POIESIS
REVISTA DE FILOSOFIA
ISSN:2448-3095

ENTRE A DESTRUIÇÃO E A REDENÇÃO: A I. A. E A FALSA SIMETRIA ENTRE TECNOFÓBICOS E TECNÓFILOS

BETWEEN DESTRUCTION AND REDEMPTION:
A I. A. AND THE FALSE SYMMETRY BETWEEN TECHNOPHOBICS AND TECHNOPHILES

Fernando Meireles Monegalha Henriques[1]

Resumo: Neste artigo, traço algumas considerações sobre o avanço da Inteligência Artificial, baseando-me para tanto num artigo recente de Eliezer Yudkowsky que causou bastante controvérsia. Lanço mão a partir disso, para justificar o ponto de vista desse autor, da ideia de uma *heurística do temor*, tal como foi aventada principalmente por Hans Jonas. A conclusão a que intento chegar é que o desenvolvimento de novas tecnologias deveria ser refreado enquanto não tivermos dados suficientes para afastar os maus prognósticos que ainda versam sobre elas e seus efeitos disruptivos – por mais utópico, sabemos, que seja querer impor freios ao desenvolvimento de novas tecnologias.

Palavras-chave: Yudkowsky – Hans Jonas – Inteligência Artificial – Heurística do Temor

Abstract: In this article, I outline some considerations about the advancement of Artificial Intelligence, based on a recent article by Eliezer Yudkowsky that caused a lot of controversy. From this, I draw on the idea of a *heuristic of fear*, as proposed mainly by Hans Jonas, to justify Yudkowsky's point of view. The conclusion I intend to reach is that the development of new technologies should be curbed until we have enough data to rule out the bad prognoses that still concern them and their disruptive effects – however utopian, we know, it may be to want to impose brakes on the development of new technologies.

Keywords: Yudkowsky – Hans Jonas – Artificial Intelligence – Heuristic of Fear

1. INTRODUÇÃO

Gostaria de iniciar este artigo propondo que a divisão habitualmente imposta entre tecnofóbicos e tecnófilos é de muita pouca valia e que ela poderia ser reenquadrada de outra

[1] Doutor e mestre em Filosofia pela Universidade Federal de São Carlos (UFSCar), Professor do curso de graduação em Filosofia da Universidade Federal de Alagoas (UFAL) e membro do PPGFIL/IFAL. E-mail: fernando.meireles@ichca.ufal.br

forma, a saber, fazendo-se uma distinção entre uma perspectiva *crítica* e outra *dogmática* da inteligência artificial e, mais do que isso, da técnica em geral. Mas, mesmo que se adote a divisão usual entre tecnofóbicos e tecnófilos, eu sustentarei que a simetria que normalmente se estabelece entre eles é falsa, na medida em que um dos lados parece ter uma clara e nítida vantagem argumentativa sobre o outro, como eu tentarei explicar, tomando como base a obra *O princípio responsabilidade*, do filósofo alemão (de origem judia) Hans Jonas, o qual, como veremos, está bem longe de ser um tecnofóbico. As reflexões de Hans Jonas me parecem completamente essenciais para se pensar o nosso tempo histórico e, mais do que isso, o advento da Inteligência Artificial. O que eu direi sobre estes pontos não é evidentemente nenhuma novidade, ao menos para os especialistas nas áreas. Dito isso, passemos à primeira parte deste artigo, em que trago alguns pontos importantes para se entender os receios e temores atualmente trazidos pela Inteligência Artificial.

1. POR QUE YUDKOWKSI TEM RAZÃO, E POR QUE SEUS AVISOS NÃO DERAM EM (QUASE) NADA

Para iniciar, deixe-me explicar o estado da questão. No meio de junho de 2023, mais de trinta mil pessoas já tinham assinado uma carta aberta proposta pelo Instituto para o Futuro da Vida (Future of Life Institute), intitulada "Pausar experimentos gigantes de Inteligência Artificial: uma carta aberta". Os signatários iniciais desta carta aberta, lançada em 22 de março de 2023, incluíam alguns pesos-pesados da indústria de tecnologia e de bens em geral, tais como Elon Musk, fundador da TESLA e da SPACE-X ou Steve Wozniak, co-fundador da Apple, além do escritor de best-sellers Yuval Harari. Em comum, todos eles pediam uma pausa de seis meses no desenvolvimento de novos projetos de Inteligência Artificial com poder computacional maior do que o ChatGPT 4. Alguns dos pontos levantados pela carta são os seguintes:

