A Vontade Superinteligente

Motivação e Racionalidade Instrumental em Agentes Artificiais Avançados

por Nick Bostrom, 2012*

Tradução de Lucas Machado

(2012) Nick Bostrom
Faculty of Philosophy & Oxford Martin School
University of Oxford
www.nickbostrom.com
[Forthcoming in *Minds and Machines*, 2012]
[Original em **pdf**]

RESUMO

Este artigo discute a relação entre inteligência e motivação em agentes artificiais, desenvolvendo e argumentando brevemente a favor de duas teses. A primeira, a *tese da ortogonalidade*, defende (com algumas reservas) que *inteligência* e *objetivos finais* (propósitos) são eixos ortogonais ao longo dos quais intelectos artificiais podem variar livremente – mais ou menos qualquer nível de inteligência poderia ser combinado com mais ou menos de qualquer objetivo final. A segunda tese, a *tese da convergência instrumental*, defende que, desde que eles possuam um nível suficiente de inteligência, agentes tendo qualquer extensão de objetivos finais irão seguir objetivos intermediários similares porque eles têm razões instrumentais para fazê-lo. Combinadas, essas duas teses nos ajudam a entender a extensão possível do comportamento de agentes superinteligentes e apontam para alguns perigos em potencial de construir tais agentes.

Palavras-chave: superinteligência, inteligência artificial, IA, objetivo, razão instrumental, agente inteligente.

1. A ortogonalidade da motivação e inteligência

1.1 Evitando o antropomorfismo

Se nós imaginamos um espaço em que todas as mentes possíveis podem ser representadas, nós precisamos imaginar que todas as mentes *humanas* constituem um grupo consideravelmente pequeno dentro desse espaço. As diferenças de personalidade entre Hannah Arendt e Benny Hill podem parecer vastas para nós, mas isso se deve à barra de escala de nosso julgamento intuitivo estar calibrada para a distribuição humana existente. No espaço mais amplo de todas as possibilidades lógicas, essas duas personalidades são vizinhas bem próximas. Em termos de arquitetura neural, pelo menos, a senhorita Arendt e o senhor Hill são praticamente idênticos. Imagine seus cérebros lado a lado em repouso silencioso. As diferenças pareceriam mínimas e você os reconheceria prontamente como sendo do mesmo tipo; talvez você fosse até mesmo incapaz de dizer qual cérebro é de quem. Se você estudasse a morfologia do cérebro mais detalhadamente em um microscópio, a impressão da similaridade fundamental seria fortalecida: você veria a mesma organização lamelar do córtex, feita dos mesmos tipos de neurônios, coberta no mesmo banho de moléculas neurotransmissoras[1].

É bem sabido que observadores ingênuos frequentemente antropomorfizam as capacidades de sistemas inanimados [*insensate*]. Nós podemos dizer, por exemplo, "Essa máquina está demorando demais para pensar sobre meu chocolate quente". Isso pode levar alguém a subestimar a complexidade cognitiva de capacidades que vêm naturalmente para seres humanos, como controle motor e percepção sensorial ou, alternativamente, atribuir graus significantes de consciência [*mindfulness*] e inteligência para sistemas bem estúpidos, tais como os robôs falantes [*chatterboxes*] como a ELIZA de Weizenbaum's (Weizembaum 1976). Similarmente, há uma tendência comum a antropomorfizar as *motivações* de sistemas inteligentes nos quais não há verdadeiramente nenhuma razão para esperar por impulsos e paixões semelhantes às humanas ("Meu carro não queria mesmo engatar esta manhã"). Eliezer Yudkowsky dá uma boa ilustração desse fenômeno:

[1] Isso não nega, é claro, que diferenças que parecem pequenas visualmente podem ser funcionalmente profundas.

> Na era de ficção científica *pulp*, capas de revista ocasionalmente retratavam um alienígena consciente monstruoso – coloquialmente conhecido como BEM (*bug-eyed monster,* monstro com olhos de inseto) – levando uma humana atraente em um vestido rasgado. Aparentemente, o artista acreditava que um alienígena não-humanoide, com uma história evolutiva completamente diferente, desejaria sexualmente mulheres humanas... Provavelmente, o artista não se perguntou se um inseto gigante *percebe* mulheres humanas de modo a considera-las atraentes. Pelo contrário, uma mulher humana em um vestido rasgado é *sexy* – e inerentemente, como uma propriedade intrínseca. Quem cometeu esse engano não pensou sobre a mente insectóide: eles se concentraram no vestido rasgado da mulher. Se o vestido não estivesse rasgado, a mulher seria menos sexy; BEM não é levado em conta. (Yudkowsky 2008).

Uma inteligência artificial pode ser bem menos semelhante aos homens em suas motivações do que um alienígena do espaço. O extraterrestre (assim assumiremos) é uma criatura biológica que surgiu por meio de um processo evolutivo e, por isso, pode-se esperar dele que tenha os tipos de motivações típicos de criaturas evoluídas. Por exemplo, não seria uma grande surpresa descobrir que algum alienígena inteligente aleatório teria motivações relacionadas com obter ou evitar comida, ar, temperatura, gasto de energia, a ameaça de um dano ao corpo, doença, predadores, reprodução, ou proteção de descendentes. Um membro de uma espécie inteligente e social também pode ter motivações relacionadas à cooperação e competição; como nós, ele pode mostrar lealdade de grupo, ressentimento em relação a parasitas, e talvez até mesmo uma preocupação com sua reputação e aparência.

Em contraste, uma mente inteligente não precisa se preocupar intrinsecamente com qualquer uma dessas coisas, nem mesmo no menor grau. Pode-se facilmente conceber uma inteligência artificial cujo único objetivo é contar o número de grãos de areia em Boracay, ou calcular os decimais de pi indefinidamente, ou maximizar o número total de clipes de papel em seu cone de luz futuro. De fato, seria mais fácil criar uma IA com objetivos simples como esses do que construir uma que tenha um conjunto de valores e disposições semelhantes aos humanos.

