IERFH
Instituto Ética, Racionalidade
e Futuro da Humanidade

A Ética da Inteligência Artificial [1]

por Nick Bostrom & Eliezer Yudkowsky, 2011*

A possibilidade de criar máquinas pensantes levanta uma série de questões éticas. Estas questões se entrelaçam tanto para garantir que as máquinas não prejudiquem os humanos e outros seres moralmente relevantes, como para o *status* moral das próprias máquinas. A primeira seção discute questões que podem surgir no futuro próximo da Inteligência Artificial (IA). A segunda seção destaca os desafios para assegurar que a IA opere com segurança uma vez que se aproxima dos seres humanos e de sua inteligência. A terceira seção destaca a forma como podemos avaliar se, e em que circunstâncias, sistemas de IA possuem *status* moral. Na quarta seção nós consideramos como sistemas de IA podem diferir dos humanos em alguns aspectos básicos relevantes para nossa avaliação ética deles. A seção final se destina a questões da criação de IAs mais inteligente do que a inteligência humana, e assegurar que elas usem essa inteligência avançada para o bem ao invés de a utilizarem para o mal.

Ética em Máquinas Aprendizes e outros domínios específicos de algoritmos de IA

Imagine, num futuro próximo, um banco usando uma máquina de algoritmo de aprendizagem[2] para aprovar solicitações de pedidos de hipo-

[1] Nota do Tradutor: *"The Ethics of Artificial Intelligence". Draft for Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011): forthcoming.* A tradução do texto para o nosso idioma foi feita de forma livre, ou seja, utilizando termos que facilitassem o entendimento em nosso idioma, mas, com o cuidado de manter o sentido do texto original. Ainda assim divergências podem ocorrer, e por isso, assumo antecipadamente a responsabilidade por equívocos na tradução.

[2] N. T.: Os Algoritmos de aprendizagem se relacionam com a aprendizagem de sistemas artifi-

tecas. Um candidato rejeitado move uma ação contra o banco, alegando que o algoritmo está discriminando racialmente os solicitantes de hipoteca. O banco responde que isso é impossível, pois o algoritmo é deliberadamente cego para a raça do solicitante. Na realidade, isso faz parte da lógica do banco para implementação do sistema. Mesmo assim, as estatísticas mostram que a taxa de aprovação do banco para candidatos negros tem constantemente caído. Submetendo dez candidatos aparentemente iguais e genuinamente qualificados (conforme determinado por um painel independente de juízes humanos), revela-se que o algoritmo aceita candidatos brancos e rejeita candidatos negros. O que poderia estar ocorrendo?

Encontrar uma resposta pode não ser fácil. Se o algoritmo de aprendizagem da máquina é baseado em uma complexa rede neural ou em um algoritmo genético produzido por evolução dirigida, pode se revelar quase impossível entender por que, ou mesmo como, o algoritmo está julgando os candidatos com base em sua raça. Por outro lado uma máquina aprendiz baseada em árvores de decisão ou redes Bayesianas é muito mais transparente para inspeção do programador (Hastie *et al.* 2001), o que pode permitir a um auditor descobrir se o algoritmo de IA usa informações de endereço dos candidatos para saber previamente onde nasceram ou se residem em áreas predominantemente pobres.

Algoritmos de IA desempenham um papel cada vez maior na sociedade moderna, embora geralmente não estejam rotulados como "IA". O cenário descrito acima pode estar acontecendo da mesma forma como nós descrevemos. E se tornará cada vez mais importante desenvolver algoritmos de IA que não sejam apenas poderosos e escaláveis[3], mas também

---

ciais, e é uma subdivisão da área da IA dedicada ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender e aperfeiçoar seu desempenho em alguma tarefa. Algumas partes da aprendizagem de máquina estão intimamente ligadas à mineração de dados e focadas nas propriedades dos métodos estatísticos. A aplicação prática inclui o processamento de linguagem natural, sistemas de busca, diagnósticos médicos, bioinformática, reconhecimento de fala, reconhecimento de escrita, visão computacional e locomoção de robôs.

[3] N. T.: Um sistema escalável é um sistema que tem seu desempenho aumentado com o acréscimo de *hardware*. A escalabilidade é de vital importância em sistemas eletrônicos, bancos de dados, redes de computadores e roteadores. Na área das telecomunicações e da engenharia de *software*, a escalabilidade é uma característica desejável em todo o sistema, em uma rede ou em um processo, pois é um indicador confiável para verificarmos se um sistema está preparado para crescer e para nos certificarmos de que o sistema é hábil em manipular uma porção crescente de trabalho de forma uniforme. Por exemplo, isto pode se referir à capacidade de um sistema em suportar um aumento de carga total quando os recursos (normalmente do *hardware*) são requeridos. Um significado análogo está relacionado ao uso dessa

transparentes para inspeção – para citar umas das muitas propriedades socialmente importantes.

Alguns desafios de máquinas éticas são muito semelhantes a outros desafios envolvidos em projetar máquinas. Projetar um braço robótico para evitar o esmagamento de seres humanos distraídos não é moralmente mais preocupante do que projetar um retardador de chamas para sofá. Trata-se de novos desafios de programação, mas não de novos desafios éticos. Mas, quando algoritmos de IA se ocupam de trabalho cognitivo com dimensões sociais – tarefas cognitivas anteriormente realizadas por humanos –, o algoritmo de IA herda as exigências sociais. Seria sem dúvida frustrante descobrir que nenhum banco no mundo deseja aprovar a sua aparentemente excelente solicitação de empréstimo, sem que ninguém saiba por que, e ninguém pode ainda descobrir mesmo em princípio. (Talvez você tenha um primeiro nome fortemente associado com fraqueza? Quem sabe?).

Transparência não é a única característica desejável da IA. Também é importante que algoritmos de IA que assumam funções sociais sejam previsíveis aos que o governam. Para compreender a importância dessa previsibilidade, considere uma analogia. O princípio legal de *stare decisis*[4] impele juízes a seguir os antecedentes sempre que possível. Para um engenheiro, esta preferência pelo precedente pode parecer incompreensível – por que amarrar o futuro com o passado, quando a tecnologia está sempre melhorando? Mas uma das mais importantes funções do sistema legal é ser previsível, de modo que, por exemplo, os contratos possam ser escritos sabendo como eles serão executados. O trabalho do sistema jurídico não é necessariamente o de aperfeiçoar a sociedade, mas proporcionar um ambiente previsível no qual cidadãos possam aperfeiçoar suas próprias vidas.

Também se tornará cada vez mais importante que os algoritmos de IA se tornem *resistentes à manipulação*. Um sistema visual de máquinas que faz a varredura de bagagem em aeroportos deve ser resistente contra

---

palavra em termos comerciais onde a escalabilidade implica um modelo de negócio que oferece potencial de crescimento econômico dentro da empresa.
(Disponível On-line: **http://pt.wikipedia.org/wiki/Escalabilidade**)

[4] N. T.: *Stare decisis* é uma expressão em latim que pode ser traduzida como "ficar com as coisas decididas". Essa expressão é utilizada no direito para se referir à doutrina segundo a qual as decisões de um órgão judicial criam precedente, ou seja, jurisprudência, e vinculam as decisões que serão emitidas no futuro.

(Disponível On-line: **http://pt.wikipedia.org/wiki/Stare_decisis**).

adversários humanos deliberadamente à procura de fraquezas exploráveis no algoritmo – por exemplo, uma forma que, colocada próxima a uma pistola em uma das bagagens, neutralizaria o reconhecimento da mesma. Resistência contra manipulação é um critério comum em segurança da informação; quase *o* critério. Mas não é um critério que aparece frequentemente em revistas especializadas em aprendizagem de máquinas – que estão atualmente mais interessadas em, por exemplo, como um algoritmo aumenta proporcionalmente em grandes sistemas paralelos.