> Os sistemas de IA com inteligência competindo com a [inteligência] humana podem colocar riscos profundos para a sociedade e a humanidade [...] Como afirmado nos amplamente endossados Princípios de IA Asilomar, "a IA Avançada pode representar uma mudança profunda na história da vida na Terra e deve ser planejada e gerenciada com cuidados e recursos proporcionais". Infelizmente, esse nível de planejamento e gerenciamento não está acontecendo, uma vez que os últimos meses tenham visto os laboratórios de IA condenados (*locked*) a uma corrida fora de controle para desenvolver e implantar mentes digitais cada vez mais poderosas que ninguém – nem mesmo seus criadores – pode entender, prever ou controlar de forma confiável. (tradução nossa) (FUTURE OF LIFE INSTITUTE, 2023)

Após esse diagnóstico, a carta aberta lança uma série de perguntas:

> "devemos deixar as máquinas inundarem nossos canais de informação com propaganda e inverdade? Devemos automatizar todos os trabalhos, incluindo os que nos satisfazem? Devemos desenvolver mentes não humanas que possam eventualmente superar em número, ser mais inteligentes, nos tornar obsoletos e nos substituir? Devemos correr o risco de perder o controle de nossa civilização? Tais decisões não devem ser delegadas a líderes da tecnologia não eleitos [os presidentes das big techs]. Sistemas poderosos de IA devem ser desenvolvidos apenas quando estivermos confiantes de que seus efeitos serão positivos e seus riscos gerenciáveis. (tradução nossa) (FUTURE OF LIFE INSTITUTE, 2023)

Cabe salientar, contudo, que um dos grandes inspiradores da carta, o especialista em Inteligência Artificial e teórico da decisão Eliezer Yudkowsky, co-fundador do Instituto de Pesquisa sobre Inteligência de Máquina (Machine Intelligence Research Institute), uma instituição privada sem fins lucrativos baseada em Berkeley – Califórnia, não assinou a carta. Suas alegações para tanto vieram sob a forma de um impactante artigo na revista Time, publicado em inglês e intitulado, numa tradução livre, *O único modo de lidar com a ameaça da IA: desligue tudo*. Eu apresento as teses gerais deste artigo a partir de agora.

Para Yudkowsky, o ponto-chave não é, como coloca a carta em seu primeiro parágrafo, "os sistemas de IA com inteligência competindo com a [inteligência] humana". Para ele, o ponto-chave, e o grande problema, surge quando as IAs se tornam *mais inteligentes do que a inteligência humana*. Para Yudkowsky, isso pode ocorrer a qualquer momento num laboratório de pesquisa, sem que os pesquisadores nem mesmo se deem conta – trata-se de um limiar muito pouco claro, mas que aparentemente não está muito longe de ser alcançado. Cabe salientar que toda argumentação de Yudkowsky que vem a seguir não depende do fato de haver uma IA no sentido forte do termo, isto é, de construirmos uma máquina com capacidades sencientes ou conscientes. Não, basta apenas que haja uma IA no sentido fraco do termo, isto é, uma IA que apenas simule ou emule a inteligência humana, ou algo além dela, para que já tenhamos a possibilidade de enormes problemas, como veremos a seguir.

Que problemas seriam estes? Yudkowsky os resume numa passagem algo bombástica:

> Muitos pesquisadores mergulhados nessas questões, inclusive eu, esperam que o mais provável resultado da construção de uma inteligência artificial super-humana, sob qualquer coisa remotamente parecida com as atuais circunstâncias, é que literalmente todos na Terra morrerão. Não como em "talvez possivelmente haja

> alguma chance remota", mas como em "essa é a coisa óbvia que aconteceria". (YUDKOWSKY, 2023)

Antes de prosseguir, abramos aqui um parêntese: Yudkowsky apressa-se em dizer que não é que não possamos, em tese, conviver com uma inteligência muito mais avançada do que a nossa. O problema é que isso exige uma preparação, uma precisão e uma cautela muito maiores do que as que observamos hoje em dia. Exige, entre outras coisas, uma IA efetivamente *alinhada* com os interesses e valores humanos. Com efeito, não podemos nos esquecer que o autor em questão é especialista também na questão do alinhamento das IAs. O alinhamento ocorre quando uma IA atende aos objetivos e interesses de seus criadores, uma IA é dita desalinhada quando isso não ocorre. Um exemplo clássico de uma IA desalinhada é o de uma máquina criada para construir palitos de dentes. Vamos supor que, para cumprir sua tarefa, ela destrua todas as árvores na Terra. Com efeito, esse foi um efeito não previsível dessa simples tarefa básica. O mesmo ocorre com a questão de uma inteligência mais avançada do que a humana: como garantir que ela esteja efetivamente alinhada com valores básicos como, digamos, a declaração universal dos direitos humanos ou, se se quiser, com uma certa simpatia com seres sencientes em geral? Yudkowsky levanta de certo modo este ponto em seu artigo quando indaga:

> Sem essa precisão e preparação, o resultado mais provável é uma IA que não faz o que queremos e não se importa conosco nem com a vida senciente em geral. Esse tipo de cuidado é algo que poderia, em princípio, ser imbuído em uma IA, mas não estamos prontos e não se sabe atualmente como. (YUDKOWSKY, 2023)

Mas fechemos aqui o nosso parêntese. Depois desta primeira afirmação, Yudkowsky continua em seu artigo com algumas outras afirmações bombásticas, que precisam melhor o que ele tinha em mente com a primeira afirmação:

> Para visualizar uma IA sobre-humana hostil, não imagine um pensador inteligente e sem vida morando dentro da internet e enviando e-mails mal-intencionados. Visualize uma civilização alienígena inteira, pensando milhões de vezes mais rápido que a velocidade humana, inicialmente confinado a computadores - em um mundo de criaturas que são, do seu ponto de vista, muito estúpidas e muito lentas. Uma IA suficientemente inteligente não ficará confinada aos computadores por muito tempo. No mundo de hoje, você pode enviar sequências de DNA por e-mail a laboratórios que produzirão proteínas sob demanda, permitindo uma IA inicialmente confinada à internet construir formas de vida artificiais. (YUDKOWSKY, 2023)

Parece cenário de ficção científica, e em certa medida o é, na medida em que a nossa realidade se tornou já faz algum tempo uma ficção científica ruim. A consequência básica desse cenário proposto por Yudkowsky é o seguinte: "Se alguém construir uma IA muito poderosa, nas condições atuais, espero que cada membro da espécie humana e toda a vida biológica na Terra morra logo em seguida." (YUDKOWSKY, 2023)

Descrito esse cenário de horror absoluto (extinção não apenas da vida humana, mas de toda a vida na Terra), compreendemos melhor porque o autor em questão (Yudkowsky) não se prestou a assinar a carta mencionada mais acima: nós não temos nenhuma possibilidade, segundo ele, no atual estado de nosso desenvolvimento científico e tecnológico, de garantir que isso não ocorra. Uma moratória de seis meses de forma alguma garantiria que esse problema básico do alinhamento da máquina aos valores humanos seja garantido. Como afirma o autor: "Não vamos preencher essa lacuna em seis meses." (YUDKOWSKY, 2023) Daí a insuficiência da carta, segundo o autor. Seria necessário, para tanto, um prazo muito mais estendido, e até mesmo indefinido, para se alcançar o nível de segurança necessário para a existência de uma inteligência super-humana:

> Demorou mais de 60 anos entre quando a noção de Inteligência Artificial foi pela primeira vez proposta e estudada, para chegarmos às capacidades de hoje. Resolver a *segurança* da inteligência sobre-humana – não segurança perfeita, segurança no sentido de "não matar literalmente todos" – poderia razoavelmente levar pelo menos metade desse tempo. E a coisa sobre tentar isso com inteligência sobre-humana é que se você errar na primeira tentativa, você não pode aprender com seus erros, porque você estará morto. (YUDKOWSKY, 2023)

Yudkowsky salienta que esse é um temor de muitos pesquisadores da área de IA, que não chegam a expor publicamente seus pontos de vista. Contudo, sabe-se de uma pesquisa, publicada em 2022, que fez a seguinte pergunta a especialistas em IA: "'Qual é, na sua estimativa, a probabilidade de os seres humanos não conseguirem impedir que sistemas de IA avançados do futuro provoquem a extinção humana ou uma expropriação permanente e grave da espécie humana, semelhante à sua extinção?' A resposta mediana foi de 10%" (KLEIN, 2023b)

O que fazer, diante de um quadro potencialmente desolador como esse? A proposta de Yudkowsky a esse respeito é: decretar uma moratória indefinida e global. Uma moratória, por exemplo, que encerre a competição desenfreada entre países como Estados Unidos e China neste domínio, com base no convincente argumento de que ela potencialmente pode gerar a morte de todos os estadunidenses e chineses, assim de todos seres humanos na face da Terra.