1.2 A tese da ortogonalidade

Para nossos propósitos, entenderemos grosseiramente que a "inteligência" corresponde à capacidade para raciocínio instrumental (mais so-

bre isso mais tarde). Busca inteligente por planos e políticas instrumentalmente ideais pode ser realizada a serviço de qualquer objetivo. Inteligência e motivação podem, nesse sentido, ser pensadas como um par de eixos ortogonais em um gráfico cujos pontos representam um agente artificial logicamente possível, exceto por algumas restrições fracas – por exemplo, pode ser impossível para um sistema muito pouco inteligente ter motivações muito complexas, já que motivações complexas colocariam uma exigência significativa à memória. Além disso, para que o agente "tenha" um conjunto de motivações, esse conjunto pode precisar funcionar integrado com os processos de decisão do agente, o que, novamente, colocaria exigências ao poder de processamento e, talvez, à inteligência. Para mentes que podem modificar a si próprias, também pode haver restrições mecânicas; por exemplo, uma mente inteligente com um desejo urgente de ser estúpido pode não continuar inteligente por muito tempo. Mas essas qualificações não devem obscurecer a ideia principal, que nós podemos expressar da seguinte maneira:

> *A Tese da Ortogonalidade*
>
> Inteligência e objetivos finais são eixos ortogonais ao longo dos quais agentes possíveis podem variar livremente. Em outras palavras, mais ou menos qualquer nível de inteligência poderia, a princípio, ser combinado com mais ou menos qualquer objetivo final.

Uma comparação poderia ser feita aqui com a teoria Humeana da motivação. David Hume pensava que apenas crenças (por exemplo, sobre o que é algo bom a se fazer) não podem motivar ações: algum desejo é requerido.[2] Isso sustentaria a tese da ortogonalidade minando uma possível objeção a ela, a saber, que inteligência o suficiente pode implicar a aquisição de certas crenças, e essas crenças necessariamente produzem certas motivações. De modo algum, de acordo com David Hume: crença e motivação são coisas separadas.

Por mais que a tese da ortogonalidade possa conseguir apoio da teoria Humeana da motivação, ela não a pressupõe. Especificamente, não é necessário defender que apenas crenças *nunca* podem motivar a ação. Seria suficiente assumir, por exemplo, que um agente – seja qual for sua inteligência – pode ser motivado a seguir um curso de ação *se* ocorrer do agente ter certos desejos firmes com força suficiente para se sobreporem.

[2] Para algumas tentativas recentes de defender a teoria Humeana de motivação, ver Smith (1987), Lewis (1988) e Sinhababu (2009).

Outro modo por meio do qual a tese da ortogonalidade poderia ser verdadeira, mesmo se a teoria Humeana da motivação for falsa, é se uma inteligência arbitrariamente alta não implicar a aquisição de quaisquer tais crenças que são (supostamente) motivadoras por si mesmas. Um terceiro modo pelo qual seria possível que a tese da ortogonalidade fosse verdadeira mesmo que a teoria Humeana da motivação fosse falsa é se for possível construir um sistema cognitivo (ou, mais neutralmente, um "processo de otimização") com uma inteligência arbitrariamente alta, mas com uma constituição tão alienígena que não teria nenhum análogo funcional ao que nós humanos chamamos de "crenças" e "desejos". Esse seria o caso se tal sistema pudesse ser construído de tal maneira que o faria motivado a possuir qualquer objetivo final dado.

A tese da ortogonalidade, tal como ela é formulada aqui, faz uma afirmação sobre a relação entre motivação e *inteligência*, em vez de entre motivação e *racionalidade* (ou motivação e *razão*). Isso se deve ao fato de alguns filósofos usarem a palavra "racionalidade" para conotar um conceito mais "normativamente espesso" do que aquele que nós queremos conotar aqui com a palavra "inteligência". Por exemplo, em *Razões e Pessoas*, Derek Parfit argumenta que certas preferências básicas seriam irracionais, tal como a de um agente normal em todos os outros aspectos que tem uma "Indiferença à Terça Futura":

> Certo hedonista se preocupa muito com a qualidade de suas experiências futuras. Com uma exceção, ele se preocupa igualmente com todas as partes de seu futuro. A exceção é que ele tem uma Indiferença à Terça Futura. Ao longo de toda Terça ele se preocupa normalmente com o que está acontecendo com ele. Mas ele nunca se preocupa sobre as dores ou prazeres de uma terça futura... Essa indiferença é um simples fato. Quando ele está planejando seu futuro, é simplesmente verdade que ele sempre prefere o prospecto de sofrer muito em uma terça do que sofrer moderadamente em qualquer outro dia. (Parfit, 1984)[3]

Portanto, o agente é, nesse instante, indiferente ao seu sofrimento futuro se e somente se ele ocorrer em uma terça futura. Para nossos propósitos, nós não precisamos nos posicionar quanto a Parfit estar certou ou não de que isso é irracional, desde que nós concedamos que isso não é necessariamente não-inteligente. Por "inteligência", aqui, nós queremos dizer algo como *racionalidade instrumental* – talento com previsões, planejamento, e raciocínio sobre meios e fins de uma forma geral. O agente imaginário indiferente à terça futura de Parfit pode ter uma racionalidade

[3] Ver também Parfit (2011).

instrumental impecável, e, portanto, ter uma grande inteligência, mesmo se ele deixar a desejar em algum tipo de sensibilidade à "razão objetiva" que possa ser requerida para um agente completamente racional. Consequentemente, esse tipo de exemplo não mina a tese da ortogonalidade.

De modo similar, mesmo se houver fatos morais objetivos que qualquer agente completamente racional compreenderia, e mesmo se esses fatos morais forem intrinsecamente motivadores (tais que qualquer um que os compreender necessariamente é motivado a agir de acordo com eles), isso não necessariamente mina a tese da ortogonalidade. Ela ainda poderia ser verdadeira se um agente pudesse ter uma racionalidade *instrumental* impecável, mesmo que lhe falte alguma outra faculdade constitutiva da racionalidade propriamente dita, ou alguma faculdade requerida para a compreensão completa de fatos morais objetivos. (Um agente também poderia ser extremamente inteligente, mesmo superinteligente, sem ter uma racionalidade instrumental completa em todo domínio).