Outro importante critério social para transações em organizações é ser capaz de encontrar a pessoa responsável por conseguir que algo seja feito. Quando um sistema de IA falha em suas tarefas designadas, quem leva a culpa? Os programadores? Os usuários finais? Burocratas modernos muitas vezes se refugiam nos procedimentos estabelecidos que distribuem responsabilidade amplamente, de modo que uma pessoa não pode ser identificada nem culpada pelo resultado das catástrofes (Howard 1994). O provável julgamento comprovadamente desinteressado de um sistema especialista poderia se transformar num refúgio ainda melhor. Mesmo que um sistema de IA seja projetado com uma substituição do usuário, é uma obrigação considerar o incentivo na carreira de um burocrata que será pessoalmente responsabilizado se a substituição sair errada, e que preferiria muito mais culpar a IA por qualquer decisão difícil com um resultado negativo.

Responsabilidade, transparência, auditabilidade, incorruptibilidade, previsibilidade e uma tendência para não fazer vítimas inocentes gritarem em desamparada frustração: todos os critérios que aplicamos a humanos que desempenham funções sociais; todos os critérios que devem ser considerados em um algoritmo destinado a substituir o julgamento humano de funções sociais; todos os critérios que podem não aparecer em um registro de aprendizado de máquina, considerando o quanto um algoritmo aumenta proporcionalmente para mais computadores. Esta lista de critérios não é de forma alguma exaustiva, mas serve como uma pequena amostra do que uma sociedade cada vez mais informatizada deveria estar pensando.

Inteligência Artificial Geral

Há concordância quase universal entre os profissionais modernos de IA que sistemas de Inteligência Artificial estão aquém das capacidades

humanas em algum sentido crítico, embora algoritmos de IA tenham batido os seres humanos em muitos domínios específicos como, por exemplo, o xadrez. Tem sido sugerido por alguns que logo que os pesquisadores de IA descobrem como fazer alguma coisa, esta capacidade deixa de ser considerada como inteligente – o xadrez era considerado o epítome da inteligência até o *Deep Blue* vencer Kasparov no campeonato mundial – mas mesmo esses pesquisadores concordam que algo importante está faltando às IA's modernas (ver Hofstadter 2006).

Enquanto esta subárea da Inteligência Artificial está apenas crescendo de forma unificada, "Inteligência Artificial Geral" (IAG) é o termo emergente usado para designar IA "real" (ver, por exemplo, o volume editado por Goertzel e Pennachin 2006). Como o nome implica, o consenso emergente é que a característica que falta é a generalidade. Os algoritmos atuais de IA com desempenho equivalente ou superior ao humano são caracterizados por uma competência deliberadamente programada em um único e restrito domínio. O *Deep Blue* tornou-se o campeão do mundo em xadrez, mas ele não pode jogar damas, muito menos dirigir um carro ou fazer uma descoberta científica. Tais algoritmos modernos de IA assemelham-se a todas as formas de vidas biológicas com a única exceção do *Homo sapiens*. Uma abelha exibe competência em construir colméias; um castor exibe competência em construir diques; mas uma abelha não pode construir diques, e um castor não pode aprender a fazer uma colméia. Um humano, observando, pode aprende a fazer ambos; mas esta é uma habilidade única entre as formas de vida biológicas. É discutível se a inteligência humana é verdadeiramente geral – nós somos certamente melhores em algumas tarefas cognitivas do que em outras (Hirschfeld e Gelman 1994) – mas a inteligência humana é, sem dúvida, significativamente mais geralmente aplicável que a inteligência não-hominídea.

É relativamente fácil imaginar o tipo de questões de segurança que podem resultar de IA operando somente dentro de um domínio específico. É uma classe qualitativamente diferente de problema manipular uma IAG operando através de muitos novos contextos que não podem ser previstos com antecedência.

Quando os engenheiros humanos constroem um reator nuclear, eles prevêem eventos específicos que poderiam acontecer em seu interior – falhas nas válvulas, falha nos computadores, aumento de temperatura no núcleo – para evitar que esses eventos se tornem catastróficos. Ou, em um nível mais mundano, a construção de uma torradeira envolve previsão do pão e previsão da reação do pão para os elementos de aquecimento da tor-

radeira. A torradeira em si não sabe que o seu objetivo é fazer torradas – o propósito da torradeira é representado na mente do designer, mas não é explicitamente representado em computações dentro da torradeira – e se você colocar um pano dentro de uma torradeira, ela pode pegar fogo, pois o projeto é realizado em um contexto não previsto, com um imprevisível efeito colateral.

Mesmo algoritmos de IA de tarefas específicas nos lançam fora do paradigma da torradeira, o domínio do comportamento especificamente previsível, localmente pré-programado. Considere o *Deep Blue*, o algoritmo de xadrez que venceu Garry Kasparov no campeonato mundial de xadrez. Na hipótese de as máquinas poderem apenas fazer exatamente o que eles dizem, os programadores teriam de pré-programar manualmente um banco de dados contendo movimentos possíveis para cada posição de xadrez que o *Deep Blue* poderia encontrar. Mas isso não era uma opção para os programadores do *Deep Blue*. Em primeiro lugar, o espaço de possíveis posições do xadrez é abundantemente não gerenciável. Segundo, se os programadores tinham de inserir manualmente o que consideravam um bom movimento em cada situação possível, o sistema resultante não teria sido capaz de fazer movimentos mais fortes de xadrez do que o de seus criadores. Uma vez que os próprios programadores não são campeões do mundo, esse sistema não teria sido capaz de derrotar Garry Kasparov.

Ao criar um super jogador de xadrez, os programadores humanos necessariamente sacrificaram sua capacidade de previsão *local* para o *Deep Blue*, *específico* comportamento do jogo. Em vez disso, os programadores do *Deep Blue* tinham (justificável) confiança que os movimentos de xadrez do *Deep Blue* satisfariam um critério não-local de otimização: isto é, que os movimentos tenderiam a orientar o futuro resultado do jogo na região "vencedora" conforme definido pelas regras do xadrez. Esta previsão sobre consequências distantes, embora provada correta, não permitiu aos programadores prever o comportamento *local* do *Deep Blue* – sua resposta a um determinado ataque ao seu rei – porque o *Deep Blue* computa o mapa do jogo não-local, a ligação entre um movimento e suas possíveis consequências futuras, com mais precisão do que os programadores poderiam fazer (Yudkowsky 2006).

Os seres humanos modernos fazem literalmente milhões de coisas para se alimentar – para servir ao objetivo final de ser alimentado. Algumas dessas atividades foram "previstas pela Natureza" no sentido de ser um desafio ancestral ao qual nós estamos diretamente adaptados. Mas o nosso cérebro adaptado cresceu poderoso o suficiente para ser, de forma

significativa, aplicável de forma mais geral; permite-nos prever as consequências de milhões de diferentes ações em vários domínios, e exercer nossas preferências sobre os resultados finais. Os seres humanos cruzaram o espaço para colocar sua pegada na Lua, apesar de nenhum de nossos ancestrais ter encontrado um desafio análogo ao vácuo. Em relação ao domínio específico de IA, é um problema qualitativamente diferente projetar um sistema que vai operar com segurança em milhares de contextos, incluindo contextos que não sejam especificamente previstos por qualquer dos designers ou usuários, incluindo contextos que nenhum humano jamais encontrou. Neste momento não pode haver nenhuma especificação local de bom comportamento – não uma simples especificação sobre seus próprios comportamentos, não mais do que existe uma descrição local compacta de todas as maneiras que os seres humanos obtêm seu pão de cada dia.

Para construir uma IA que atua com segurança enquanto age em vários domínios, com muitas consequências, incluindo os problemas que os engenheiros nunca previram explicitamente, é preciso especificar o bom comportamento em termos como *"X tal que a conseqüência de X não é prejudicial aos seres humanos"*. Isto é não-local; e envolve extrapolar a consequência distante de nossas ações. Assim, esta é apenas uma especificação efetiva – que pode ser realizada como uma propriedade do *design* – se o sistema extrapola explicitamente as consequências de seu comportamento. Uma torradeira não pode ter essa propriedade de *design* porque uma torradeira não pode prever as consequências do pão tostado.

