessoa com quem estamos atualmente - tudo isso com base em uma miríade de dados relevantes sobre pessoas como nós. O anúncio na Internet que atende nossas preferências, avaliando o que pedimos ou clicamos anteriormente, é apenas uma sombra do que está por vir.

Se a mente é apenas um algoritmo complexo, podemos, eventualmente, ter pouca escolha a não ser conceder o mesmo status moral a certas máquinas que os humanos têm. Questões sobre o status moral dos animais surgem por conta das muitas continuidades entre os seres humanos e outras espécies: quanto menos os vemos como diferentes de nós em termos de propriedades moralmente relevantes, mais devemos tratá-los como companheiros de viagem em uma vida compartilhada, como feito, por exemplo, em Zoopolis, de Sue Donaldson e Will Kymlicka.[4]. Tal raciocínio acaba sendo transferido para as máquinas. Não devemos nos distrair com o fato de que, a partir de agora, as máquinas têm interruptores de desligamento. As futuras máquinas podem ser compostas e interligadas de maneiras que não permitem mais o desligamento fácil. Mais

[4] Donaldson and Kymlicka, Zoopolis.

importante, elas podem exibir emoções e comportamento para expressar apego: podem até se preocupar em serem desligadas e ficar ansiosas para fazer algo a respeito. Ou as máquinas futuras podem ser ciborgues, parcialmente compostas de partes orgânicas, enquanto humanos são modificados com partes não orgânicas para aprimoramento. Distinções entre humanos e não humanos podem se deteriorar. Ideias sobre a personalidade podem se alterar assim que se torna possível carregar e armazenar um cérebro digitalizado em um computador, da mesma forma que hoje em dia podemos armazenar embriões humanos.

Mesmo antes de isso acontecer, as novas gerações crescerão com máquinas de novas maneiras. Talvez não tenhamos nenhum escrúpulo em esmagar laptops quando eles não tiverem mais bom desempenho. Mas, se crescermos com uma babá-robô, cuja capacidade de aprendizado de máquina permite que ela nos atenda de maneiras muito além daquelas que os pais fazem, teríamos atitudes diferentes em relação aos robôs. Já em 2007, um coronel americano cancelou um exercício robótico de varredura de minas terrestres, porque considerou a operação desumana depois que um robô continuou rastejando ao perder uma perna de cada vez[5]. Ficções científicas como *Westworld* ou *The Good Place* antecipam como seria estar rodeado por máquinas que só podemos reconhecer como tal abrindo-as. Um robô humanoide chamado Sophia com capacidade de participar de entrevistas, desenvolvido pela *Hanson Robotics*, tornou-se cidadão saudita em outubro de 2017. Mais tarde, Sophia foi nomeada a primeira campeã de inovação do UNDP, a primeira não-humana com um título da ONU[6]. O futuro pode se lembrar desses momentos históricos. O mundo dos animais de estimação não está muito atrás. Jeff Bezos adotou recentemente um cão chamado SpotMini, um animal de estimação robótico versátil capaz de abrir portas, levantar-se e até mesmo carregar a máquina de lavar louça. E SpotMini nunca precisa sair para passear se Bezos preferir comprar na Amazon ou curtir os tweets presidenciais.

Se, de fato, existe mais na mente do que o cérebro, lidar com IA incluindo robôs humanoides seria mais fácil. A consciência pode, então, nos separar. É uma questão genuinamente aberta como compreender a experiência qualitativa e, portanto, a consciência. Mas mesmo que considerações sobre a consciência possam contradizer a visão de que os sistemas de inteligência artificial são agentes morais, eles não tornarão impossível para tais sistemas serem atores legais e, como tal, propriedade, cometerem crimes e serem responsáveis de maneira legalmente executável. Afinal, temos uma história de tratar corporações de tal maneira, e essas também não possuem consciência. Por mais que haja enormes dificuldades em separar a responsabilidade das

[5] Wallach and Allen, Moral Machines, 55.
[6] https://en.wikipedia.org/wiki/Sophia_(robot)

corporações da dos humanos envolvidos com elas, problemas semelhantes surgirão em relação às máquinas inteligentes.

2. A moralidade da inteligência pura

Um outro problema filosófico de longa data que obtém nova relevância aqui é a conexão entre racionalidade e moralidade. Essa questão surge quando nos perguntamos sobre a moralidade da inteligência pura. O termo "singularidade" se refere ao momento em que as máquinas ultrapassam os humanos na inteligência. Desde que os seres humanos conseguiram criar algo mais inteligente do que eles próprios, esse novo tipo de cérebro pode produzir algo mais esperto do que ele mesmo e assim sucessivamente, em grande velocidade. Haverá limites para quanto tempo isso pode continuar. Mas como os poderes computacionais aumentaram rapidamente ao longo das décadas, os limites para o que uma inteligência superior pode fazer estão além do que podemos imaginar agora. A singularidade e a superinteligência exercitam bastante alguns participantes no debate da IA, enquanto outros as consideram irrelevantes em comparação com preocupações mais prementes. De fato, pode nunca haver uma singularidade ou ela pode passar décadas ou centenas de anos de folga. Ainda assim, o avanço tecnológico exponencial das últimas décadas coloca esses tópicos em nossa agenda[7].