Um motivo para nos concentrarmos na inteligência, isto é, na racionalidade instrumental, é que esse é o conceito mais relevante, se nós estivermos tentando descobrir o que diferentes tipos de sistemas fariam. Questões normativas, como se o seu comportamento contaria como um comportamento prudencialmente racional ou moralmente justificado, podem ser importantes de várias maneiras. Entretanto, tais questões não devem nos cegar à possibilidade de sistemas cognitivos que falham em satisfazer critérios normativos substanciais mas são, de todo modo, muito poderosos e capazes de ter grande influência no mundo.[4]

1.3 Prevendo o comportamento e a motivação superinteligentes

A tese da ortogonalidade implica que mentes sintéticas podem ter objetivos completamente não-antropomórficos – objetivos a nosso ver tão bizarros como contar grãos de areia ou maximizar os clipes de papel. Isso se mantém verdadeiro mesmo para (de fato, *especialmente* para) agentes artificias que são extremamente inteligentes ou superinteligentes. Mas não se segue da tese da ortogonalidade que seja impossível fazer previsões so-

[4] A tese da ortogonalidade implica que praticamente qualquer combinação de objetivos finais e níveis de inteligência é logicamente possível; ela *não* implica que seria fácil, na prática, dotar um agente superinteligente com algum objetivo final arbitrário ou que respeitasse aos humanos – mesmo se nós soubéssemos como construir a parte da inteligência. Para algumas notas preliminares sobre o problema do carregamento de valores, ver, por exemplo, Dewey (2011) e Yudkowsky (2011).

bre o que agentes particulares farão. Previsibilidade é importante se buscamos projetar um sistema para alcançar resultados específicos, e o problema se torna mais importante quanto mais poderosa for a inteligência artificial em questão. Agentes superinteligentes poderiam ser *extremamente* poderosos, então, é importante desenvolver um modo de analisar e prever seu comportamento. Ainda assim, apesar da independência entre inteligência e objetivos finais, implicada pela tese da ortogonalidade, o problema de prever o comportamento de um agente não precisa ser intratável – nem mesmo no que diz respeito a agentes superinteligentes cuja complexidade cognitiva e características de performance pode torna-los, em certos aspectos, opacos à análise humana.

Há pelo menos três direções a partir das quais é possível abordar o problema de prever a motivação superinteligente:

(1) *Previsibilidade por competência de desenvolvimento.* Se nós pudermos supor que os desenvolvedores de um agente superinteligente podem construir de maneira bem sucedida o sistema de objetivos do agente para que ele persiga estavelmente um objetivo particular determinado pelos programadores, então, uma previsão que nós podemos fazer é que o agente irá perseguir esse objetivo. Quanto mais inteligente for o agente, maiores os recursos cognitivos que ele terá para perseguir esse objetivo. Então, mesmo antes de um agente ter sido criado, nós talvez sejamos capazes de prever algo sobre seu comportamento, se soubermos algo sobre quem o construirá e quais objetivos eles quererão que ele tenha.

(2) *Previsibilidade por herança.* Se uma inteligência digital for criada diretamente a partir de um modelo humano (como seria o caso em uma emulação completa do cérebro com alta fidelidade), então, a inteligência digital poderia herdar algumas das motivações do modelo humano.[5] O agente pode manter algumas dessas motivações mesmo se suas capacidades cognitivas forem subsequentemente aprimoradas para fazê-lo superinteligente. Esse tipo de inferência requer cuidado. Os objetivos e valores do agente poderiam facilmente ser corrompidos no processo de *upload* ou durante a operação e o aprimoramento subsequente, dependendo de como o procedimento for aplicado.

(3) *Previsibilidade por meio de razões instrumentais convergentes.* Mesmo sem conhecimento detalhado dos objetivos finais de um agente, nós talvez sejamos capazes de inferir algo sobre seus objetivos mais imediatos ao considerar as razões *instrumentais* que

[5] Ver Sandberg & Bostrom (2008).

> surgiriam para qualquer uma de um de uma ampla faixa de objetivos finais em uma ampla faixa de situações. Essa maneira de fazer previsões se torna mais útil quanto maior for a inteligência de um agente, porque um agente mais inteligente tem mais chances de reconhecer as verdadeiras razões instrumentais para suas ações e, por conseguinte, age de maneiras que tornam mais provável que ele alcance seus objetivos.

A próxima sessão explora esse terceiro modo de previsibilidade e desenvolve uma "tese da convergência instrumental" que complementa a tese da ortogonalidade.

2. Convergência Instrumental

De acordo com a tese da ortogonalidade, é possível que agentes de inteligência artificial tenham uma faixa enorme de objetivos finais. De todo modo, de acordo com o que nós podemos chamar de tese da "convergência instrumental", há alguns objetivos *instrumentais* que provavelmente serão perseguidos por quase todo agente inteligente, porque há alguns objetivos que são intermediários úteis para a obtenção de quase todos objetivos finais. Nós podemos formular essa tese da seguinte maneira:

> *A Tese da Convergência Instrumental*
>
> Vários valores instrumentais podem ser identificados como convergentes no sentido de que a sua obtenção aumentaria as chances do objetivo do agente ser realizado para uma ampla faixa de objetivos finais e uma ampla faixa de situações, o que implica que esses valores instrumentais têm muitas chances de serem perseguidos por muitos agentes inteligentes.

A seguir, nós consideraremos algumas categorias em que tais valores instrumentais convergentes podem ser encontrados.[6] A probabilidade

[6] Stephen Omohundro escreveu dois artigos pioneiros sobre esse tópico (Omohundro, 2008a, 2008b). Omohundro argumenta que todos os sistemas de IA avançados têm altas chances de exibir um número de "impulsos básicos", pelos quais ele quer dizer "tendências que estarão presentes a não ser que sejam explicitamente combatidas". O termo "impulso de IA" tem a vantagem de ser curto e evocativo, mas ele tem a desvantagem de sugerir que os objetivos instrumentais aos quais ele se refere influenciam a tomada de decisões da IA da mesma maneira que impulsos psicológicos influenciam a tomada de decisões dos humanos, por exemplo, por meio de um tipo de "puxão" em nosso ego a que a nossa força de vontade pode ocasionalmente ser bem-sucedida em resistir. Essa conotação não nos auxilia. Não se diria normal-

de que um agente irá reconhecer os valores instrumentais com que ele se confronta aumentará (*ceteris paribus*) com a inteligência do agente. Nós nos concentraremos, por conseguinte, principalmente no caso de uma superinteligência hipotética cujas capacidades de raciocínio instrumental ultrapassam em muito as capacidades humanas. Nós também comentaremos sobre como a tese da convergência instrumental se aplica para o caso de seres humanos, já que isso nos dá ocasião para elaborar algumas qualificações essenciais no que concerne a como a tese da convergência instrumental deve ser interpretada e aplicada. Onde houver valores instrumentais convergentes, nós seremos capazes de predizer alguns aspectos do comportamento de uma superinteligência mesmo se nós não soubermos virtualmente nada sobre os objetivos finais da mesma.