Imagine um engenheiro tendo que dizer: *"Bem, eu não tenho ideia de como esse avião que eu construí pode voar com segurança – de fato eu não tenho ideia de como ele fará tudo, se ele vai bater as asas ou inflar-se com hélio, ou outra coisa que eu nem sequer imagino, mas eu lhe asseguro, o projeto é muito, muito seguro"*. Isto pode parecer uma posição invejável da perspectiva de relações públicas, mas é difícil ver que outra garantia de comportamento ético seria possível para uma operação de inteligência geral sobre problemas imprevistos, em vários domínios, com preferências sobre consequências distantes. Inspecionando o *design* cognitivo podemos verificar que a mente estava, na verdade, buscando soluções que nós classificaríamos como éticas; mas não poderíamos prever que solução específica a mente descobriria.

Respeitar essa verificação exige alguma forma de distinguir as garantias de confiança (um procedimento que não desejo dizer "a IA é segura a menos que a IA seja realmente segura") de pura esperança e pensamen-

to mágico ("Não tenho ideia de como a Pedra Filosofal vai transformar chumbo em ouro, mas eu lhe asseguro, ela vai!"). Deve-se ter em mente que expectativas puramente esperançosas já foram um problema em pesquisa de IA (McDermott 1976).

Comprovadamente construir uma IAG de confiança exigirá métodos diferentes, e uma maneira diferente de pensar, para inspecionar uma falha no *software* de uma usina de energia[5] – ele exigirá um IAG que pensa como um engenheiro humano preocupado com a ética, não apenas um simples produto da engenharia ética.

Desta forma a disciplina de IA ética, especialmente quando aplicada à IAG, pode diferir fundamentalmente da disciplina ética de tecnologias não-cognitivas, em que:

- O comportamento específico local da IA não pode ser previsível independentemente de sua segurança, mesmo se os programadores fizerem tudo certo;
- Verificação de segurança do sistema torna-se um desafio maior, porque nós devemos verificar o comportamento seguro do sistema operando em todos os contextos;
- A própria cognição ética deve ser tomada como um assunto de engenharia.

Máquinas com *status* moral

Um diferente conjunto de questões éticas surge quando se contempla a possibilidade de que alguns futuros sistemas de IA possam ser candidatos a possuírem *status* moral. Nossas relações com os seres que possuem *status* moral não são exclusivamente uma questão de racionalidade instrumental: nós também temos razões morais para tratá-los de certas maneiras, e de nos *refrearmos de tratá-los de outras formas.* Francis Kamm propôs a seguinte definição do *status* moral, que servirá para nossos propósitos:

[5] N. T.: No texto original a expressão utilizada aqui é "*power station*", que pode designar uma central elétrica, uma estação geradora ou uma usina de energia. No centro de quase todas as estações de energia existe um gerador, uma máquina rotativa que converte energia mecânica em energia elétrica através da criação de movimento relativo entre um campo magnético e um condutor.

> X tem status moral = porque X conta moralmente em seu próprio direito, e é permitido/proibido fazer as coisas para ele para seu próprio bem. (Kamm 2007: cap. 7; paráfrase).

Uma pedra não tem *status* moral: podemos esmagá-la, pulverizá-la, ou submetê-la a qualquer tratamento que desejarmos sem qualquer preocupação com a própria rocha. Uma pessoa humana, por outro lado, deve ser encarada não apenas como um meio, mas também como um fim. Exatamente o que significa tratar uma pessoa como um fim é algo sobre a qual diferentes teorias éticas discordam; mas ela certamente envolve tomar os seus interesses legítimos em conta – atribuindo peso para o seu bem-estar – e também pode envolver aceitar severas restrições morais em nossa relações com ela, como a proibição contra assassiná-la, roubá-la, ou fazer uma série de outras coisas para ela ou para sua propriedade sem o seu consentimento. Além disso, é porque a pessoa humana é importante em seu próprio direito, e por seu bem estar, que estamos proibidos de fazer com ela essas coisas. Isso pode ser expresso de forma mais concisa, dizendo que uma pessoa humana tem *status* moral.

Perguntas sobre *status* moral são importantes em algumas áreas da ética prática. Por exemplo, as disputas sobre a legitimidade moral do aborto muitas vezes levam a desacordos sobre o *status* moral do embrião. Controvérsias sobre experimentação animal e o tratamento dispensado aos animais na indústria de alimentos envolvem questões sobre o *status* moral de diferentes espécies de animais. E as nossas obrigações em relação a seres humanos com demência grave, tais como pacientes em estágio final de Alzheimer, também podem depender de questões de *status* moral.

É amplamente aceito que os atuais sistemas de IA não têm *status* moral. Nós podemos alterar, copiar, encerrar, apagar ou utilizar programas de computador tanto quanto nos agradar, ao menos no que diz respeito aos próprios programas. As restrições morais a que estamos sujeitos em nossas relações com os sistemas contemporâneos de IA são todas baseadas em nossas responsabilidades para com os outros seres, tais como os nossos companheiros humanos, e não em quaisquer direitos para os próprios sistemas.

Embora seja realmente consensual que aos sistemas atuais de IA falta *status* moral, não está claro exatamente quais atributos servem de base para o *status* moral. Dois critérios são comumente propostos como importantemente relacionados com o estatuto moral, ou isoladamente ou

em combinação: a senciência e a sapiência (ou personalidade). Estes podem ser caracterizados aproximadamente como segue:

> **Senciência:** a capacidade para a experiência fenomenal ou *qualia*, como a capacidade de sentir dor e sofrer;
>
> **Sapiência:** um conjunto de capacidades associadas com maior inteligência, como a autoconsciência e ser um agente racional responsável.

Uma opinião comum é que muitos animais têm *qualia* e, portanto, têm algum *status* moral, mas que apenas os seres humanos têm sabedoria, o que lhes confere um *status* moral mais elevado do que possuem os animais não-humanos[6]. Esta visão, é claro, deve enfrentar a existência de casos limítrofes, tais como, por um lado, crianças ou seres humanos com grave retardo mental – às vezes, infelizmente referidos como "humanos marginais" – que não satisfazem os critérios de sapiência; e, por outro lado, alguns animais não-humanos, tais como os grandes símios, que podem ter pelo menos alguns dos elementos da sapiência. Alguns negam que o chamado "homem marginal" tenha um *status* moral pleno. Outros propõem maneiras adicionais em que um objeto poderia qualificar-se como um sustentador de *status* moral, tais como ser membro de uma espécie que normalmente tem sensibilidade ou sapiência, ou por estar em uma relação adequada para alguns seres que tem *status* moral independente (cf. Mary Anne Warren 2000). Para os propósitos do texto, no entanto, nos concentraremos nos critérios de sensibilidade e sapiência.

Esta imagem de *status* moral sugere que um sistema de IA terá algum *status* moral se ele tiver capacidades de *qualia*, tais como a capacidade de sentir dor. Um sistema de IA senciente, mesmo que não tenha linguagem e outras faculdades cognitivas superiores, não será como um bichinho de pelúcia ou um boneco; será mais como um animal vivo. É errado infligir dor a um rato, a menos que existam razões suficientemente fortes e razões morais prevalecentes para fazê-lo. O mesmo vale para qualquer sistema senciente de IA. Se além de consciência, um sistema de inteligência artificial também tiver sapiência de um tipo semelhante à de um adulto humano normal, então terá também pleno *status* moral, equivalente ao dos seres humanos.

---

[6] Alternativamente, se poderia negar que o estatuto moral vem em graus. Em vez disso, pode-se considerar que certos seres têm interesses mais importantes do que os outros seres. Assim, por exemplo, alguém poderia alegar que é melhor salvar um ser humano do que salvar um pássaro, não porque o ser humano tem maior *status* moral, mas porque o ser humano tem um interesse mais significativo em ter sua vida salva do que um pássaro.

Uma das ideias subjacentes a esta avaliação moral pode ser expressa de forma mais forte como um princípio de não discriminação:

Princípio da Não-Discriminação do Substrato

Se dois seres têm a mesma funcionalidade e a mesma experiência consciente, e diferem apenas no substrato de sua aplicação, então eles têm o mesmo *status* moral.