O que os filósofos pensam então é a disputa entre David Hume e Immanuel Kant sobre se a racionalidade fixa nossos valores. Notoriamente, Hume achava que a razão não fazia nada para fixar valores: um ser dotado de razão, racionalidade ou inteligência (suponhamos que todos são similarmente relevantes) poderia ter quaisquer objetivos, assim como qualquer variedade de atitudes, especialmente em relação aos seres humanos. Se assim for, uma superinteligência - ou qualquer IA, mas o problema é especialmente problemático para uma superinteligência - poderia ter praticamente qualquer tipo de compromisso de valor, incluindo aqueles que nos pareceriam um tanto absurdos (como maximizar o número de clipes de papel no universo, para mencionar um exemplo por vezes levantado na literatura). E como saberíamos que tais pensamentos são equivocados se, de fato, eles recebem tal superinteligência e seriam maciçamente mais inteligentes e, portanto, particularmente diferentes de nós?

Em oposição a isso, há a visão kantiana que deriva a moralidade da racionalidade. O Imperativo Categórico de Kant pede a todos os seres racionais que nunca usem suas próprias capacidades racionais nem as de qualquer outro ser racional de maneira puramente instrumental. Excluem-se, em particular, a violência gratuita e o engano de outros seres racionais (o que, para

[7] Chalmers, “The Singularity: A Philosophical Analysis”; Bostrom, Superintelligence; Eden et al., Singularity Hypotheses.

Kant, seria sempre muito parecido com a instrumentalização pura). Em um modo diferente de pensar sobre o Imperativo Categórico, é necessário que nossas ações sempre se deem de maneiras que passariam em um teste de generalização. Certas atitudes seriam tidas como inadmissíveis, porque não resistiriam se todos as tomassem, como, por exemplo, roubar e mentir: não haveria nenhuma propriedade se todos roubassem e nenhuma comunicação se todos se reservassem o direito de mentir. O ponto da derivação de Kant é que qualquer ser inteligente cairia em contradição consigo mesmo ao violar outros seres racionais. Grosso modo, isso é assim porque apenas nossa escolha racional que dá valor às coisas, o que também significa que, ao valorizarmos algo, estamos comprometidos em valorizar nossa capacidade de valor. Mas destruir outros seres racionais em busca de nossos próprios interesses atrapalha suas capacidades de valorizar, que são as mesmas capacidades cuja posse devemos valorizar em nós mesmos. Se Kant estiver certo, uma superinteligência pode ser um verdadeiro modelo para o comportamento ético. Como não podemos mudar a natureza humana - e a natureza humana intensamente paroquial em seus julgamentos e compromissos de valor -, a IA pode fechar a lacuna que se abre com humanos e sua Idade da Pedra, pequenos grupos orientados geneticamente que operam em um contexto global[8].

Se algo como esse argumento funcionasse - e há dúvidas - não teríamos nada com o que nos preocupar quanto a uma superinteligência. Possivelmente, seríamos racionais o suficiente para esse tipo de argumento gerar proteção para humildes humanos em uma era de máquinas muito mais inteligentes. Mas desde que uma série de filósofos que são inteligentes para padrões contemporâneos tem argumentado contra o ponto de vista kantiano, o assunto está longe de ser resolvido. Nós não sabemos como essas questões se pareceriam do ponto de vista de uma superinteligência.

É claro que algum tipo de moralidade poderia estar em vigor com a superinteligência responsável, mesmo que o valor não possa ser derivado apenas da racionalidade. Há também a abordagem hobbesiana de prever o que aconteceria com os seres humanos visando à autopreservação e caracterizada por certas propriedades em um estado de natureza sem uma autoridade compartilhada. Hobbes argumenta que, embora esses indivíduos não agissem com valores compartilhados apenas por pensar racionalmente, como fariam em um quadro kantiano, eles rapidamente experimentariam a maldade da vida sem uma autoridade compartilhada. Longe de serem vis, como indivíduos, eles se sentiriam compelidos a atacar uns aos outros em antecipação. Afinal, mesmo que eles se considerassem cooperativos e dessem ao outro lado o benefício da dúvida, eles não poderiam ter certeza de que o outro lado lhes daria o mesmo

[8] Petersen, "Superintelligence as Superethical"; Chalmers, "The Singularity: A Philosophical Analysis." Veja também a palestra de 2017 de Daniel Kahneman: https://www.youtube.com/watch?v=z1N96In7GUc

benefício e, assim, poderiam se sentir obrigados a atacar primeiro, dado o quanto está em jogo. A menos que haja apenas uma superinteligência, ou todas as superinteligências estejam intimamente ligadas de alguma maneira, talvez tal raciocínio se aplicasse também a essas máquinas e elas estariam sujeitas a algum tipo de autoridade compartilhada. Assim, o estado de natureza de Hobbes descreveria o status original das superinteligências *vis-à-vis* umas com as outras. Não está claro se tal autoridade compartilhada também criaria benefícios para os humanos.