Tal moratória inclui, entre outras coisas, desligar todos os grandes *clusters* de GPUs (fazendas de computadores onde são treinadas as mais poderosas IAs), colocar um teto em quanto poder computacional alguém está autorizado a usar no treinamento de um sistema de IA, rastrear todas as GPUs vendidas, fazer acordos multinacionais para impedir que as atividades proibidas se movam para outros lugares e países e, caso isso ocorra, bombardear e destruir os *datacenters* por meio de ataques aéreos. Para Yudkowsky, essas medidas são necessárias na medida em que os riscos trazidos pela IA chegam a ser *maiores* do que aqueles trazidos pelas armas nucleares (de fato, podemos dizer que esses dois riscos andam juntos!).

O autor então finaliza: "Nós não estamos prontos. Não estamos no caminho certo para estarmos significativamente prontos num futuro próximo. Se formos adiante, todos morrerão, inclusive as crianças que não escolheram isso e não fizeram nada de errado. Desliguem tudo." (YUDKOWSKY, 2023)

*

* *

O artigo de Yudkowsky (que apresentamos somente por cima aqui), como era de se esperar foi logo classificado com epítetos pouco elogiosos, tal como "ludita" ou, como é de praxe, "tecnofóbico" ou "tecnófobo". Ele estaria, segundo alguns de seus detratores mais apressados nas redes sociais, simplesmente pintando um cenário clássico de ficção científica para expor então sua profunda desconfiança em relação à tecnologia, mais especificamente ao alinhamento de uma Inteligência Artificial Geral, acreditando assim demasiadamente em suas próprias crenças e argumentos, e estando pouco aberto a criticismos. Acerca dessa divisão entre "tecnofóbicos" e "tecnófilos", cabem aqui algumas ponderações, nós cremos. Abramos, portanto, outro parêntese. Um procedimento comum em filosofia é servir-se de oposições simples para expor visões divergentes sobre um determinado assunto, esse é o caso com a divisão entre "tecnofóbicos" e "tecnófilos". O risco a que se chega aí, na prática, é o de uma discussão infinita, uma negatividade pura e simples, sem nenhuma possibilidade de resolução dialética ou, ainda, de alguma medida efetiva a ser tomada: temos a discussão pela discussão. Esse, note-se, é o modo mais simples de tornar a filosofia inoperante ou, o que é pior, colocá-la a serviço dos poderes instituídos: servir-se de oposições simples como "tecnofóbicos" ou "tecnófilos", engajá-la num pugilato intelectual interminável, sem que se derive nenhuma

resolução, tampouco nenhuma atitude prática (daí deriva, pontuemos, o horror de Deleuze à discussão filosófica interminável). Além disso, os próprios termos do debate entre "os amigos da técnica" (os tecnófilos) e os "que temem a técnica" (os tecnofóbicos) nos parecem viciados, já que essas posturas intelectuais extremadas raramente serão encontradas em estado puro e, além disso, já colocam os "tecnofóbicos" numa posição de rebaixamento e de defesa (com efeito, ouve-se muitas vezes se dizer "fulano é tecnofóbico" como uma espécie de quase-xingamento). Com efeito, poderíamos dizer que, assim como usamos a divisão entre "tecnófilos"/"tecnofóbicos", também seria igualmente válida um outro tipo de divisão: a saber, a que parte da constatação de que temos de um lado pessoas que têm algum tipo de reserva em relação ao desenvolvimento tecnológico desenfreado, e que gostariam de impor-lhe limites – uma postura *crítica*, portanto – e pessoas que, pelos mais variados motivos, acham utópico querer impor algum limite ao desenvolvimento tecnológico, aceitando resignadamente e de forma *dogmática* o seu crescimento ilimitado. Se pudéssemos, usaríamos então a divisão entre *críticos* e *dogmáticos* proposta por Kant para expor a divisão em tela. Parece-nos mais produtivo do que a divisão entre tecnófilos e tecnofóbicos.