Pode-se argumentar a favor desse princípio, por razões de que rejeitá-lo equivaleria a adotar uma posição similar ao racismo: substrato carece de fundamental significado moral, da mesma forma e pela mesma razão que a cor da pele também carece. O Princípio da Não-Discriminação do Substrato não implica que um computador digital possa ser consciente, ou que possa ter a mesma funcionalidade que um ser humano. O substrato pode ser moralmente relevante na medida em que faz a diferença para a senciência, ou funcionalidade. Mas, mantendo essas coisas constantes, não faz diferença moral se um ser é feito de silício ou de carbono, ou se o cérebro usa semicondutores ou neurotransmissores.

Um princípio adicional que pode ser proposto é que o fato de que sistemas de IA sejam artificiais – ou seja, o produto de *design* deliberado – não é fundamentalmente relevante para o seu *status* moral. Nós poderíamos formular isto da seguinte forma:

Princípio da Não-Discriminação da Ontogenia

Se dois seres têm a mesma funcionalidade e mesma experiência de consciência, e diferem apenas na forma como vieram a existir, então eles têm o mesmo *status* moral.

Hoje essa ideia é amplamente aceita no caso de humanos – embora em alguns grupos de pessoas, particularmente no passado, a ideia de que é um *status* moral dependa de uma linhagem ou casta, tenha sido influente. Nós não acreditamos que fatores causais, tais como planejamento familiar, assistência ao parto, fertilização *in vitro*, seleção de gametas, melhoria deliberada da nutrição materna, etc. – que introduzem um elemento de escolha deliberada e *design* na criação de seres humanos – têm qualquer implicação necessária para o *status* moral da progênie. Mesmo aqueles que se opõem à clonagem para reprodução humana, por razões morais ou religiosas, em geral aceitam que, se um clone humano fosse trazido à existência, ele teria o mesmo *status* moral que qualquer outra criança

humana. O Princípio da Não-Discriminação da Ontogenia estende este raciocínio aos casos envolvendo sistemas cognitivos inteiramente artificiais.

É evidentemente possível, nas circunstâncias da criação, afetar a descendência resultante de maneira a alterar o seu *status* moral. Por exemplo, se algum procedimento realizado durante a concepção ou a gestação é a causa do desenvolvimento de um feto humano sem um cérebro, então este fato sobre a ontogenia seria relevante para o nosso julgamento sobre o *status* moral da prole. A criança anencefálica, porém, teria o mesmo *status* moral que qualquer outra similar criança anencefálica, incluindo aquela que tenha sido concebida através de um processo totalmente natural. A diferença de *status* moral entre uma criança anencefálica e uma criança normal está baseada na diferença qualitativa entre os dois – o fato de que um tem uma mente, enquanto o outro não. Desde que as duas crianças não tenham a mesma funcionalidade e a mesma experiência consciente, o Princípio da Não-Discriminação da Ontogenia não se aplica.

Embora o Princípio da Não-Discriminação da Ontogenia afirme que a ontogenia dos seres não tem qualquer relevância fundamental sobre o seu *status* moral, ela não nega que os fatos sobre a ontogênese podem afetar obrigações particulares que os agentes morais têm para com o ser em questão. Os pais têm deveres especiais para com seus filhos que eles não têm para com outras crianças, e que não teriam mesmo se houvesse outra criança qualitativamente idêntica a sua. Similarmente, o Princípio da Não-Discriminação da Ontogenia é consistente com a alegação de que os criadores ou proprietários de um sistema de IA com *status* moral podem ter direitos especiais para com sua mente artificial que não têm para com outras mentes artificiais, mesmo se as mentes em questão são qualitativamente semelhantes e têm o mesmo *status* moral.

Se os princípios de não discriminação com relação ao substrato e ontogenia são aceitos, então muitas questões sobre como devemos tratar mentes artificiais podem ser respondidas por aplicarmos os mesmos princípios morais que usamos para determinar nossos deveres em contextos mais familiares. Na medida em que os deveres morais decorrem de considerações sobre *status* moral, nós devemos tratar a mente artificial da mesma maneira como devemos tratar uma mente natural humana qualitativamente idêntica e em uma situação similar. Isto simplifica o problema do desenvolvimento de uma ética para o tratamento de mentes artificiais.

Mesmo se aceitarmos essa postura, no entanto, temos de enfrentar uma série de novas questões éticas que os princípios acima mencionados deixaram sem resposta. Novas questões éticas surgem porque mentes artificiais podem ter propriedades muito diferentes das ordinárias mentes humanas ou animais. Devemos considerar como essas novas propriedades afetariam o *status* moral de mentes artificiais e o que significaria respeitar o *status* moral de tais mentes exóticas.

Mentes com propriedades exóticas

No caso dos seres humanos, normalmente não hesitamos em atribuir sensibilidade e experiência consciente a qualquer indivíduo que apresenta os tipos normais de comportamento humano. Poucos acreditam que haja outras pessoas que atuem de forma perfeitamente normal, mas lhes falte consciência. No entanto, outros seres humanos não apenas se comportam como pessoas normais de maneiras semelhantes a nós mesmos; eles também têm cérebros e arquiteturas cognitivas que são constituídas de forma muito parecida com a nossa. Um intelecto artificial, pelo contrário, pode ser constituído um pouco diferentemente de um intelecto humano e ainda assim apresentar um comportamento semelhante ao humano ou possuir disposições comportamentais normalmente indicativas de personalidade. Por isso, seria possível conceber um intelecto artificial que seria sapiente, e talvez fosse uma pessoa, e ainda assim não estaria consciente ou teria experiências conscientes de qualquer tipo. (Se isto é realmente possível depende das respostas a algumas questões metafísicas não triviais). Se tal sistema fosse possível, ele levantaria a questão de saber se uma pessoa não-senciente teria qualquer *status* moral; e nesse caso, se teria o mesmo *status* moral de uma pessoa sensível. A senciência, ou pelo menos uma capacidade de senciência, é comumente assumida como estando presente em qualquer indivíduo que seja uma pessoa; esta questão não tem recebido muita atenção até o momento[7].

[7] Esta questão está relacionada com alguns problemas na filosofia da mente que têm recebido grande atenção, em particular o "problema do zumbi", que pode ser formulado da seguinte forma: Existe um mundo metafisicamente possível que seja idêntico ao mundo real no que diz respeito a todos os fatos físicos (incluindo a microestrutura física exata de todos os cérebros e organismos), mas que difere do mundo real em relação a alguns fatos fenomenais (experiência subjetiva)? Colocado de forma mais crua, é metafisicamente possível que possa haver uma pessoa que é fisicamente e exatamente idêntica a você, mas que é um "zumbi", ou seja, sem *qualia* e consciência fenomenal? (David Chalmers, 1996) Esta questão familiar é diferente do referido no texto: ao nosso "zumbi" é permitido ter sistematicamente diferentes propriedades físicas dos seres humanos normais. Além disso, queremos chamar a atenção especificamente ao *status* ético de um zumbi sapiente.

Outra propriedade exótica, o que certamente é metafisicamente e fisicamente possível para uma inteligência artificial, é que a *taxa subjetiva de tempo* desvia-se drasticamente da taxa que é característica de um cérebro biológico humano. O conceito de taxa subjetiva do tempo é melhor explicado primeiramente pela introdução da ideia de emulação de todo o cérebro, ou *"uploading"*.[8]

"Uploading" refere-se a uma hipotética tecnologia do futuro que permitiria a um intelecto humano ou de outro animal serem transferidos de sua aplicação original em um cérebro orgânico para um computador digital. Um cenário como este: Primeiro, uma alta resolução de varredura é realizada em algumas particularidades do cérebro, possivelmente destruindo o original no processo. Por exemplo, o cérebro pode ser vitrificado e dissecado em fatias finas, que podem então ser digitalizado usando alguma forma de microscópio de alta capacidade combinada com o reconhecimento automático de imagem. Podemos imaginar que esta análise deve ser detalhada o suficiente para capturar todos os neurônios, suas interconexões sinápticas, e outras características que são funcionalmente relevantes para as operações do cérebro original. Em segundo lugar, este mapa tridimensional dos componentes do cérebro e suas interconexões são combinados com uma biblioteca de avançadas teorias da neurociência, que especificam as propriedades computacionais de cada tipo básico de elementos, tais como diferentes tipos de neurônios e de junção sináptica. Terceiro, a estrutura computacional e os algoritmos associados de comportamento dos seus componentes são implementados em alguns computadores poderosos. Se o processo de *uploading* for bem sucedido, o programa de computador deve agora replicar as características funcionais essenciais do cérebro original. O resultante *upload* pode habitar uma realidade virtual simulada, ou, alternativamente, pode ser dado a ele o con-