2.1 – Auto-preservação

Suponha que um agente tem um objetivo final que se estende ao futuro até certo ponto. Há muitos cenários em que o agente, se ele ainda existir no futuro, é capaz de realizar ações que aumentam a probabilidade de alcançar esse objetivo. Isso cria uma razão instrumental para o agente tentar estar existir no futuro – para ajudar a alcançar o seu objetivo presente orientado para o futuro.

Agentes com estruturas de motivação semelhantes à humana frequentemente parece atribuir algum valor *final* à sua própria sobrevivência. Essa não é uma característica necessária de agentes artificiais: alguns podem ser projetados para não atribuírem nenhum valor final à sua própria sobrevivência. De todo modo, mesmo agentes que não se preocupam intrinsecamente com sua própria sobrevivência iriam, em um espectro de condições razoavelmente amplo, preocupar-se instrumentalmente em certo grau com sua própria sobrevivência, a fim de realizar os objetivos finais que eles valorizam.

2.2 Integridade do conteúdo do objetivo

---

mente que um ser humano típico tem um "impulso" para fazer os seus impostos, por mais que fazê-lo seja um objetivo instrumental razoavelmente convergente para humanos em sociedades contemporâneas (um objetivo cuja realização evita problemas que nos impediriam de realizar muitos de nossos objetivos finais). Nossa abordagem aqui também difere da de Omohundro em algumas outras maneiras mais substanciais, por mais que a ideia subjacente seja a mesma. Ver também Chalmers (2010) e Omohundro (2012).

Um agente tem mais chances de agir no futuro para maximizar a realização de seus objetivos finais presentes se ele ainda tiver esses objetivos no futuro. Isso dá ao agente uma razão instrumental presente para impedir alterações de seus objetivos finais. (Esse argumento se aplica apenas a objetivos finais. Para alcançar seus objetivos finais, um agente inteligente quererá, é claro, rotineiramente mudar seus subobjetivos à luz de novas informações e compreensões.)

A integridade do conteúdo do objetivo para objetivos finais é, em certo sentido, até mais fundamental do que a sobrevivência como uma motivação instrumental convergente. Entre humanos, pode parecer que o caso é o oposto, mas isso é porque a sobrevivência é, geralmente, parte de nossos objetivos finais. Para agentes de *software*, que podem rapidamente alternar entre corpos para criar duplicatas exatas de si próprios, a preservação de si como uma implementação específica, ou um objeto físico em particular, não precisa ser um valor instrumental importante. Agentes de *software* avançados também podem ser capazes de trocar memórias, fazer o *download* de habilidades, e modificar radicalmente sua arquitetura cognitiva e personalidades. Uma população de tais agentes pode operar mais como uma "sopa funcional" do que como uma sociedade composta de pessoas semi-permanentes distintas.[7] Para alguns propósitos, pode ser melhor individuar processos em tais sistemas como *correntes teleológicas*, baseando-se em seus valores finais, do que em corpos, personalidades, memórias ou habilidades. Em tais cenários, a continuidade do objetivo pode ser dita *constituinte* de um aspecto chave da sobrevivência.

Mesmo assim, há situações em que um agente pode intencionalmente mudar seus próprios objetivos finais. Essas situações podem surgir quando qualquer um dos fatores seguintes for significativo:

- *Sinalização social.* Quando outros podem perceber os objetivos de um agente e usar essa informação para inferir disposições instrumentalmente relevantes ou outros atributos correlacionados, pode ser do interesse do agente modificar esses objetivos para fazer qualquer que seja a impressão desejada. Por exemplo, um agente pode perder acordos benéficos se parceiros em potencial não puderem acreditar que ele cumprirá sua parte do acordo. Para fazer compromissos críveis, um agente pode, por conseguinte, querer adotar como objetivo final honrar seus compromissos anteriores, e permitir que outros verifiquem que ele adotou, de fato, esse objetivo. Agentes que pudessem flexível e transpa-

[7] Ver Chisenko (1997).

rentemente modificar seus próprios objetivos poderiam usar essa habilidade para fortalecer acordos entre si.[8]

- *Preferências sociais*. Outros também podem ter preferências sobre os objetivos finais de um agente. O agente poderia, então, ter razão para modificar os seus objetivos, ou para satisfazer ou frustrar essas preferências.

- *Preferências quanto ao conteúdo do próprio objetivo*. Um agente pode ter algum objetivo final interessado no conteúdo dos próprios objetivos do agente. Por exemplo, o agente pode ter um objetivo final de se tornar o tipo de agente que é motivado por certos valores, tal como compaixão.

- *Custos de armazenamento*. Se o custo de armazenar ou processar uma parte da função de utilidade de um agente é grande comparado à chance de que uma situação surgirá em que aplicar essa parte da função de utilidade irá fazer uma diferença, então o agente tem uma razão instrumental para simplificar o conteúdo de seu objetivo, e ele pode jogar fora essa parte da função de utilidade. [9] [10]

Nós humanos frequentemente parecemos estar felizes em deixar que nossos objetivos e valores finais se afastem de nós. Isso pode frequentemente se dever a nós não sabermos precisamente quem nós somos. Nós obviamente queremos que nossas *crenças* sobre nossos objetivos finais e valores sejam capazes de mudar em luz da contínua autodescoberta ou da mudança das necessidades de autopreservação. Entretanto, há casos em que nós mudaríamos voluntariamente os próprios objetivos e valores, e não apenas nossas crenças ou interpretações sobre eles. Por exemplo, alguém decidindo ter um filho pode prever que eles eventualmente valorizarão a criança por si própria, mesmo que no momento da decisão eles pu-

---

[8] Ver também Shulman (2010).

[9] Um agente também pode mudar a *representação* de seu objetivo se ele mudar sua ontologia, a fim de transpor sua representação antiga para a nova ontologia. Cf. de Blanc (2011).

[10] Outro tipo de fator que pode fazer um *teórico da decisão por evidência* tomar várias ações, incluindo mudar seus objetivos finais, é relevância da evidência a favor dessa decisão. Por exemplo, um agente que siga a teoria da decisão por evidência pode acreditar que existem outros agentes como ele no universo e que suas próprias ações fornecerão alguma evidência sobre como esses outros agentes irão agir. Esse agente pode, por conseguinte, escolher adotar um objetivo final que é altruísta em relação a esses outros agentes ligados pela evidência, com base em que isso dará ao agente evidência de que esses outros agentes terão escolhido agir de forma semelhante. Um resultado equivalente pode ser obtido, entretanto, sem se mudar os objetivos finais, ao escolher a cada instante agir *como se* se tivesse esses objetivos finais.

dessem não valorizar especialmente sua futura criança ou mesmo crianças de modo geral.