Fechemos, contudo, este parêntese. A despeito de seus detratores apressados, os alertas de Yudkowsky parecem ter tido algum impacto, ao menos na medida em que se mesclaram positivamente a algumas outras ações que foram amplamente divulgadas pela mídia. Seguem algumas delas:

- um comunicado do Centro pela segurança da IA (*Centre for AI Safety*), assinado por pessoas como o CEO da OpenAI, Sam Altman e o CEO da Google DeepMind, Demis Hassabis, colocando, em uma linha, que "mitigar o risco de extinção da IA deve ser uma prioridade global ao lado de outros riscos em escala social, como pandemias e guerra nuclear"

- a saída do vice-presidente de engenharia do Google, Geoffrey Hinton, considerado por muitos como uma lenda em seus campos de estudos (psicologia da cognição e ciências da computação), a fim de alertar a sociedade dos riscos da IA.

- uma certa busca por regulamentação e desaceleração verificada entre os desenvolvedores e dirigentes do Vale do Silício ligados à IA (cf. KLEIN, 2023c). EUA, UE e China também já elaboraram propostas e legislações específicas para regulamentar a IA (a despeito das insuficiências de cada uma delas). E por aí vai....

Mas convenhamos: tudo isso é muito pouco, se comparado ao que precisaria ser feito, segundo preconiza Yudkowsky. Com efeito, há bons motivos para ser completamente cético

em relação à possibilidade de qualquer moratória no desenvolvimento da IA, se levarmos em consideração a própria dinâmica competitiva do capitalismo e das relações globais entre Estados e governos. Como pontua a esse respeito o jornalista Ezra Klein:

> As grandes empresas de tecnologia estão numa corrida pela hegemonia da IA. Os EUA e a China estão numa corrida pela hegemonia da IA. Está jorrando dinheiro para empresas com *expertise* em IA. Sugerir que avancemos mais devagar ou mesmo que paremos por completo já parece infantilidade. Se uma empresa for mais devagar, outra vai acelerar seus esforços. Se um país apertar o botão de pausa, os outros avançarão com mais determinação. O fatalismo conduz à inevitabilidade, e a inevitabilidade vira a justificativa da aceleração. (KLEIN, 2023d)

Por enquanto, o principal fator de risco associado à IA não é o horizonte da destruição total, mas seus impactos no mercado de trabalho. Houve, de fato, uma crescente conscientização dos riscos evidentes de desemprego que a IA coloca para os trabalhadores, principalmente para aquelas profissões ligadas ao trabalho intelectual, assim como para aquelas ligadas ao trabalho mais burocrático. É curioso notar que uma futurologia algo desconexa dizia, não muito tempo atrás, que o trabalho intelectual e criativo seria aquele que se preservaria com o advento da nova revolução tecnológica. Agora se vê que é justamente o contrário: as profissões que tendem a ser preservadas, ao menos mais imediatamente, são exatamente aquelas relacionadas ao trabalho manual: artesãos, trabalhadores da construção civil, entre outros. Com a classe média sendo afetado pela ameaça de desemprego crônico, abre-se o espaço para um período de turbulência, crise social e revoluções. Mas fechemos agora esse primeiro tópico, de apresentação da problemática, e passemos à uma discussão mais filosófica, partindo de uma breve apresentação das propostas éticas de Hans Jonas. Aos poucos ficará bastante claro sua intersecção com tudo que vimos até aqui.

2. HANS JONAS E OS PERIGOS DA TÉCNICA

Vimos até o presente momento porque os diagnósticos e prognósticos de Yudkowsky dificilmente surtirão os efeitos práticos desejados pelo autor. Mas eu me comprometi a mostrar algo além disso, a saber, que Yudkowski está certo. Para tanto, não o defenderei do ponto de vista técnico (não tenho evidentemente competência para tanto), mas com algumas considerações baseadas em parte em Hans Jonas e em seu *opus magnus O Princípio Responsabilidade*. Para isso, começarei fazendo uma breve apresentação da vida de Jonas. Hans Jonas (nascido em 1903 e falecido em 1993) foi um filósofo alemão de origem judia,