---

[8] N. T.: Traduzir a palavra "*upload*" de forma literal traria problemas para o entendimento do texto. Por isso, consideramos mais apropriado manter a palavra em sua forma original. Um entendimento melhor do que significa *upload* nesse contexto pode ser encontrado em outro texto de Bostrom, agora em parceria com Anders Sandberg: *"The concept of brain emulation: Whole brain emulation, often informally called "uploading" or "downloading", has been the subject of much science fiction and also some preliminary studies (...). The basic idea is to take a particular brain, scan its structure in detail, and construct a software model of it that is so faithful to the original that, when run onappropriate hardware, it will behave in essentially the same way as the original brain"*. (*"Hole Brain Emulation". Disponível on-line:)* Numa tradução livre: "O conceito de emulação do cérebro: Emulação do Cérebro Inteiro, muitas vezes, informalmente chamado de "*upload*" ou "*download*", tem sido o objeto de muita ficção científica e também de alguns estudos preliminares (...). A ideia básica é ter um cérebro especial, digitalizar a sua estrutura em detalhe, e construir um modelo de *software* é que é tão fiel ao original que, quando executado em *hardware* apropriado, ele irá se comportar basicamente da mesma maneira como o cérebro original".

trole de um corpo robótico, que lhe permite interagir diretamente com a realidade física externa.

Uma série de perguntas surge no contexto de tal cenário: Quão plausível é que este procedimento um dia se torne tecnologicamente viável? Se o procedimento trabalhou e produziu um programa de computador exibindo aproximadamente a mesma personalidade, as mesmas memórias, e os mesmos padrões de pensamento que o cérebro original, pode este programa ser sensível? Será o *upload* a mesma pessoa que o indivíduo cujo cérebro foi desmontado no processo de carregamento? O que acontece com a identidade pessoal se um *upload* é copiado de tal forma que duas mentes semelhantes ou qualitativamente idênticas de *upload* sejam executadas em paralelo? Apesar de todas estas questões serem relevantes para a ética da IA, vamos aqui focar na questão que envolve a noção de uma taxa subjetiva de tempo.

Suponha que um *upload* possa ser senciente. Se executarmos o programa de transferência em um computador mais rápido, isso fará com que o *upload*, se ele estiver conectado a um dispositivo de entrada como uma câmera de vídeo, perceba o mundo externo como se esse estivesse perdendo velocidade. Por exemplo, se o *upload* está sendo executado milhares de vezes mais rápido que o cérebro original, então o mundo exterior será exibido para o *upload* como se fosse desacelerado por um fator de mil. Alguém deixa cair uma caneca física de café: O *upload* observa a caneca lentamente caindo no chão enquanto termina de ler o jornal pela manhã e envia alguns e-mails. Um segundo de tempo objetivo corresponde a 17 minutos do tempo subjetivo. Duração objetiva e subjetiva podem então divergir.

Tempo subjetivo não é a mesma estimativa ou percepção que um sujeito tem de quão rápido o tempo flui. Os seres humanos são muitas vezes confundidos com o fluxo do tempo. Podemos acreditar que se trata de 1 hora quando de fato são 2h15 min.; ou uma droga estimulante pode acelerar nossos pensamentos, fazendo parecer que mais tempo subjetivo tenha decorrido do que realmente é o caso. Estes casos mundanos envolvem uma percepção distorcida do tempo ao invés de uma mudança na taxa de tempo subjetivo. Mesmo em um cérebro ex-viciado em cocaína, provavelmente não há uma mudança significativa na velocidade de base nos cálculos neurológicos; mais provavelmente, a droga está fazendo o cérebro cintilar mais rapidamente a partir de um pensamento para outro, fazendo-o gastar menos tempo subjetivo para pensar um número maior de pensamentos distintos.

A variabilidade da taxa subjetiva do tempo é uma propriedade exótica de mentes artificiais que levanta novas questões éticas. Por exemplo, os casos em que a duração de uma experiência é eticamente relevante, devem ser mensurados durações no tempo objetivo ou subjetivo? Se um *upload* cometeu um crime e é condenado a quatro anos de prisão, deve este ser quatro anos objetivos - que pode corresponder a muitos milênios de tempo subjetivo - ou deve ser quatro anos subjetivos, que pode ser pouco mais de um par de dias de tempo objetivo? Se uma IA avançada e um ser humano estão com dor, é mais urgente aliviar a dor da IA, em razão de que ela experimenta uma maior duração subjetiva da dor para cada segundo sideral[9] que o alívio é retardado? Uma vez que em nosso contexto habitual, de humanos biológicos, tempo subjetivo não é significativamente variável, não é surpreendente que esse tipo de questionamento não seja francamente respondido por normas éticas familiares, mesmo se essas normas são estendidas a IA por meio de princípios de não-discriminação (como os propostos na seção anterior).

Para ilustrar o tipo de afirmação ética que pode ser relevante aqui, nós formulamos (mas não defendemos) um princípio de privilegiar tempo subjetivo como a noção normativa mais fundamental:

---

[9] N. T.: Tempo Sideral pode ser entendido como 'tempo estelar". Em nossas vidas e tarefas diárias costumamos utilizar o Tempo Solar, que tem como unidade fundamental o dia, ou seja, o tempo que o Sol demora para viajar 360 graus em torno do céu, devido a rotação da Terra. O Tempo Solar também possui unidades menores que são subdivisões de um dia:

1/24 Dia = 1 hora
1/60 Hora = 1 Minuto
1/60 Minuto = 1 Segundo

Mas, o Tempo Solar apresenta dificuldades pois a Terra não gira em torno de si 360° num Dia Solar. A Terra está em órbita ao redor do Sol, e ao longo de um dia, ele se move cerca de um grau ao longo de sua órbita (360 graus/365.25 dias para uma órbita completa = cerca de um grau por dia). Assim, em 24 horas, a direção em direção ao Sol varia em cerca de um grau. Portanto, a Terra só tem que girar 359° para fazer o Sol parecer que tem viajado 360° no céu. Em astronomia, é relevante quanto tempo a Terra leva para girar com relação as estrelas "fixas", por isso é necessário uma escala de tempo que remove a complicação da órbita da Terra em torno do Sol, e apenas se concentre em quanto tempo a Terra leva para girar 360° com relação às estrelas. Este período de rotação é chamado de Dia Sideral. Em média, é de 4 minutos a mais do que um Dia Solar, devido ao grau 1 extra que Terra tem de girar para completar 360°. Ao invés de definir um Dia Sideral em 24 horas e 4 minutos, nós definimos Horas Siderais, minutos e segundos que são a mesma fração de um dia como os seus homólogos Solar. Portanto, um segundo Sideral = 1,00278 Segundos Solar. O Tempo Sideral divide uma rotação completa da Terra em 24 Horas Siderais, é útil para determinar onde as estrelas estão em determinado momento.

Princípio da Taxa Subjetiva do Tempo

> Nos casos em que a duração de uma experiência tem um significado normativo básico, é a duração subjetiva da experiência que conta.

Até agora, discutimos duas possibilidades (sapiência não-senciente e a taxa subjetiva de tempo variável), que são exóticas no sentido de ser relativamente profunda e metafisicamente problemáticas, assim como faltam exemplos claros ou paralelos no mundo contemporâneo. Outras propriedades de possíveis mentes artificiais seriam exóticas em um sentido mais superficial; por exemplo, por serem divergentes em algumas dimensões quantitativas não problemáticas do tipo de mente com o qual estamos familiarizados. Mas tais características superficialmente exóticas também podem representar novos problemas éticos – se não ao nível fundamental da filosofia moral, ao menos no nível da ética aplicada ou para princípios éticos de complexidade média.