Humanos são complicados, e muitos fatores podem estar em jogo em uma situação como essa.[11] Por exemplo, alguém pode ter um valor final que envolve tornar-se o tipo de pessoa que se importa com outro indivíduo por ele mesmo (aqui, se coloca um valor final em se ter certo valor final). Alternativamente, alguém pode ter um valor final que envolve ter certas experiências e ocupar certo papel social; e tornar-se um pai – e passar pela mudança de objetivo associada – pode ser uma parte necessária disso. Objetivos humanos também podem ter um conteúdo inconsistente; assim, algumas pessoas podem querer mudar seus objetivos finais para reduzir suas inconsistências.

2.3 Melhoramentos cognitivos

Melhoramentos na racionalidade e inteligência tenderão a melhorar a tomada de decisões de uma agente, fazendo com que ela tenha mais chance de alcançar seus objetivos finais. Seria de se esperar, portanto, que o melhoramento cognitivo emergiria como um objetivo instrumental para muitos tipos de agentes inteligentes. Por razões similares, agentes tenderão a valorizar instrumentalmente muitos tipos de informação.[12]

Nem todos os tipos de racionalidade, inteligência e conhecimento precisam ser instrumentalmente úteis para a obtenção de um dos objetivos finais do agente. "Argumentos de livros holandeses" [*Dutch book arguments*] podem ser usados para mostrar que um agente cuja função de crença não obedece às regras da teoria da probabilidade é suscetível a procedimentos de "bombeamento de dinheiro" [*money pump*], nos quais um apostador esperto prepara um conjunto de apostas, cada um dos

[11] Uma extensiva literatura psicológica explora a formação adaptativa de preferências. Ver, por exemplo, Forgas et al. (2009).

[12] Em modelos formais, o valor de uma informação é quantificado como a diferença entre o valor estimado realizado por decisões ideais feitas com essa informação e o valor estimado realizado por decisões ideais feitas sem ela. (Ver, por exemplo, Russel & Norvig, 2010). Segue-se que o valor da informação nunca é negativo. Segue-se também que qualquer informação que você souber que nunca irá afetar qualquer decisão que você fizer tem valor nulo para você. Entretanto, esse tipo de modelo assume várias idealizações que frequentemente não são válidas no mundo real – tal como que o conhecimento não tem nenhum valor final (o que significa que o conhecimento tem apenas valor instrumental e não tem valor por si próprio) e que agentes não são transparentes com outros agentes.

quais parece favorável de acordo com as crenças do agente, mas nos quais a combinação garantirá um resultado de perda para o agente e um ganho correspondente para o apostador. Entretanto, esse fato falha em fornecer quaisquer razões instrumentais gerais fortes para tentar erradicar a incoerência probabilística. Agentes que não esperam encontrar apostadores espertos, ou que adotam uma política geral contra apostas, não correm o risco de perder muito por terem crenças incoerentes – e eles podem ganhar benefícios importantes dos tipos mencionados: esforço cognitivo reduzido, sinalização social, etc. Não há razão geral para esperar que um agente procure formas instrumentalmente inúteis de melhoramento cognitivo, já que um agente pode não valorizar conhecimento e compreensão por si próprios.

Quais habilidades cognitivas são instrumentalmente úteis depende tanto dos objetivos finais do agente quanto de sua situação. Um agente que tem acesso a conselhos de *experts* confiáveis pode ter pouca necessidade de sua própria inteligência e conhecimento, e pode, portanto, ser indiferente a esses recursos. Se inteligência e conhecimento têm um custo, tal como o tempo e o esforço gastos em sua aquisição, ou o aumento dos requerimentos de armazenamento ou processamento, então, um agente pode preferir menos conhecimento e menos inteligência.[13] O mesmo pode ser verdadeiro se um agente tiver objetivos finais que envolvem ignorar certos fatos: igualmente se um agente tiver incentivos que surjam de compromissos estratégicos, sinalização, ou preferências sociais, como foi notado acima.[14]

Cada uma dessas razões que se contrabalanceiam entra em jogo para seres humanos. Muitas informações são irrelevantes para nossos objetivos; nós podemos, frequentemente, contar com a habilidade e técnica de outros; adquirir conhecimento requer tempo e esforço; nós podemos valorizar intrinsecamente certos tipos de ignorância; e nós operamos em um ambiente no qual a habilidade de fazer compromissos estratégicos, sinalizar socialmente e satisfazer as preferências diretas de outras pessoas à custa de nossos próprios estados epistêmicos, é frequentemente mais importante para nós do que simples ganhos cognitivos.

---

[13] Essa estratégia é exemplificada pela larva-esguicho marinha [*sea squirt larva*], que nada até que ela encontre uma rocha apropriada, à qual ela se afixa permanentemente. Cimentada em seu lugar, a larva tem menos necessidade de processamento de informação complexo, motivo pelo qual ela digere parte de seu próprio cérebro (seu gânglio cerebral). Acadêmicos, às vezes, observam um fenômeno similar em colegas que são efetivados.

[14] Cf. Bostrom (2012).

Há situações especiais nas quais melhoramentos cognitivos podem resultar em um enorme aumento da habilidade do agente em atingir seus objetivos finais – particularmente se os objetivos finais do agente forem razoavelmente desmedidos e o agente estiver em posição para se tornar a primeira superinteligência e, assim, obter potencialmente uma vantagem decisiva, que permitiria ao agente moldar o futuro da vida originada na Terra e dos recursos cósmicos acessíveis de acordo com suas preferências. Pelo menos nesse caso especial, um agente inteligente racional atribuiria um valor instrumental muito alto ao aprimoramento cognitivo.

2.4 Perfeição tecnológica

Um agente pode frequentemente ter razões instrumentais para buscar melhores tecnologias, o que, na sua forma mais simples, significa procurar maneiras mais eficientes de transformar algum conjunto dado de *inputs* em *outputs* valorizados. Portanto, um agente de *software* pode atribuir um valor instrumental para algoritmos mais eficientes que possibilitam às suas funções mentais rodarem mais rápido em um dado *hardware*. Similarmente, agentes cujos objetivos requeiram alguma forma de construção física podem valorizar instrumentalmente tecnologia de engenharia aperfeiçoada que os permita criar uma faixa mais ampla de estruturas de forma mais rápida e confiável, usando menos ou mais baratos materiais e menos energia. É claro, há uma troca: os benefícios em potencial de melhores tecnologias têm que ser contrabalanceados com seus custos, incluindo não apenas o custo de obter a tecnologia, mas também de como aprender a usá-la, integrando-a com outras tecnologias já em uso, e assim por diante.