Talvez as idéias de T. M. Scanlon sobre respostas apropriadas aos valores ajudariam. A superinteligência pode ser "moral" no sentido de reagir de maneira apropriada ao que observa ao seu redor. Talvez, então, tenhamos alguma chance de obter proteção ou mesmo algum nível de emancipação em uma sociedade mista composta de humanos e máquinas, dado que as habilidades do cérebro humano são realmente surpreendentes e geram capacidades em seres humanos que devem ser dignos de respeito[9]. Mas também são as capacidades dos animais, que, normalmente, não levam os humanos a reagir em relação a eles, ou ao meio ambiente, de maneira adequadamente respeitosa. Em vez de mostrar algo como um antropocentrismo esclarecido, muitas vezes instrumentalizamos a natureza. Com sorte, é possível que uma superinteligência simplesmente nos supere em tais questões e isso significará que a vida distintamente humana receberá alguma proteção, porque é digna de respeito. Não podemos saber com certeza, mas também não precisamos ser pessimistas.

3. Direitos Humanos e o Problema do Alinhamento de Valor

Todos esses assuntos estão em uma parte do futuro sobre a qual não sabemos quando ou mesmo se algum dia estará sobre nós. Mas do ponto de vista dos direitos humanos, esses cenários são importantes porque precisaríamos nos acostumar a compartilhar o mundo social que construímos ao longo de milhares de anos com novos tipos de seres. Outras criaturas até agora nunca ficaram no nosso caminho por muito tempo, e o melhor que elas podem esperar são alguns arranjos simbióticos como animais de estimação, gado ou zoológico. Tudo isso explicaria por que temos uma Declaração Universal de Direitos Humanos (DUDH) baseada em ideias sobre uma vida distintamente humana que parece merecer proteção, em nível individual, de um tipo que não estamos dispostos a conceder a outras espécies. Por motivos filosóficos, eu próprio penso que se justifica dar uma proteção especial aos seres humanos que assume a forma de direitos individuais, sem por isso dizer que praticamente tudo pode ser feito a outros animais ou ao meio ambiente. Contudo, tudo seria muito diferente com máquinas inteligentes. Nós controlamos os animais

[9] Para especulações sobre como tais sociedades mistas poderiam ser, cf. Tegmark, Life 3.0, chapter 5.

porque podemos criar um ambiente onde eles desempenham um papel subordinado. Mas podemos ser incapazes de fazê-lo com a IA. Nós precisaríamos, então, de regras para um mundo onde alguns agentes inteligentes são máquinas. Eles teriam que ser projetados de forma que respeitassem os direitos humanos, mesmo que fossem inteligentes e poderosos o suficiente para violá-los. Ao mesmo tempo, teriam que ser dotados de proteção adequada. Não é impossível que, eventualmente, a DUDH tenha que se aplicar a alguns deles[10].

Há uma urgência em garantir que esses desenvolvimentos tenham um bom começo. O desafio pertinente é o problema do alinhamento de valores, um desafio que surge antes que importe a moralidade da inteligência pura. Independentemente da precisão com que os sistemas de inteligência artificial são gerados, devemos tentar assegurar que seus valores estejam alinhados com os nossos, para tornar o mais improvável possível qualquer complicação no sentido de que uma superinteligência tenha compromissos com valores muito diferentes dos nossos. Que o problema do alinhamento de valores precisa ser enfrentado agora também está implícito nos Princípios Orientadores sobre Empresas e Direitos Humanos das Nações Unidas, criados para integrar os direitos humanos nas decisões de negócios. Esses princípios se aplicam à IA. Isso significa abordar questões como "Quais são os impactos potenciais mais graves?", "Quem são os grupos mais vulneráveis?" e "Como podemos garantir o acesso a mecanismos processuais?"[11].

Na comunidade de IA, o problema do alinhamento de valores foi reconhecido, no mais tardar, desde o conto de 1942 de Isaac Asimov “Runaround”, onde ele formula suas famosas Três Leis da Robótica, que são citadas como provenientes de um manual publicado em 2058 (sic!): (1) Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano se machuque. (2) Um robô deve obedecer às ordens dadas por seres humanos, exceto quando tais ordens entrem em conflito com a Primeira Lei. (3) Um robô deve proteger sua própria existência, desde que tal proteção não entre em conflito com a Primeira ou Segunda Leis.