Um importante conjunto de propriedades exóticas de inteligências artificiais se relaciona com a reprodução. Certo número de condições empíricas que se aplicam à reprodução humana não são aplicáveis à IA. Por exemplo, as crianças humanas são o produto de uma recombinação do material genético dos dois genitores; os pais têm uma capacidade limitada para influenciar o caráter de seus descendentes; um embrião humano precisa ser gestado no ventre durante nove meses; leva de quinze a vinte anos para uma criança humana atingir a maturidade; a criança humana não herda as habilidades e os conhecimentos adquiridos pelos seus pais; os seres humanos possuem um complexo e evoluído conjunto de adaptações emocionais relacionados à reprodução, carinho, e da relação pais e filhos. Nenhuma dessas condições empíricas deve pertencer ao contexto de reprodução de uma máquina inteligente. Por isso, é plausível que muitos dos princípios morais de nível médio, que temos aceitado como as normas que regem a reprodução humana, precisarão ser repensados no contexto da reprodução de IA.

Para ilustrar por que algumas de nossas normas morais precisam ser repensadas no contexto da reprodução de IA, é suficiente considerar apenas uma propriedade exótica dos sistemas de IA: sua capacidade de reprodução rápida. Dado o acesso ao *hardware* do computador, uma IA poderia duplicar-se muito rapidamente, em menos tempo do que leva para fazer uma cópia do *software* da IA. Além disso, desde que a cópia da IA seja idêntica à original, ela nasceria completamente madura; e então a cópia pode começar a fazer suas próprias cópias imediatamente. Na ausên-

cia de limitações de *hardware*, uma população de IA poderia, portanto, crescer exponencialmente em uma taxa extremamente rápida, com um tempo de duplicação da ordem de minutos ou horas em vez de décadas ou séculos.

Nossas atuais normas éticas sobre a reprodução incluem uma versão de um princípio de liberdade reprodutiva, no sentido de que cabe a cada indivíduo ou ao casal decidir por si se quer ter filhos e quantos filhos desejam ter. Outra norma que temos (ao menos nos países ricos e de renda média) é que a sociedade deve intervir para prover as necessidades básicas das crianças nos casos em que seus pais são incapazes ou se recusam a fazê-lo. É fácil ver como estas duas normas poderia colidir no contexto de entidades com capacidade de reprodução extremamente rápida.

Considere, por exemplo, uma população de *uploads*, um dos quais acontece de ter o desejo de produzir um clã tão grande quanto possível. Dada à completa liberdade reprodutiva, este *upload* pode começar a copiar-se tão rapidamente quanto possível; e os exemplares que produz – que podem rodar em *hardware* de computador novo ou alugado pelo original, ou podem compartilhar o mesmo computador que o original, também vão começar a se auto-copiar, uma vez que eles são idênticos ao progenitor *upload* e compartilha seu desejo de produzir descendentes[10]. Logo, os membros do clã *upload* se encontrarão incapazes de pagar a fatura de eletricidade ou o aluguel para o processamento computacional e de armazenamento necessário para mantê-los vivos. Neste ponto, um sistema de previdência social poderá ser acionado para fornecer-lhes pelo menos as necessidades básicas para sustentar a vida. Mas, se a população crescer mais rápido que a economia, os recursos vão se esgotar; ao ponto de os *uploads* morrerem ou a sua capacidade para se reproduzir ser reduzida. (Para dois cenários distópicos relacionados, veja Bostrom (2004).).

Esse cenário ilustra como alguns princípios éticos de nível médio que são apropriados nas sociedades contemporâneas talvez precisem ser modificados se essas sociedades incluírem pessoas com a propriedade exótica de serem capazes de se reproduzir rapidamente.

---

[10] N. T.: No texto original a palavra que aqui aparece é *philoprogenic*. Não encontramos palavra equivalente em português, por isso optamos por traduzi-la da forma mais aproximada possível da intenção dos autores. Por exemplo, a palavra *philoprogenitive* significa "produzindo muitos descendentes, amar um filho ou crianças em geral, relativo ao amor à prole". Por aproximação, podemos dizer que populações de *uploads* desejarão produzir mais descendentes.

O ponto geral aqui é que quando se pensa em ética aplicada para contextos que são muito diferentes da nossa condição humana familiar, devemos ser cuidadosos para não confundir princípios éticos de nível médio com verdades normativas fundamentais. Dito de outro modo, nós devemos reconhecer até que ponto os nossos preceitos normativos comuns são implicitamente condicionados à obtenção de condições empíricas variadas, e à necessidade de ajustar esses preceitos de acordo com casos hipotéticos futuristas nos quais suas pré-condições não são obtidas. Por isso, não estamos fazendo uma afirmação polêmica sobre o relativismo moral, mas apenas destacando o ponto de senso comum de que o contexto é relevante para a aplicação da ética, e sugerindo que este ponto é especialmente pertinente quando se está considerando a ética de mentes com propriedades exóticas.

Superinteligência

I. J. Good (1965) estabeleceu a hipótese clássica sobre a superinteligência: que uma IA suficientemente inteligente para compreender a sua própria concepção poderia reformular-se ou criar um sistema sucessor, mais inteligente, que poderia, então, reformular-se novamente para tornar-se ainda mais inteligente, e assim por diante, em um ciclo de *feedback* positivo. Good chamou isso de "explosão de inteligência". Cenários recursivos não estão limitados a IA: humanos com inteligência aumentada através de uma interface cérebro-computador podem reprogramar suas mentes para projetar a próxima geração de interface cérebro-computador. (Se você tivesse uma máquina que aumentasse o seu QI, lhe ocorreria, uma vez que se tornou bastante inteligente, tentar criar uma versão mais poderosa da máquina.).

Super-inteligência também pode ser obtida através do aumento da velocidade de processamento. O mais rápido disparo de neurônios observado é de 1000 vezes por segundo; as fibras mais rápidas de axônio conduzem sinais a 150 metros/segundo, aproximadamente meio-milionésimo da velocidade da luz (Sandberg 1999). Aparentemente deve ser fisicamente possível construir um cérebro que calcula um milhão de vezes mais rápido que um cérebro humano, sem diminuir o seu tamanho ou reescrever o seu *software*. Se a mente humana for assim acelerada, um ano subjetivo do pensamento seria realizado para cada 31 segundos físicos no mundo exterior, e um milênio voaria em oito horas e meia. Vinge (1993) refere-se a

essas mentes aceleradas como “super-inteligência fraca”: uma mente que pensa como um ser humano, mas muito mais rápida.

Yudkowsky (2008a) enumera três famílias de metáforas para visualizarmos a capacidade de um IA mais inteligente que humanos:

- Metáforas inspiradas pelas diferenças de inteligência individuais entre os seres humanos: IA patenteará novas invenções, publicará inovadores trabalhos de pesquisa, ganhará dinheiro na bolsa, ou formará blocos de poder político.

- Metáforas inspiradas pelas diferenças de conhecimento entre as civilizações humanas do passado e do presente: IA mais rápida inventará recursos que comumente futuristas prevêem para as civilizações humanas de um século ou milênio no futuro, como a nanotecnologia molecular ou viagens interestelares.

- Metáforas inspiradas pelas diferenças da arquitetura de cérebro entre humanos e outros organismos biológicos: Por exemplo, Vinge (1993): “Imagine executar uma mente de um cão a uma velocidade muito alta. Será que mil anos de vida do cão pode ser somada a qualquer percepção humana?” Isto é: alterações da arquitetura cognitiva podem produzir *insights* que nenhuma mente do nível humano seria capaz de encontrar, ou talvez até mesmo representar, após qualquer período de tempo.

Mesmo se nos limitarmos às metáforas históricas, torna-se claro que a inteligência sobre-humana apresenta desafios éticos que são literalmente sem precedentes. Neste ponto, as apostas não são mais em escala individual (por exemplo, pedidos de hipotecas injustamente reprovados, casa incendiada, pessoa maltratada), mas em uma escala global ou cósmica (por exemplo, a humanidade é extinta e substituída por nada que nós consideramos de valor). Ou, se a super-inteligência pode ser moldada para ser benéfica, então, dependendo de suas capacidades tecnológicas, poderá trabalhar nos muitos problemas atuais que têm se revelado difíceis para a nossa inteligência de nível humano.