que estudou filosofia e teologia nas universidades de Freiburg, Berlim e Heidelberg. Obteve seu doutoramento na universidade de Marburg, em 1928, com uma tese sobre o antigo gnosticismo orientada por Martin Heidegger. Em Marburg conheceu também Hannah Arendt, com quem manteria um vínculo de amizade pessoal pelo resto de suas vidas. Com a afiliação e Heidegger ao partido nazista, em 1933, Jonas rompe seus contatos com ele, chegando a repudiar publicamente Heidegger em 1964. Devido à ascensão do partido nazista, ele deixou a Alemanha em 1933, movendo-se para a Inglaterra e depois para a Palestina. Em 1940 ele voltou para a Inglaterra, a fim de integrar uma brigada especial do exército inglês, composta de judeus que quisessem lutar contra Hitler. Após ao fim da guerra, ao saber que sua mãe tinha sido enviada para uma câmara de gás, ele se recusa a voltar para a Alemanha. Ele volta para a Palestina, onde toma parte nos conflitos da guerra árabe-isralense (1948), seguindo posteriormente para o Canadá (1950) e para Nova Iorque/EUA (1955), onde ele se estabeleceria até o fim dos seus dias, atuando de 1955 a 1976 como professor da *New School for Social Research*. Do ponto de vista filosófico, a obra de Jonas é vasta e muito importante, principalmente no que tange aos campos da ética, da bioética e da filosofia ecológica, onde ela teve um impacto profundo, mas também no que tange aos seus estudos sobre o gnosticismo antigo. Seu livro mais difundido e lido foi sem dúvida *O princípio responsabilidade – ensaio de uma ética para a civilização tecnológica*, publicado em alemão em 1979 e em inglês em 1984. Mas temos também outros livros fundamentais de Jonas que foram também traduzidos para o português, tais como *Técnica, medicina e ética – sobre a prática do princípio responsabilidade* (de 1985) e *O princípio vida – fundamentos para uma biologia filosófica* (1966).

No restante deste artigo, concentrar-me-ei em alguns poucos elementos de *O princípio responsabilidade*, um livro atualíssimo para as discussões éticas envolvendo a IA. Como Jonas expõe logo no prefácio do livro: “A tese de partida deste livro é que a promessa da tecnologia moderna se converteu em ameaça” (JONAS, 2006, p. 21). Evidentemente, quando Jonas escrevia em 1979, ainda não havia, como há hoje, esse risco iminente causado pela possibilidade de destruição em massa pela IA. Mas já havia outros riscos proporcionados pela técnica, como aqueles relacionados ao uso de armamentos nucleares, além de uma incipiente preocupação ecológica, que iria se desenvolver ao longo dos anos com a crise climática

global. Como Jonas coloca no prefácio da obra, esses fatores colocam problemas éticos inteiramente novos, que a ética tradicional não está preparada para resolver. Jonas denomina então *O princípio responsabilidade* de um "tractatus technologico-eticus" (JONAS, 2006, p. 23), uma obra que tentará pensar os desdobramentos éticos das novas tecnologias, pensando a necessidade de se recuperar o tema da responsabilidade – o tema central do livro – que devemos ter para com as futuras gerações. Para tanto, Jonas propõe um interessante *tournant* metafísico da ética, refundando-a numa doutrina que põe a investigação ontológica sobre o Ser em seu centro. Não temos, contudo, como desenvolver este tema aqui, que é, contudo de fundamental importância. Para o que vem, concentrar-me-ei na crítica de Jonas ao imperativo categórico kantiano, que pode ser um ponto de partida interessante

Como vocês sabem, Kant formula o seu famoso imperativo categórico de três formas, sendo a mais famosa formulação aquela que diz: "Aja como se a máxima de tua ação devesse tornar-se, através da tua vontade, uma lei universal.". O que é genial no imperativo categórico, é que ele não nos dá nenhuma ação concreta que devamos ou não executar – como fazem as éticas tradicionais – mas nos dá um critério para poder distinguir se uma conduta é ética ou não, se ela é correta ou não. Tomando o clássico exemplo da mentira, podemos constatar que a máxima subjetiva "tu deves mentir" não passa pelo teste da universalização, uma vez que uma sociedade em que todos mentem está condenada à dissolução. Pelo teste do imperativo categórico, compreendemos que "Tu deves mentir" então não é uma máxima eticamente aceitável, uma vez que ela não implica uma conduta universalizável. Logo, a máxima "tu deves mentir" não é correta. O gênio de Kant consistiu em buscar encontrar um critério para a correta conduta humana sem apelar para nenhum princípio transcendente (uma divindade, por exemplo), mas unicamente se baseando na capacidade racional que cada ser humano traz consigo. Para Kant, a ética está vinculada à razão humana.