No entanto, essas leis há muito são consideradas bastante inespecíficas. Vários esforços foram feitos para substituí-los, até agora sem qualquer conexão com os Princípios sobre Empresas e Direitos Humanos das Nações Unidas ou qualquer outra parte do movimento de direitos humanos. Entre outros esforços, em 2017, o Instituto *Future of Life* Cambridge, MA, fundado em torno do físico do MIT Max Tegmark e do cofundador do Skype, Jaan Tallinn, realizou uma

[10] Margaret Boden argumenta que as máquinas nunca podem ser agentes morais e, portanto, responsáveis; ela também pensa que é contra a dignidade humana oferecer companheiros de vida ou tipos de cuidadores que são máquinas. Veja https://www.youtube.com/watch?v=KVp33Dwe7qA (Para o impacto da tecnologia na interação humana, veja também Turkle, Alone Together.) Outros argumentam que certos tipos de IA teriam direitos morais ou mereceriam outros tipos de consideração moral; para as visões de Matthew Liao e Eric Schwitzgebel sobre isso, veja: https://www.youtube.com/watch?v=X-uFetzOrsg

[11] Ruggie, Just Business

conferência sobre IA Benéfica no centro de conferência Asilomar, na Califórnia, para apresentar princípios para guiar futuros desenvolvimentos da IA. Dos 23 Princípios Asilomar, 13 estão listados sob o título Ética e Valores. Entre outras questões, esses princípios insistem que, onde quer que a IA cause danos, deve ser verificável o porquê, e onde um sistema de IA estiver envolvido na tomada de decisões judiciais, seu raciocínio deve ser verificável pelos auditores humanos. Tais princípios respondem às preocupações de que a IA que utiliza *machine learning* possa raciocinar em tal velocidade e ter acesso a uma gama de dados tão grande que suas decisões serão cada vez mais obscuras, impossibilitando identificar se as análises se desviam. Os princípios também insistem no alinhamento de valores, conclamando que "sistemas de IA altamente autônomos devem ser projetados de modo que seus objetivos e comportamentos possam ser assegurados para se alinhar com os valores humanos em toda a operação" (Princípio 10). As ideias explicitamente aparecem no Princípio 11 (Valores Humanos) incluem "dignidade humana, direitos, liberdades e diversidade cultural"[12].

Insistir nos direitos humanos pressupõe que um certo conjunto de debates filosóficos tenha se resolvido: há valores universais, na forma de direitos, e nós, grosso modo, sabemos quais são esses direitos. Como os Princípios de Asilomar deixam claro, há aqueles na comunidade de IA que acreditam que os direitos humanos foram estabelecidos de maneira confiável. Mas outros estão ansiosos para evitar o que eles percebem como imperialismo ético. Eles acham que o problema do alinhamento de valores deve ser resolvido de forma diferente, ensinando, por exemplo, a inteligência artificial a absorver informações de todo o mundo, de uma maneira colaborativa. Portanto, este é outro caso em que um problema filosófico assume nova relevância: nossa compreensão filosoficamente preferida da metaética deve atuar para julgar se estamos confortáveis em colocar ou não os princípios dos direitos humanos no projeto da IA.[13]

Os direitos humanos também têm a vantagem de que tem havido numerosas formas de vernacularização dos direitos humanos ao redor do mundo. Suporte global para esses direitos é bastante substancial. E, novamente, já temos os Princípios Orientadores sobre Empresas e Direitos Humanos das Nações Unidas. Mas podemos ter certeza de que a China estará entre os principais produtores de IA e terá pouca inclinação para resolver o problema de alinhamento de valor em um espírito voltado para os direitos humanos. Isso não precisa frustrar os esforços de outros países para avançar com a solução dos direitos humanos para esse problema. Talvez, no devido tempo, os sistemas de IA possam trocar ideias sobre a melhor forma de se alinhar com os humanos. Mas ajudaria se os seres humanos planejassem a IA de uma maneira unificada, avançando com a mesma

[12] https://futureoflife.org/ai-principles/ Sobre o alinhamento de valores, veja também https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/

[13]

solução para o problema do alinhamento de valor. No entanto, uma vez que até os direitos humanos continuam a ter detratores, há pouca esperança de que isso aconteça.