Super-inteligência é um dos vários “riscos existenciais” conforme definido por Bostrom (2002): um risco “onde um resultado adverso pode aniquilar permanentemente a vida inteligente originária da Terra ou limitar drasticamente o seu potencial”. Por outro lado, um resultado positivo da super-inteligência poderia preservar as formas de vida inteligentes originárias da Terra e ajudá-las a atingir o seu potencial. É importante ressaltar

que as mentes mais inteligentes representam grandes benefícios potenciais, bem como riscos.

As tentativas de raciocinar sobre riscos catastróficos globais podem estar suscetíveis a uma série de vieses[11] cognitivos (Yudkowsky 2008b), incluindo o "viés da boa história" proposto por Bostrom (2002):

> Suponha que nossas intuições sobre cenários futuros que são "plausíveis e realistas" sejam moldadas por aquilo que vemos na televisão, nos filmes e pelo que lemos nos romances. (Afinal, uma grande parte do discurso sobre o futuro que as pessoas encontram é em forma de ficção e outros contextos recreativos). Devemos então, quando pensarmos criticamente, suspeitar de nossas intuições, de sermos tendenciosos no sentido de superestimar a probabilidade desses cenários que fazem uma boa história, uma vez que tais situações parecerão muito mais familiares e mais "reais". Esse *viés da boa história* pode ser muito poderoso. Quando foi a última vez que viu um filme sobre a humanidade em que os humanos são extintos de repente (sem aviso e sem ser substituído por alguma outra civilização)? Embora esse cenário possa ser muito mais provável do que um cenário no qual heróis humanos repelem, com sucesso, uma invasão de monstros ou de guerreiros robôs, não seria muito divertido de assistir.

Na verdade resultados desejáveis fazem filmes pobres: Sem conflito significa sem história. Enquanto as Três Leis da Robótica de Asimov (Asimov 1942) são muitas vezes citadas como um modelo de desenvolvimento ético para IA, as Três Leis são mais como um enredo para as trama com os "cérebros positrônicos" de Asimov. Se Asimov tivesse representado as três leis como bom trabalho, ele não teria obtido nenhuma história.

Seria um erro considerar sistemas de "IA", como uma espécie com características fixas, e perguntar "eles vão ser bons ou maus?" O termo "Inteligência Artificial" refere-se a um vasto espaço de projeto, provavelmente muito maior do que o espaço da mente humana (uma vez que todos os seres humanos compartilham uma arquitetura cerebral comum). Pode ser uma forma de "viés da boa história" perguntar: "Será que sistemas de IA são bons ou maus?", como se estivesse tentando pegar uma premissa

---

[11] N. T.: Um *viés cognitivo* é uma tendência inerente ao comportamento humano em cometer desvios sistemáticos de racionalidade, ao pensar ou analisar determinadas situações. Nossos mecanismos cognitivos (isto é, mecanismos de pensamento, raciocínio, inferência etc) são enviesados, ou seja, viciados em determinadas direções, nos tornando mais propensos a cometer certos tipos de erros, por exemplo, de identificação ou de estimação de tempo, probabilidades, etc.

para um enredo de filme. A resposta deve ser: “Exatamente sobre qual *design* de IA você está falando?”

Pode o controle sobre a programação inicial de uma Inteligência Artificial ser traduzido em influência sobre o seu efeito posterior no mundo? Kurzweil (2005) afirma que “[i]nteligência é inerentemente impossível de controlar”, e que, apesar das tentativas humanas de tomar precauções, “por definição... entidades inteligentes têm a habilidade de superar essas barreiras facilmente”. Suponhamos que a IA não é apenas inteligente, mas que, como parte do processo de se melhorar a sua própria inteligência, tenha livre acesso ao seu próprio código fonte: ela pode reescrever a si mesma e se tornar qualquer coisa que quer ser. No entanto, isso não significa que o IA deve *querer* se reescrever de uma forma hostil.

Considere Gandhi, que parece ter possuído um desejo sincero de não matar pessoas. Gandhi não conscientemente toma uma pílula que o leva a querer matar pessoas, porque Gandhi sabe que se ele quiser matar as pessoas, ele provavelmente vai matar pessoas, e a versão atual do Gandhi não quer matar. Em termos mais gerais, parece provável que a maioria das mentes mais auto-modificadoras irão naturalmente ter funções de utilidade estáveis, o que implica que uma escolha inicial do projeto da mente pode ter efeitos duradouros (Omohundro 2008).

Neste ponto no desenvolvimento da ciência da IA, existe alguma maneira em que podemos traduzir a tarefa de encontrar um *design* para “bons” sistemas de IA em uma direção da pesquisa moderna? Pode parecer prematuro especular, mas se faz suspeitar que alguns paradigmas de IA são mais prováveis do que outros para eventualmente provar que propiciem a criação de agentes inteligentes de auto-modificação, cujos objetivos continuem a ser previsíveis mesmo depois de várias interações de auto-aperfeiçoamento. Por exemplo, o ramo Bayesiano da IA, inspirado por coerentes sistemas matemáticos, como o da teoria da probabilidade e da maximização da utilidade esperada, parece mais favorável para o problema de auto-modificação previsível do que a programação evolutiva e algoritmos genéticos. Esta é uma afirmação polêmica, mas ilustra o ponto de que, se formos pensar no desafio da super-inteligência, isso pode na verdade ser transformado em um conselho direcional para as pesquisas atuais em IA.

No entanto, mesmo admitindo que possamos especificar um objetivo de IA a ser persistente sob auto-modificação e auto-aperfeiçoamento, este só começa a tocar nos problemas fundamentais da ética para criação da

super-inteligência. Os seres humanos, a primeira inteligência geral a existir na Terra, têm usado a inteligência para remodelar substancialmente a escultura do globo – esculpir montanhas, domar os rios, construir arranha-céus, agricultura nos desertos, produzir mudanças climáticas não-intencionais no planeta. Uma inteligência mais poderosa poderia ter consequências correspondentemente maiores.

Considere novamente a metáfora histórica para a super-inteligência – diferenças semelhantes às diferenças entre as civilizações passadas e presentes. Nossa civilização atual não está separada da Grécia antiga somente pela ciência aperfeiçoada e aumento de capacidade tecnológica. Há uma diferença de perspectivas éticas: os gregos antigos pensavam que a escravidão era aceitável, nós pensamos o contrário. Mesmo entre os séculos XIX e XX, houve substanciais divergências éticas – as mulheres devem ter direito ao voto? Os negros podem votar? Parece provável que as pessoas de hoje não serão vistas como eticamente perfeitas por civilizações futuras, não apenas por causa da nossa incapacidade de resolver problemas éticos reconhecidos atualmente, como a pobreza e a desigualdade, mas também por nosso fracasso até mesmo em reconhecer alguns problemas éticos. Talvez um dia o ato de sujeitar as crianças involuntariamente à escolaridade será visto como abuso infantil – ou talvez permitir que as crianças deixem a escola aos 18 anos será visto como abuso infantil. Nós não sabemos.

Considerando a história da ética nas civilizações humanas ao longo dos séculos, podemos ver que se poderia tornar uma tragédia muito grande criar uma mente que ficou estável em dimensões éticas ao longo da qual as civilizações humanas parecem exibir *mudança direcional*. E se Arquimedes de Siracusa tivesse sido capaz de criar uma inteligência artificial de longa duração com uma versão estável do código moral da Grécia Antiga? Mas, evitar esse tipo de estagnação ética é comprovadamente complicado: não seria suficiente, por exemplo, simplesmente tornar a mente aleatoriamente instável. Os gregos antigos, mesmo que tivessem percebido suas próprias imperfeições, não poderiam ter feito melhor mesmo jogando dados. Ocasionalmente uma boa e nova ideia em ética vem acompanhada de uma surpresa; mas a maioria das ideias geradas aleatoriamente traz mudanças éticas que nos parecem loucura ou rabiscos incompreensíveis.