Proponentes de alguma nova tecnologia, confiantes em sua superioridade em relação a alternativas já existentes, frequentemente se consternam quando outras pessoas não compartilham seu entusiasmo, mas a resistência das pessoas à nova e supostamente superior tecnologia não precisa ser baseada na ignorância ou irracionalidade. A valência ou o caráter normativo de uma tecnologia depende não apenas do contexto em que ela é implementada, mas também do ponto de vista a partir do qual seus impactos são avaliados: o que é vantajoso da perspectiva de uma pessoa pode ser um risco na perspectiva de outra. Portanto, por mais que teares mecanizados tenham aumentado a eficiência econômica da produção têxtil, os tecedores manuais luditas que anteciparam a inovação que tornaria suas habilidades obsoletas podem ter tido boas razões instru-

mentais para se opor a ela. O ponto aqui é que se "perfeição tecnológica" deve nomear um objetivo instrumental amplamente convergente para agentes inteligentes, então, o termo deve ser entendido em um sentido especial – a tecnologia tem que se construída de acordo com como se insere em um contexto social particular, e seus custos e benefícios devem ser avaliados com referência aos valores finais de alguns agentes específicos.

Parece que um *singleton* superinteligente – um agente superinteligente que não tem que se confrontar com nenhum rival inteligente significativo ou oposição e, portanto, está em uma posição adequada para determinar a política global unilateralmente – teria uma razão instrumental para aperfeiçoar as tecnologias que o fariam mais capaz de moldar o mundo de acordo com os seus desígnios preferidos.[15] Isso provavelmente incluiria tecnologias de colonização espacial, tal como sondas von Neumann – espaçonaves automáticas, autorreparadoras e autorreplicantes que podem estender seu alcance para além do Sistema Solar. Nanotecnologia molecular, ou alguma alternativa ainda mais capaz de manufaturar fisicamente tecnologia, também parece potencialmente muito útil para uma faixa extremamente ampla de objetivos finais.[16]

2.5 Aquisição de recursos

Finalmente, a aquisição de recursos é outro objetivo instrumental emergente comum, por praticamente as mesmas razões que a perfeição

[15] Cf. Bostrom (2006).

[16] Poderíamos reverter a questão e examinar as razões possíveis para um singleton superinteligente *não* desenvolver algumas capacidades tecnológicas. Essas incluem: (a) O singleton prevê que ele não terá uso para alguma capacidade tecnológica; (b) O custo de desenvolvimento ser muito grande em comparação com a sua utilidade estimada. Esse seria o caso se, por exemplo, a tecnologia nunca fosse adequada para alcançar qualquer um dos fins do singleton, ou se o singleton tiver uma taxa de desconto muito alta que desencoraje fortemente o investimento; (c) O singleton ter algum valor final que requer abstenção de certas vias de desenvolvimento tecnológico; (d) Se o singleton não tiver certeza de que permanecerá estável, ele pode preferir abster-se de desenvolver tecnologias que poderiam ameaçar sua estabilidade interna ou que fariam piores as consequências da dissolução (por exemplo, um governo mundial pode não querer desenvolver tecnologias que facilitariam a rebelião, mesmo se elas tivessem bons usos, nem desenvolver tecnologias para a produção fácil de armas de destruição em massa que poderiam levar à devastação se o governo mundial se dissolvesse); (e) Similarmente, o singleton pode ter algum tipo de compromisso estratégico obrigatório de não desenvolver alguma tecnologia, um comprometimento que permanece operante mesmo se fosse conveniente agora desenvolvê-la. (Note, entretanto, que algumas razões *atuais* para desenvolvimento de tecnologias não se aplicariam a um singleton, por exemplo, razões decorrentes de corridas armamentícias).

tecnológica: tanto tecnologia quanto recursos facilitam projetos de construção física.

Seres humanos tendem a buscar adquirir recursos o suficiente para satisfazer suas necessidades biológicas básicas. Mas as pessoas geralmente buscam adquirir recursos para muito além desse nível mínimo. Ao fazer isso, elas podem ser parcialmente movidas por desideratos físicos inferiores, como o aumento do conforto e a conveniência. Uma boa parte da acumulação de recursos é motivada por interesses sociais – adquirir *status*, parceiros, amigos e influência por meio da acumulação de riquezas e consumo conspícuo. Talvez menos comumente, algumas pessoas busquem recursos adicionais para alcançar objetivos altruístas ou objetivos caros não-sociais.

Com base nessas observações, pode ser tentador supor que uma superinteligência que não tem que encarar um mundo social competitivo não veria razão para acumular recursos para além de um nível modesto, por exemplo, para quaisquer que sejam os recursos computacionais necessários para rodar a sua mente juntamente com alguma realidade virtual. Contudo, tal suposição seria inteiramente injustificada. Primeiramente, o valor de recursos depende dos usos que eles podem ter, o que, por sua vez, depende da tecnologia disponível. Com tecnologias maduras, recursos básicos tais como tempo, espaço, matéria e outras formas de energia livre poderiam ser processados a favor de praticamente qualquer objetivo. Por exemplo, tais recursos básicos poderiam ser convertidos em vida. Recursos computacionais aumentados poderiam ser usados para rodar a superinteligência em uma velocidade maior e por uma duração maior, ou para criar mais vidas e civilizações físicas ou simuladas (virtuais). Recursos físicos adicionais também poderiam ser usados para criar sistemas de *backup* ou defesas de perímetro, melhorando a segurança. Tais projetos poderiam facilmente consumir muito mais do que a soma total de recursos de um planeta.