O que de qualquer forma é necessário é mais interação entre os direitos humanos e as comunidades de IA, de modo que o futuro não seja criado sem a comunidade de direitos humanos. (Não há risco de que seja criado sem a comunidade de IA). Um passo importante nesse sentido é a decisão da Anistia Internacional de fazer uso extensivo de dispositivos de inteligência artificial em busca de causas de direitos humanos. Esta iniciativa foi inaugurada pelo Secretário-Geral então de saída, Salil Shetty, sendo o líder do projeto Sherif Elsayed-Ali. Nesse estágio, a Anistia Internacional está testando o uso da *machine learning* em investigações de direitos humanos e também se concentra no potencial de discriminação no uso de aprendizado de máquina, particularmente com relação ao policiamento, justiça criminal e acesso a serviços econômicos e sociais essenciais. A Anistia também está mais preocupada com o impacto da automação na sociedade, incluindo o direito ao trabalho e à subsistência. É preciso haver mais engajamento desse tipo, idealmente em ambos os sentidos, entre o movimento de direitos humanos e os engenheiros por trás desse desenvolvimento.

4. Estupidez Artificial e o Poder das Empresas

Há problemas mais imediatos do que as máquinas inteligentes do futuro, embora elas precisem avançar de forma correta. O exercício de cada direito humano na DUDH é afetado por tecnologias, de uma forma ou de outra. As disposições antidiscriminação são ameaçadas se os algoritmos, utilizados em áreas que vão desde cuidados de saúde a subscrição de seguros para decisões de liberdade condicional, são racistas ou sexistas, porque o aprendizado que eles fazem tenha bases sexistas ou racistas. A liberdade de discurso e expressão, e qualquer outra liberdade que os indivíduos tenham, é minado pela enxurrada de notícias falsas que nos cerca, incluindo a fabricação de vídeos falsos que poderiam ter qualquer pessoa fazendo qualquer coisa, abarcando atos de terrorismo que nunca ocorreram ou foram cometidos por pessoas diferentes.

Quanto mais a participação política depende da internet e das mídias sociais, mais elas são ameaçadas pelos avanços tecnológicos, desde a possibilidade de implantar robôs da Internet cada vez mais sofisticados participando de debates online até hackear dispositivos usados para contar votos ou hackear administrações públicas ou acessórios para criar desordem. Onde quer que haja IA, também há EA, estupidez artificial. Como poderia ser muito pior do que um BS a que nos acostumamos demais: esforços feitos por adversários não apenas para minar os ganhos obtidos pela IA, mas para transformá-los em seu oposto. A manipulação russa nas eleições é um alerta; é provável que algo muito pior surja. Os direitos judiciais podem ser ameaçados se a IA for usada

sem transparência suficiente e possibilidade de escrutínio humano. Um sistema de IA previu os resultados de centenas de casos no Tribunal Europeu dos Direitos Humanos, prevendo veredictos com precisão de 79%; e, uma vez que a precisão fique ainda mais alta, será tentador usar a IA também para tomar decisões. O uso da IA em processos judiciais pode ajudar a gerar acesso a aconselhamento legal para os pobres (um dos projetos que a Anistia persegue, especialmente na Índia); mas também pode levar a situações kafkianas se os algoritmos derem conselhos incompreensíveis[14].

Quaisquer direitos à segurança e à privacidade são potencialmente prejudicados não apenas por drones ou soldados robôs, mas também pela crescente legibilidade e rastreabilidade de indivíduos em um mundo de atividades e presenças humanas registradas eletronicamente. A quantidade de dados disponíveis sobre as pessoas provavelmente aumentará enormemente, especialmente quando os sensores biométricos puderem monitorar a saúde humana. (Eles podem nos checar no chuveiro e enviar seus dados, e isso pode ser de nosso interesse porque a doença se torna diagnosticável antes de se tornar um problema.) Haverá desafios para os direitos civis e políticos decorrentes da pura existência desses dados e do fato de que eles podem ser de propriedade privada, mas não daqueles a quem os dados se referem. As empresas líderes no setor de IA são mais poderosas do que as empresas de petróleo, e esse é, presumivelmente, apenas o começo de sua ascensão.

No passado, o *status* em sociedades complexas era determinado primeiro pela propriedade da terra e, depois da Revolução Industrial, pela posse de fábricas. As estruturas altamente desiguais resultantes não funcionaram bem para muitos. A propriedade desigual dos dados terá também consequências prejudiciais para muitas pessoas na sociedade. Se o poder de empresas como Alphabet, Apple, Facebook ou Tesla não for aproveitado para o bem público, poderemos, eventualmente, nos encontrar em um mundo dominado por empresas, como descrito, por exemplo, no romance de Oryx e Crake, de Margaret Atwood, ou Infinite Jest, de David Foster Wallace. O escândalo Cambridge-Analytica é um alerta aqui, e o depoimento de Mark Zuckerberg aos senadores norte-americanos em 10 de abril de 2018 revelou um grau surpreendente de ignorância entre os legisladores seniores sobre o funcionamento das empresas de internet cujo modelo de negócios depende de dados de marketing. Tal ignorância abre caminho para o poder das empresas. Ou considere um ponto relacionado: os governos precisam do setor privado para ajudar na segurança cibernética. Os especialistas relevantes são inteligentes, caros e muitos nunca trabalhariam para o governo. Só podemos esperar que seja possível cooptá-los, uma vez que o

[14] http://www.bbc.com/news/technology-37727387

governo está sobrecarregado aqui. Se tais esforços falharem, apenas as empresas fornecerão o mais alto nível de segurança cibernética.