Contudo, por mais interessante que seja o imperativo categórico kantiano, Jonas encontrará algumas brechas nele, que o levarão a propor um outro tipo de imperativo. Basicamente, a crítica que Jonas endereça a Kant se baseia no caráter instantaneista e individualista do imperativo kantiano: segundo Jonas, não é contraditório pensar numa conduta universalizável, mas que traga danos ou até mesmo sacrifique as gerações futuras. No capítulo intitulado "Velhos e novos imperativos" do *Princípio responsabilidade*, Jonas não chega a dar nenhum exemplo concreto, mas poderíamos pensar, por exemplo, no crescimento econômico: "Deve haver crescimento econômico" parece ser uma máxima bastante razoável,

se aplicada no contexto atual. Mas isso pode levar a problemas ecológicos gravíssimos, que afligirão principalmente as futuras gerações. Diferentemente de Heidegger, para quem o ser-para-a-morte é um existencial fundamental, a preocupação de Jonas incide principalmente sobre a natalidade, sobre a vida que desponta a cada geração humana (e não-humana). E é para com as gerações futuras, segundo Jonas, que a responsabilidade humana deve se voltar. Daí decorre sua proposta de reformulação do imperativo categórico, que se tornaria basicamente o seguinte: "Aja de modo a que os efeitos da tua ação sejam compatíveis com a permanência de uma autêntica vida humana sobre a Terra" (JONAS, 2006, p. 47) ou, expresso negativamente, "Aja de modo a que os efeitos de tua ação não sejam destrutivos para a possibilidade futura de uma tal vida" (JONAS, 2006, p. 47-48). Como se pode ver, o que Jonas busca é "desdobrar" o imperativo categórico no *tempo*, retirando-o de seu instantaneísmo, a fim de fazê-lo abarcar também a dimensão temporal do *futuro*. Trata-se também de uma tentativa deliberada de fazer com que o imperativo abranja não somente a esfera individual (como parece ser o caso em Kant), mas alcance igualmente a esfera das políticas públicas.

Evidentemente, alguém poderia argumentar com Groucho Marx aqui: mas o que que as futuras gerações já fizeram por nós? Elas nem existem ainda! Jonas reconhece, nesta etapa do seu livro, que é necessário um trabalho de fundamentação dessa exigência teórica da responsabilidade para com as futuras gerações, o que exigiria um recuo à metafísica que não temos a possibilidade de fazer aqui. A essa altura do campeonato, ele simplesmente nos pede que aceitemos esse imperativo "sem justificativa, como um axioma" (JONAS, 2006, p. 48). Axioma bastante convincente, contudo! Pois parece haver uma inclinação natural do ser humano para essa preocupação com as gerações que estão despontando, que deveria naturalmente se estender às gerações mais longínquas. Quem já é mãe ou pai, sabe disso na pele, mas mesmo quem já cuidou de um irmão ou irmã ou parente menor sabe do que estou falando. Um sinal bastante eloquente disso no artigo de Yudkowsky que mencionei mais acima é que ele lance mão do argumento do perigo que corre sua filha, e todos os nossos filhos, com o desenvolvimento desenfreado da Inteligência Artificial. O pano de fundo do artigo de Yudkowsky me parece muito próximo da argumentação de Hans Jonas. Essa é a preocupação de Jonas, de Yudkowsky, mas é também a minha preocupação. Que mundo estamos deixando para nossos filhos (se é que estamos deixando um mundo para eles)?