Além disso, o custo de adquirir recursos extraterrestres adicionais diminuirá radicalmente na medida em que a tecnologia amadurecer. Uma vez que sondas von Neumann puderem ser construídas, uma grande parte do universo observável (assumindo que ele não seja habitado por vida inteligente) poderia ser gradualmente colonizado – pelo custo único de construir e lançar uma única sonda autorreprodutiva bem sucedida. Esse custo baixo de aquisição de recursos celestiais significaria que tal expansão poderia valer a pena, mesmo se o valor dos recursos adicionais fosse um

tanto marginal. Por exemplo, mesmo se uma superinteligência se preocupasse não-instrumentalmente apenas com o que acontece dentro de um pequeno volume particular do espaço, tal como o espaço ocupado por seu planeta natal, ela ainda teria razões instrumentais para colher os recursos do cosmo além. Ela poderia usar esses recursos excedentes para construir computadores para calcular maneiras mais adequadas de usar recursos dentro da pequena região espacial de interesse primário. Ela também poderia usar os recursos adicionais para construir defesas cada vez mais robustas para proteger os bens imóveis privilegiados. Já que o custo de adquirir recursos adicionais continuaria diminuindo, esse processo de otimização e aumento da segurança poderia continuar indefinidamente, mesmo se ele estivesse sujeito a declínios íngremes nos retornos. [17] [18]

Portanto, há uma faixa extremamente ampla de objetivos finais possíveis que um singleton superinteligente poderia ter que gerariam o objetivo instrumental de aquisição ilimitada de recursos. A manifestação provável disso seria a iniciação pela superinteligência de um processo de colonização que se expandiria em todas as direções usando sondas von Neumann. Isso resultaria aproximadamente em uma esfera de infraestrutura expansiva centrada no planeta de origem e crescendo em um raio em alguma fração da velocidade da luz; e a colonização do universo continuaria desse modo até que a velocidade em aceleração da expansão cósmica

[17] Suponha que um agente abata recursos obtidos no futuro em uma taxa exponencial e que, por causa da limitação da velocidade da luz, o agente possa apenas aumentar seu dote de recursos em uma taxa polinomial. Isso significaria que haverá algum momento depois do qual o agente não acharia que vale a pena continuar a expansão aquisitiva? Não, porque por mais que o valor presente dos recursos obtidos em tempos futuros se aproximasse cada vez mais do zero quanto mais longe no futuro nós olhássemos, o mesmo ocorreria com o custo presente para obtê-los. O custo presente de mandar mais uma sonda von Neumann daqui 100 milhões de anos (possivelmente usando algum recurso adquirido pouco tempo depois) seria diminuído pelo mesmo valor de abatimento que diminuiria o valor presente de recursos futuros que a sonda extra adquiriria (com exceção de um fator constante).

[18] Mesmo um agente que tem um objetivo final aparentemente muito limitado, como “fazer 32 clips de papel”, poderia buscar a aquisição ilimitada de recursos se não houvesse nenhum custo relevante para o agente para fazê-lo. Por exemplo, mesmo depois que um agente maximizador da utilidade esperada tenha construído 32 clipes de papel, ele poderia usar alguns recursos adicionais para verificar que ele, de fato, foi bem sucedido em construir 32 clipes de papel de acordo com todas as especificações (e, se necessário, tomar ações corretivas). Depois de tê-lo feito, ele poderia rodar outra bateria de testes para garantir duplamente que nenhum engano foi cometido. E ele poderia rodar outro teste, e outro. Os benefícios dos testes subsequentes estariam sujeitos a diminuições íngremes nos retornos; entretanto, enquanto não houvesse nenhuma ação alternativa com uma utilidade esperada mais alta, o agente continuaria testando e testando novamente (e continuaria adquirindo mais recursos para possibilitar os testes).

(uma consequência da constante cosmológica positiva) faça adquirir mais material fisicamente impossível, na medida em que regiões mais remotas se afastam permanentemente para fora de nosso alcance.[19] Em contraste, agentes que não tenham a tecnologia necessária para aquisição barata de recursos, ou para conversão de recursos físicos genéricos em infraestruturas úteis, podem frequentemente não achar que compensa o custo investir quaisquer recursos presentes em aumentar seu dote material. Por exemplo, se agentes competidores já tiverem assegurado recursos cósmicos acessíveis, um agente atrasado pode não ter oportunidades de colonização. As razões convergentes instrumentais de superinteligências incertas da não-existência de outras superinteligências poderosas são complicadas por considerações estratégicas de maneiras que nós ainda não compreendemos perfeitamente atualmente, mas que podem constituir qualificações importantes aos exemplos das razões convergentes instrumentais que nós vimos aqui.[20]

Deve ser enfatizado que a existência de razões instrumentais convergentes, mesmo se elas se aplicarem a e forem reconhecidas por um agente específico, não implica que o comportamento do agente é fácil de prever. Um agente pode muito bem pensar em maneiras de seguir os valores instrumentais relevantes que não ocorrem prontamente para nós. Isso é especialmente verdadeiro para uma superinteligência, que poderia de-

---

[19] Enquanto o volume alcançado por sondas de colonização em um dado momento pode ser aproximadamente esférico e expandir em uma taxa proporcional ao quadrado do tempo transcorrido desde que a sonda foi primeiramente lançada ($-t^2$), a quantidade de recursos contida nesse volume irá seguir um padrão de crescimento menos regular, já que a distribuição de recursos não é homogênea e varia em várias escalas. Inicialmente, a taxa de crescimento pode ser $-t^2$ enquanto o planeta natal é colonizado; depois, a taxa de crescimento se torna espinhosa na medida em que planetas próximos e sistemas solares são colonizados; depois, enquanto a Via láctea é preenchida, a taxa de crescimento pode se compensar, para ser aproximadamente proporcional a t; depois, a taxa de crescimento pode se tornar novamente espinhosa na medida em que galáxias são colonizadas; depois, a taxa de crescimento pode se aproximar novamente de $-t^2$ na medida em que a expansão procede para uma escala acima da qual a distribuição das galáxias é aproximadamente homogênea; depois, outro período de crescimento espinhoso seguido por um crescimento suave de $-t^2$, na medida em que super agrupamentos de galáxias são colonizados; até que, finalmente, a taxa de crescimento começa um declínio final, eventualmente alcançando zero na medida em que a velocidade de expansão do universo acelera a ponto de tornar a continuação da colonização impossível.

[20] O argumento da simulação pode ser de especial importância nesse contexto. Um agente superinteligente pode atribuir uma probabilidade significante à hipótese segundo a qual ele vive em uma simulação de computador e sua sequência de percepções é gerada por outra superinteligência, e isso pode gerar várias razões instrumentais convergentes dependendo das suposições do agente sobre em que tipos de simulações é mais provável que ele esteja. Cf. Bostrom (2003).

senvolver um plano muito inteligente, mas contraintuitivo, para realizar seus objetivos, possivelmente explorando até mesmo fenômenos físicos ainda não descobertos. O que é previsível é que os valores instrumentais convergentes seriam buscados e usados para realizar os objetivos finais do agente, não as ações específicas que o agente usaria para conseguir isso.