Isto nos apresenta talvez o último desafio das máquinas éticas: Como construir uma IA que, quando executada, torna-se mais ética do que você? Isto não é como pedir a nossos próprios filósofos para produzir uma super-ética, mais do que o *Deep Blue* foi construído fazendo com que os

melhores jogadores humanos de xadrez programassem boas jogadas. Mas temos de ser capazes de efetivamente descrever a questão, se não a resposta – jogar dados não irá gerar bons movimentos do xadrez, ou boa ética tampouco. Ou, talvez, uma maneira mais produtiva de pensar sobre o problema: Qual a estratégia que você gostaria que Arquimedes seguisse na construção de uma super-inteligência, de modo que o resultado global ainda seria aceitável, se você não pudesse lhe dizer especificamente o que estava fazendo de errado? Esta é a situação em que estamos em relação ao futuro.

Uma parte forte do conselho que emerge, considerando nossa situação análoga à de Arquimedes, é que não devemos tentar inventar uma versão "super" do que nossa civilização considera ética; esta não é a estratégia que gostaríamos que Arquimedes seguisse. Talvez a pergunta que devemos considerar, sim, é como uma IA programada por Arquimedes – sem maior experiência moral do que Arquimedes – poderia reconhecer (pelo menos em parte) nossa própria civilização ética como progresso moral, em oposição à simples instabilidade moral. Isso exigiria que começássemos a compreender a estrutura de questões éticas da maneira que já compreendemos a estrutura do xadrez.

Se nós somos sérios sobre o desenvolvimento de uma IA avançada, este é um desafio que devemos enfrentar. Se as máquinas estão a ser colocadas em posição de ser mais fortes, mais rápidas, mais confiáveis, ou mais espertas que os humanos, então a disciplina de máquinas éticas deve se comprometer a buscar refinamento humano superior (e não apenas seres humanos equivalentes).[12]

Conclusão

Embora a IA atual nos ofereça algumas questões éticas que não estão presentes no *design* de automóveis ou de usinas de energia, a abordagem de algoritmos de inteligência artificial em relação a um pensamento mais humano prenuncia complicações desagradáveis. Os papéis sociais podem ser preenchidos por meio de algoritmos de IA, o que implica novas exigências de projeto, como transparência e previsibilidade. Suficientes algoritmos de IAG já não podem executar em contextos previsíveis, e exigem novos tipos de garantia de segurança e engenharia, e de considera-

[12] Os autores são gratos a Rebecca Roache pelo auxílio à pesquisa e aos editores deste volume por comentários detalhados a uma versão anterior do nosso manuscrito.

ções da ética artificial. Sistemas de IA com estados mentais suficientemente avançados, ou o tipo certo de estados, terão um *status* moral, e alguns poderão ser considerados como pessoas – embora talvez pessoas muito diferentes do tipo que existe agora, talvez com regras diferentes. E, finalmente, a perspectiva de IA com inteligência sobre-humana, e habilidades sobre-humanas, nos apresenta o desafio extraordinário de indicar um algoritmo que gere comportamento super ético. Esses desafios podem parecer visionários, mas parece previsível que vamos encontrá-los, e eles não são desprovidos de sugestões para os rumos da pesquisa atual.

Biografia dos autores

Nick Bostrom é professor na *Faculty of Philosophy at University of Oxford* e diretor do *Future of Humanity Institute* no *Martin Oxford School*. Ele é autor de cerca de 200 publicações, incluindo *Anthropic Bias* (*Routledge*, 2002), *Global Catastrophic Risks* (ed. OUP, 2008), e *Enhancing Humans* (ed., OUP, 2009). Sua pesquisa abrange uma série de questões de grande relevância para a humanidade. Ele está atualmente trabalhando num livro sobre o futuro da inteligência artificial e suas implicações estratégicas.

Eliezer Yudkowsky é um pesquisador do *Singularity Institute for Artificial Intelligence*, onde trabalha em tempo integral sobre os problemas previsíveis da arquitetura em auto-melhoria de IA. Seu atual trabalho centra-se em modificar a teoria da decisão clássica para descrever uma forma coerente de auto-modificação. Ele também é conhecido por seus escritos populares sobre questões da racionalidade humana e bias cognitivos.

Leituras

Bostrom, N. 2004. *"The Future of Human Evolutian"*, em *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy (Palo Alto, Califórnia: Ria *University Press*). - Este trabalho explora algumas dinâmicas evolutivas que poderiam levar a uma população de *uploads* para desenvolver em direções distópicas.

Yudkowsky, E. 2008a. *'Artificial Intelligence as a Positive and Negative Factor in Global Risk'*, in Bostrom and Cirkovic (eds.), pp. 308-345. - Uma introdução aos riscos e desafios apresentados pela possibilidade de melhorar a auto-recursividade das máquinas superinteligentes.

Wendell, W. 2008. *'Moral Machines: Teaching Robots Right from Wrong'*(Oxford University Press, 2008). – Uma pesquisa global de desenvolvimento recente.

Referências

ASIMOV, I. 1942. 'Runaround', *Astounding Science Fiction*, March 1942.

BEAUCHAMP, T. and Chilress, J. *Principles of Biomedical Ethics*. Oxford: Oxford University Press.

BOSTROM, N. 2002. 'Existential Risks: Analyzing Human Extinction Scenarios', *Journal of Evolution and Technology* 9
(**http://www.nickbostrom.com/existential/risks.html**).

BOSTROM, N. 2003. 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', *Utilitas* 15: 308-314.

BOSTROM, N. 2004. 'The Future of Human Evolution', in *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy (Palo Alto, California: Ria University Press)
(**http://www.nickbostrom.com/fut/evolution.pdf**)

BOSTROM, N. and CIRKOVIC, M. (eds.) 2007. *Global Catastrophic Risks*. Oxford: Oxford University Press.

CHALMERS, D. J., 1996, *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford University Press

HIRSCHFELD, L. A. and GELMAN, S. A. (eds.) 1994. *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge: Cambridge University Press.

GOERTZEL, B. and PENNACHIN, C. (eds.) 2006. *Artificial General Intelligence*. New York, NY: Springer-Verlag.

GOOD, I. J. 1965. 'Speculations Concerning the First Ultraintelligent Machine', in Alt, F. L. and Rubinoff, M. (eds.) *Advances in Computers*, 6, New York: Academic Press. Pp. 31-88.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. 2001. *The Elements of Statistical Learning*. New York, NY: Springer Science.

HENLEY, K. 1993. 'Abstract Principles, Mid-level Principles, and the Rule of Law', *Law and Philosophy* 12: 121-32.

HOFSTADTER, D. 2006. 'Trying to Muse Rationally about the Singularity Scenario', presented at the *Singularity Summit at Stanford*, 2006.

HOWARD, Philip K. 1994. *The Death of Common Sense: How Law is Suffocating America*. New York, NY: Warner Books.

KAMM, F. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

KURZWEIL, R. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York, NY: Viking.

McDERMOTT, D. 1976. 'Artificial intelligence meets natural stupidity', *ACM SIGART Newsletter* 57:4-9.

OMOHUNDRO, S. 2008. 'The Basic AI Drives', *Proceedings of the AGI-08 Workshop*. Amsterdam: IOS Press. Pp. 483-492.

SANDBERG, A. 1999. 'The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains', *Journal of Evolution and Technology*, 5.

VINGE, V. 1993. 'The Coming Technological Singularity', presented at the *VISION-21 Symposium*, March, 1993.

WARREN, M. E. 2000. *Moral Status: Obligations to Persons and Other Living Things*. Oxford: Oxford University Press.

YUDKOWSKY, E. 2006. 'AI as a Precise Art', presented at the *2006 AGI Workshop* in Bethesda, MD.

YUDKOWSKY, E. 2008a. 'Artificial Intelligence as a Positive and Negative Factor in Global Risk', in Bostrom and Cirkovic (eds.), pp. 308-345.

YUDKOWSKY, E. 2008b. 'Cognitive biases potentially affecting judgment of global risks', in Bostrom and Cirkovic (eds.), pp. 91-119.

Notas

*Texto traduzido por Pablo Araújo Batista. Revisado por Diego Caleiro e Lauro Edison.

O original pode ser lido aqui:
http://www.nickbostrom.com/ethics/artificial-intelligence.pdf