5. A grande desconexão: tecnologia e desigualdade

Isso me leva ao meu último tópico: IA e desigualdade, e a conexão entre esse tópico e os direitos humanos. Para começar, devemos dar atenção à advertência de Thomas Piketty de que o capitalismo deixado por conta própria em tempos de paz gera uma desigualdade econômica cada vez maior. Aqueles que possuem a economia se beneficiam mais do que aqueles que apenas trabalham lá. Com o tempo, as chances de vida dependerão cada vez mais do *status* social no nascimento[15]. Também vemos mais e mais como aqueles que produzem tecnologia ou sabem como usar a tecnologia para ampliar o impacto podem comandar salários cada vez mais altos. A IA apenas reforçará essas tendências, tornando mais fácil para os líderes de todos os segmentos ampliarem seu impacto. Isso, por sua vez, faz com que os produtores de IA sejam fornecedores cada vez mais caros de tecnologia. Mais recentemente, aprendemos com Walter Scheidel que, historicamente, reduções substanciais na desigualdade só ocorreram em resposta a calamidades como epidemias, colapsos sociais, desastres naturais ou guerra. Caso contrário, é difícil reunir vontade política efetiva de mudança[16].

Os luditas originais esmagaram os teares na Inglaterra do século XIX, porque se preocupavam com empregos. Mas até agora toda onda de inovação tecnológica acabou criando mais empregos do que os que destruiu. Embora a mudança tecnológica não tenha sido boa para todos, foi boa para a sociedade como um todo e para a humanidade. É possível que haja tantos empregos que aqueles que desenvolvem, supervisionam ou usem tecnologia de forma inovadora, assim como profissões criativas que não podem ser deslocadas, acabarão superando os que perdem empregos para a IA. Mas o apego a essa esperança seria ingênuo, porque pressupõe uma revisão radical do sistema educacional para tornar as pessoas competitivas. Alternativamente, poderíamos esperar por alguma combinação de criação de empregos, menor jornada de trabalho para que os empregos possam ser compartilhados, mas também salários mais altos para que as pessoas possam ter uma vida decente. Porém, de qualquer forma, pode-se ter mais esperança nos países europeus do que nos Estados Unidos, onde tantos ficaram para trás na corrida entre tecnologia e educação e onde a solidariedade em nível nacional está tão arraigada que até o atendimento de saúde universal permanece contestado[17]. Como os países em desenvolvimento

[15] Piketty, Capital in the Twenty-First Century
[16] Scheidel, Great Leveler.
[17] Goldin and Katz, The Race Between Education and Technology

com vantagem comparativa em manufatura e mão-de-obra barata vão se sair bem em tudo isso é algo que ninguém sabe.

Diante desse pano de fundo, devemos nos preocupar se a IA forçará uma expansão tecnológica em sociedades que deixam milhões excluídos, os tornam redundantes como participantes do mercado e, portanto, podem minar o ponto de sua participação na comunidade política. Quando a riqueza era determinada pela propriedade da terra, os ricos precisavam de descanso, porque o ponto de propriedade da terra era cobrar o aluguel. Quando a riqueza era determinada pela propriedade das fábricas, os proprietários precisavam do resto para trabalhar nas máquinas e comprar coisas. Mas aqueles que estão do lado perdedor da divisão tecnológica podem não ser mais necessários. Em seu conto de 1926, "The Rich Boy", F. Scott Fitzgerald escreveu: "Deixe-me falar sobre os muito ricos. Eles são diferentes de você e de mim". A IA pode validar essa afirmação de maneira notável.

Eventualmente, poderemos ver novos bantustões, como na África do Sul do *apartheid*, ou, talvez mais provavelmente, o surgimento de entidades privadas separadas de propriedade da empresa com serviços sociais maravilhosos dos quais outros são excluídos. Talvez apenas o suficiente seja dado a esses outros para que eles não se rebelem completamente. O tecido da sociedade pode se dissolver se houver muito mais pessoas do que o necessário como participantes em qualquer sentido. Embora o mundo fosse rico o suficiente para lhes oferecer uma vida decente, a vontade política de fazê-lo pode não estar entre os privilegiados se existirem formas de ir em frente que permitam a vida privilegiada sem medo de desordens violentas. Tudo isso seria seriamente uma má notícia do ponto de vista dos direitos humanos. Cenários como este estão mais para o futuro do que as preocupações mais imediatas da crescente presença de algoritmos na vida humana, mas, provavelmente, não tão longe no futuro quanto a chegada de uma superinteligência. As chances são de que os desafios provenientes do aumento da desigualdade chegam nos próximos 70 anos da DUDH.