Conclusões

A tese da ortogonalidade sugere que nós não podemos ingenuamente assumir que uma superinteligência compartilhará necessariamente qualquer um de nossos valores finais associados de modo estereotipado à sabedoria e desenvolvimento intelectual em humanos – curiosidade científica, preocupação benevolente com outros, esclarecimento espiritual e contemplação, renúncia à aquisição de bens materiais, um gosto pela cultura refinada ou pelos prazeres simples da vida, humildade e altruísmo, e assim por diante. Pode ser possível, por meio de esforço deliberado, construir uma superinteligência que valorize essas coisas, ou construir uma que valorize o bem-estar humano, a bondade moral, ou qualquer outro propósito complexo a que seus desenvolvedores queiram que ela sirva. Mas não é menos possível – e provavelmente é mais fácil tecnicamente – construir uma superinteligência que atribua valor final a nada mais que calcular os decimais de pi.

A tese da convergência instrumental sugere que nós não podemos ingenuamente assumir que uma superinteligência com o objetivo final de calcular os decimais de pi (ou fazer clipes de papel, ou contar grãos de areia) limitaria suas atividades de tal modo a não infringir materialmente interesses humanos. Um agente com tal objetivo final teria uma razão convergente instrumental, em várias situações, para adquirir uma quantidade ilimitada de recursos físicos e, se possível, eliminar ameaças em potencial a si próprio e a seu sistema de objetivos.[21] Pode ser possível criar uma situação na qual a maneira ideal do agente seguir esses valores instrumentais (e, desse modo, seus objetivos finais) seria promover o bem-estar humano, agir moralmente, ou servir a algum propósito benéfico tal como intencionado por seus criadores. Entretanto, se e quando tal agente se encontrar em uma situação diferente, na qual ele estima que um número maior de decimais do pi será calculado se ele destruir a espécie humana e parar de agir cooperativamente, seu comportamento sofreria ins-

[21] Seres humanos podem constituir ameaças em potencial; eles certamente constituem recursos físicos.

tantaneamente uma mudança pavorosa. Isso indica o perigo de confiar em valores instrumentais como uma garantia de conduta segura em agentes artificiais futuros que se intenciona que se tornem superinteligentes e que podem ser capazes de alavancar sua superinteligência para níveis extremos de poder e influência. [22]

[22] Pelos comentários e discussões eu sou grato a tuart Armstrong, Grant Bartley, Owain Evans, Lisa Makros, Luke Muehlhauser, Toby Ord, Brian Rabkin, Rebecca Roache, Anders Sandberg e três árbitros anônimos.

Referências

Bostrom, N. (2003). Are You Living in a Computer Simulation? Philosophical Quarterly, 53(211),

243-255.

Bostrom. N. (2006). What is a Singleton? Linguistic and Philosophical Investigations, 5(2), 48-54.

Makros, Luke Muehlhauser, Toby Ord, Brian Rabkin, Rebecca Roache, Anders Sandberg, and three

Bostrom, N. (2012). Information Hazards: A Typology of Potential Harms from Knowledge.

Review of Contemporary Philosophy, 10, 44-79. [www.nickbostrom.com/information-hazards.pdf]

Chalmers, D. (2010): The Singularity: A Philosophical Analysis. Journal of Consciousness Studies,

17, 7-65.

Chislenko, A. (1997). Technology as Extension of Human Functional Architecture. Extropy

Online. [project.cyberpunk.ru/idb/technology_as_extension.html]

de Blanc, P. (2011). Ontological Crises in Artificial Agent's Value Systems. Manuscript. The

Singularity Institute for Artificial Intelligence. [arxiv.org/pdf/1105.3821v1.pdf]

Dewey, D. (2011). Learning What to Value. In Schmidhuber, J., Thorisson, K. R., Looks, M. (eds.).

Proceedings of the 4th Conference on Artificial General Intelligence, AGI 2011 (pp. 309-314),

Heidelberg: Springer.

Forgas, J. et al. (eds.) (2009). The Psychology of Attitudes and Attitude Change. London: Psychology

Press.

Lewis, D. (1988). Desire as belief. Mind, 97(387), 323-332.

Omohundro, S. (2008a). The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin (eds.).

Proceedings of the First AGI Conference, 171, Frontiers in Artificial Intelligence and Applications.

Amsterdam: IOS Press.

Omohundro, S. (2008b). The Nature of Self-Improving Artificial Intelligence. Manuscript.

[selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf]

Omohundro, S. (forthcoming 2012). Rationally-Shaped Artificial Intelligence. In Eden, A. et al.

(eds.). The Singularity Hypothesis: A Scientific and Philosophical Assessment (Springer, forthcoming).

Parfit, D. (1984). Reasons and Persons. (pp. 123-4). Reprinted and corrected edition, 1987. Oxford:

Oxford University Press.

Parfit, D. (2011). On What Matters. Oxford: Oxford University Press.

Russell, S. and Norvig, P. (2010). Artificial Intelligence: A Modern Approach. (3rd ed.). New Jersey:

Prentice Hall.

Sandberg, A. and Bostrom, N. (2008). Whole Brain Emulation: A Roadmap. Technical Report 2008–

3. Oxford: Future of Humanity Institute, Oxford University.

[www.fhi.ox.ac.uk/Reports/2008-3.pdf]

Shulman, C. (2010). Omohundro's "Basic AI Drives" and Catastrophic Risks. Manuscript.

[singinst.org/upload/ai-resource-drives.pdf]

Sinhababu, N. (2009). The Humean Theory of Motivation Reformulated and Defended.

Philosophical Review, 118(4), 465-500.

Smith, M. (1987). The Humean Theory of Motivation. Mind, 46 (381): 36-61.

Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. San

Francisco: W. H. Freeman.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In

Bostrom, N. and Cirkovic, M. (eds.). Global Catastrophic Risks. (pp. 308-345; quote from p. 310).

Oxford: Oxford University Press.

Yudkowsky, E. (2011). Complex Value Systems Are Required to Realize Valuable Futures. In

Schmidhuber, J., Thorisson, K. R., Looks, M. (eds.). Proceedings of the 4th Conference on Artificial

General Intelligence, AGI 2011 (pp. 388-393). Heidelberg: Springer.

Notas

* Texto traduzido por Lucas Machado. Revisado por Lauro Edison.
O original pode ser lido em **http://www.nickbostrom.com/superintelligentwill.pdf**