Mas voltemos a Jonas. De posse desse novo imperativo, ele não tarda a retirar suas consequências. Como se sabe, Jonas critica fortemente todo ideal utópico que vê o desenvolvimento tecnológico como necessariamente fonte de progresso e aperfeiçoamento. Não, o desenvolvimento tecnológico pode ser tanto fonte de excelentes realizações quanto de catástrofes inimagináveis. Não é que Jonas condene o progresso tecnológico em bloco. Como bem lembra Jelson Oliveira (2022), Jonas não é um "tecnofóbico". Essa foi uma pecha que se atribuiu erroneamente a ele no Brasil por causa do famoso artigo de Gérard Lebrun – *Sobre a tecnofobia*. Para Jonas, não se trata de condenar a técnica, mas de criarmos condições para um desenvolvimento tecnológico responsável, em que a aceleração tecnológica não seja um mandamento por si mesmo. Para tanto, Jonas define alguns deveres da ética do futuro, que decorrem do imperativo acima proposto. Primeiramente, é necessário que haja uma projeção dos efeitos a longo prazo dos impactos das novas tecnologias: precisamos projetar no futuro as consequências benéficas e maléficas de uma nova técnica disruptiva, de um novo saber, e analisar as possibilidades positivas e negativas que essa técnica e esse saber trazem consigo. Esse é o primeiro dever da ética do futuro preconizada por Jonas. Além disso, é preciso que haja algum "sentimento adequado" a essa ética do futuro, um sentimento que mobilize a nossa responsabilidade para com a vida na Terra, com a humanidade e com as futuras gerações. Esse sentimento é para Jonas o *temor*. O temor não é o medo (embora tenha sido assim traduzido na edição brasileira): ele não é um estado patológico que deve nos levar a uma certa paralisia da ação – como é o caso do medo –, mas uma disposição ativa do espírito, que faz com que nós assumamos a responsabilidade de nossas ações, precavendo-nos de uma esperança infundada que poderia nos fazer acabar cometendo um erro potencialmente fatal. O segundo dever da ética do futuro é, portanto, estabelecer essa *heurística do temor*, que nos permita antever consequências ruins de uma determinada técnica ou de determinado saber. Por fim, se a ética tem efetivamente algum papel a desempenhar, é necessário tomar uma decisão – tal técnica será ou não implementada, caso já tenha sido implementada, seu desenvolvimento será ou não permitido? Para tanto, precisamos nos ater não somente ao seu desenvolvimento atual, mas projetar suas consequências futuras. Precisamos operar não somente no campo do real, mas do *possível*. E aqui, o mau prognóstico deve ter claramente uma primazia sobre o bom prognóstico. Como afirma Jonas: "*grosso modo... é necessário dar mais ouvidos à profecia da desgraça do que à profecia da salvação*" (JONAS, 2006, p. 48). O motivo disso fica evidente se pensarmos no artigo de Yudkowsky que mencionamos acima:

operando tão somente no campo das possibilidades, podemos dizer que o desenvolvimento da Inteligência Artificial precisaria ser pausado, pois há uma possibilidade efetiva dela causar a destruição da humanidade. Note-se bem: é um perigo tão somente possível mas, se aceitamos os termos propostos por Jonas, isso já deveria ser motivo para se pausar o seu desenvolvimento. Mas isso seria o caso, evidentemente, num mundo em que a ética tivesse algum lugar e alguma importância mínima, um mundo que permitisse o agir racional e zelasse pela proteção das gerações futuras.

Tudo isso me parece puro bom senso. Mas o mundo atual é regido cada vez menos pelo bom senso. É por isso, de um modo geral, que Yudkowski e os críticos da IA estão certos. A diferença básica é: se um crítico da IA estiver certo e o desenvolvimento da IA não for pausado, você e seus filhos morrem, e as futuras gerações jamais virão a lume. Se uma pessoa com uma crença dogmática na tecnologia estiver certa, e ainda assim ocorresse essa moratória por tempo indeterminado, tudo que ocorreria seria uma pausa no desenvolvimento tecnológico e na dinâmica competitiva do capital. Mas isso parece cada vez mais impossível. Pois, de fato, é mais fácil imaginar o fim do mundo do que colocar um freio no capital e no desenvolvimento da técnica.

REFERÊNCIAS:

FUTURE OF LIFE INSTITUTE. **Pause Giant AI Experiments:** An Open Letter. Disponível em <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>. Publicado em 22/03/2023. Acessado em 06/12/2024.

JONAS, H. **O Princípio Responsabilidade**: ensaio de uma ética para uma civilização tecnológica. Rio de Janeiro: PUC Rio, 2006.

_____. **O Princípio Vida:** fundamentos para uma biologia filosófica. Petrópolis: Vozes, 2002.

KLEIN, E. Humanidade precisa se adaptar à inteligência artificial ou decidir desacelerá-la. **Folha de São Paulo**. São Paulo, 14 de março de 2023. Disponível em: <

https://www1.folha.uol.com.br/colunas/ezra-klein/2023/03/humanidade-precisa-se-adaptar-a-inteligencia-artificial-ou-decidir-desacelera-la.shtml> Acessado em 06/12/2024.

LEBRUN, G. **Sobre a tecnofobia**. In: LEBRUN, G. A filosofia e sua história. Organização de Carlos Alberto Ribeiro de Moura, Maria Lúcia M. O. Cacciola e Marta Kawano. São Paulo: Cosac Naify, 2006. p. 481-508.

OLIVEIRA, J. Para uma *ethical turn* da tecnologia: por que Hans Jonas não é um tecnofóbico. **Trans/form/ação**. Marília, v.45 (02), abr/jun 2022.

YUDKOWSKY, E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. **TIME magazine**. Disponível em: <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>. Publicado em 29/03/2023. Acessado em 06/12/2024.