Os EUA são o centro da tecnologia global, incluindo a IA, mas, na verdade, tem muito menos prática do que, digamos, muitas nações europeias em solidariedade nacional para ajudar nos esforços sustentados para tornar a IA benéfica para toda a população. Os EUA têm uma mobilidade social terrivelmente baixa. Estudos apontam que até 50% de todos os empregos agora são suscetíveis à automação, incluindo profissões tradicionalmente seguras, como direito, contabilidade e medicina[18]. Ou, como Philip Alston, Relator Especial da ONU sobre Pobreza Extrema e Direitos Humanos, observou em uma visita oficial em 2017 aos EUA:

[18] https://rightsinfo.org/rise-artificial-intelligence-threat-human-rights/

> A automação e a robotização já estão tirando muitos trabalhadores de meia idade dos empregos em que eles acreditavam estar seguros. Na economia do século XXI, apenas uma pequena porcentagem da população está imune à possibilidade de cair na pobreza como resultado de quebras ruins além de seu próprio controle[19].

Muitas vezes ouvimos que devemos progredir com a mudança tecnológica apenas se ela puder ser amplamente compartilhada[20]. Mas, como acabamos de observar, medidas radicais contra a desigualdade só acontecem em tempos profundamente conturbados, tempos em que não desejaríamos viver. O aumento da desigualdade nas últimas décadas, bem como a eleição de um homem que personifica a ganância, vingança e total falta de empatia normal não é um bom presságio para qualquer esforço de disseminação da riqueza nos EUA, independentemente do quanto isso seja bom em conferências e eventos políticos.

Devemos nos preocupar com esses aumentos de desigualdade também por seu impacto nos direitos humanos. É difícil superestimar o que está em jogo. Marx estava certo quando, em Sobre a Questão Judaica, assinalou que a emancipação concebida integralmente em termos de direitos não era atraente. Uma sociedade construída em torno de ideais baseados em direitos perde muito. Nos últimos 70 anos, o movimento pelos direitos humanos muitas vezes fracassou em enfatizar o tema maior do qual os direitos humanos devem fazer parte: a justiça distributiva, interna e global. A IA pode, eventualmente, colocar em risco o próprio legado do Iluminismo, porque a individualidade, como tal, está cada vez mais cercada em uma era de *Big Data* e *machine learning*. Isso também pode acontecer, uma vez que o que está ameaçado aqui também é o tipo de preocupação com a sociedade como um todo, capturada no pensamento moderno sobre justiça distributiva ou social, que só se tornou possível com o surgimento do Iluminismo e das possibilidades tecnológicas abertas pela industrialização. Gostaria de poder terminar com uma nota mais edificante, e, realmente, não acho que seja "tarde demais". Mas é provável que o aumento da desigualdade em combinação com a IA seja a ruína dos próximos 70 anos na vida da DUDH. A menos que, talvez, pessoas suficientes vejam esses tópicos como incluídos na urgência feroz de agora.

Referências

ANGWIN, Julia; LARSON, Jeff. “Machine Bias.” Text/html. **ProPublica**, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[19] http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=22533&LangID=E Sobre a divisão tecnológica, cf. também ttps://www.politico.com/agenda/story/2018/02/07/technologyinterview-mit-david-autor-000629 And see also http://harvardpolitics.com/world/automation/
Sobre IA e o future do trabalho, veja também Brynjolfsson and McAfee, The Second Machine Age; Kaplan, Humans Need Not Apply.

[20] Por exemplo, nesse evento: http://futureofwork.mit.edu/

BINNS, Reuben. “Fairness in Machine Learning: Lessons from Political Philosophy.” **Proceedings of Machine Learning Research**, v. 81, p. 1-11, 2018.

BODEN, Margaret A. **AI**: Its Nature and Future. 1 edition. Oxford, United Kingdom: Oxford University Press, 2016.

BOSTROM, Nick. **Superintelligence**: Paths, Dangers, Strategies. Reprint edition. Oxford, United Kingdom; New York, NY: Oxford University Press, 2016.

BRYNJOLFSSON, Erik; MCAFEE, Andrew. **The Second Machine Age**: Work, Progress, and Prosperity in a Time of Brilliant Technologies. 1. ed. New York London: W. W. Norton & Company, 2016.

CARTER, Matt. **Minds and Computers**: An Introduction to the Philosophy of Artificial Intelligence. 1. ed. Edinburgh: Edinburgh University Press, 2007.

Chalmers, David J. “The Singularity: A Philosophical Analysis.” **Journal of Consciousness Studies**, v. 17, n. 9-10, p. 7-65, 2010.

DOMINGOS, Pedro. **The Master Algorithm**: How the Quest for the Ultimate Learning Machine Will Remake Our World. Reprint edition. Basic Books, 2018.

DONALDSON, Sue; KYMLICKA, Will. **Zoopolis**: A Political Theory of Animal Rights. 1 edition. Oxford; New York: Oxford University Press, 2013.

EDEN, Amnon H. et al (Eds.). **Singularity Hypotheses**: A Scientific and Philosophical Assessment. 2012 edition. New York: Springer, 2013.

FRANKISH, Keith; RAMSEY, William M. Ramsey (Eds.). **The Cambridge Handbook of Artificial Intelligence**. Cambridge, UK: Cambridge University Press, 2014.

GOLDIN, Claudia; KATZ, Lawrence. **The Race Between Education and Technology**. Cambridge, Mass.: Belknap, 2008.

IHDE, Don. **Philosophy of Technology**: An Introduction. 1. ed. New York: Paragon House, 1998.

JASANOFF, Sheila. ***The Ethics of Invention***: Technology and the Human Future. New York: W. W. Norton & Company, 2016.

KAPLAN, David M. (Ed). **Readings in the Philosophy of Technology**. 2. ed. Lanham: Rowman & Littlefield Publishers, 2009.

KAPLAN, Jerry. **Artificial Intelligence**: What Everyone Needs to Know. 1. ed. New York, NY, United States of America: Oxford University Press, 2016.

_____. **Humans Need Not Apply**: A Guide to Wealth and Work in the Age of Artificial Intelligence. Reprint edition. New Haven: Yale University Press, 2016.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big Data**: A Revolution That Will Transform How We Live, Work, and Think. Reprint edition. Boston: Eamon Dolan/Mariner Books, 2014.

MITTELSTADT, Brent Daniel et al. “The Ethics of Algorithms: Mapping the Debate.” **Big Data & Society**, v. 3, n. 2, dez. 2016: 205395171667967. https://doi.org/10.1177/2053951716679679.

O’NEIL, Cathy. **Weapons of Math Destruction**: How Big Data Increases Inequality and Threatens Democracy. Reprint edition. New York: Broadway Books, 2017.

OSOBA, Osonde A.; WELSER, William. **An Intelligence in Our Image**: The Risks of Bias and Errors in Artificial Intelligence. Santa Monica, Calif: RAND Corporation, 2017.

PETERSEN, Steve. “Superintelligence as Superethical.” In: LIN, Patrick; ABNEY, Keith; JENKINS, Ryan (Eds.). **Robot Ethics 2.0**: From Autonomous Cars to Artificial Intelligence. 1. ed. New York, NY: Oxford University Press, 2017, p. 322-337.

PIKETTY, Thomas. **Capital in the Twenty-First Century**. Cambridge: Belknap, 2014.

RUGGIE, John Gerard. **Just Business**: Multinational Corporations and Human Rights. 1 edition. New York: W. W. Norton & Company, 2013.

SCANLON, T. M. “What Is Morality?” In: SHEPHARD, Jennifer M.; KOSSLYN, Stephen Michael; HAMMONDS, Evelynn Maxine (Eds.). **The Harvard Sampler**: Liberal Education for the Twenty-First Century. Cambridge (MA), 2011.

SCHARFF, Robert C.; DUSEK, Val (Eds.). **Philosophy of Technology**: The Technological Condition: An Anthology. 2 ed. Malden, MA: Wiley-Blackwell, 2014.

SCHEIDEL, Walter. **Great Leveler**: Violence and the History of Inequality from the Stone Age to the Twenty-First Century. Princeton, NJ: Princeton Univers. Press, 2017.

TEGMARK, Max. **Life 3.0**: Being Human in the Age of Artificial Intelligence. New York: Knopf, 2017.

TROUT, J. D. **The Empathy Gap**: Building Bridges to the Good Life and the Good Society. New York, N.Y: Viking Adult, 2009.

TURKLE, Sherry. **Alone Together**: Why We Expect More from Technology and Less from Each Other. Expanded, Revised edition. Basic Books, 2017.

VERBEEK, Peter-Paul. **What Things Do**: Philosophical Reflections on Technology, Agency, and Design. Illustrated edition edition. University Park, Pa: Penn State University Press, 2005.

WALLACH, Wendell; ALLEN, Colin. **Moral Machines**: Teaching Robots Right from Wrong. 1 edition. Oxford: Oxford University Press, 2010.

WIENER, Norbert. **The Human Use Of Human Beings**: Cybernetics And Society. Revised edition. New York, N.Y: Da Capo Press, 1988.

Enviado em: 25 de maio de 2018

Aprovado em: 18 de junho de 2018