Studies in Social Sciences Review
ISSN: 2764-085X

Superinteligência artificial: utopia ou distopia tecnológica?

Artificial superintelligence: utopia or technological dystopia?

DOI: 10.54018/sssrv3n1-003

Recebimento dos originais: 30/11/2021
Aceitação para publicação: 23/12/2021

Pablo de Araújo Batista
Bacharel em Filosofia pela Universidade São Judas Tadeu
Universidade São Judas Tadeu
E-mail: pablobits@gmail.com

RESUMO

O artigo analisa a possibilidade do surgimento de uma superinteligência artificial em um evento denominado *Singularidade*. A evolução da capacidade computacional da Inteligência Artificial segue um processo dependente da trajetória. Processos dependentes da trajetória possuem alto grau de previsibilidade, sendo que em tais processos todas as probabilidades de ocorrência de um evento são equiprováveis. Em tais sistemas o passado assume exponencialmente mais importância. A inserção de uma superinteligência capaz de se auto compreender e se auto aprimorar num processo dependente da trajetória, alterará de forma drástica nossa capacidade de previsibilidade e será a causa de uma ruptura no sistema - *tipping point*. Esse ponto de inflexão altera drasticamente o rumo de todo o sistema, aumentando seu grau de incerteza e de imprevisibilidade. Com a concretização de tal cenário, seremos incapazes de prever com precisão se o surgimento dessa Inteligência terá como consequência uma utopia ou uma distopia tecnológica.

Palavras-chave: Superinteligência Artificial, Singularidade, Dependência da Trajetória, Tipping Point.

ABSTRACT

The article examines the possibility of the emergence of an artificial superintelligence during an event called Singularity. The evolution of the computing power of Artificial Intelligence follows a path-dependent process. Path-dependent processes have a high degree of predictability, and in such cases all the probabilities of occurrence of an event are equal. In such systems, the past exponentially assumes more importance. Inserting a superintelligence capable of self-understanding and self-improvement during a path-dependent process will drastically change our predicting capacity and will be the cause of a breakdown in the system – a tipping point. This inflection point dramatically alters the course of the entire system, increasing its degree of uncertainty and unpredictability. With the realization of such a scenario, we will be unable to accurately predict whether the emergence of this Intelligence will result in a technological utopia or dystopia.

Keywords: Artificial Superintelligence, Singularity, Path Dependence, Tipping Point.

> A evolução tem sido vista como um drama de um bilhão de anos que levou inexoravelmente à sua maior criação: a inteligência humana. Nas primeiras décadas do século XXI, a emergência de uma nova forma de inteligência na Terra que possa competir com a inteligência humana, e no fim das contas superá-la de modo significativo, será um desenvolvimento de maior importância do que a criação da inteligência que a criou, e terá profundas implicações em todos os aspectos do esforço humano, incluindo a natureza do trabalho, o aprendizado humano, o governo, a guerra, as artes e nosso conceito de nós mesmos.
>
> - Ray Kurzweil

1 INTRODUÇÃO

O evento denominado Singularidade mudará significativamente a condição humana, principalmente devido ao advento de uma superinteligência artificial. As previsões mais otimistas estimam que a Singularidade ocorrerá em meados de 2045 quando o poder computacional das máquinas será equivalente a todos os cérebros humanos conectados[1].

As consequências da Singularidade ainda são uma incógnita, pois poderão trazer inúmeros benefícios à humanidade (Utopia), como também poderão ser o início do fim dos organismos humanos (Distopia). Sistemas altamente complexos de inteligência artificial poderão dominar o planeta, automatizando todas astarefas que atualmente desejamos abandonar ou até mesmo tornando nossa existência totalmente obsoleta. Quando isso ocorrer, questões que hoje consideramos importantes como a consciência das máquinas, se tornarão irrelevantes. As questões mais importantes estarão diretamente relacionadas à ética aplicada a esses seres, bem como o impacto social, político e econômico resultante do advento dessas novas formas de inteligência[2].

[1] KURZWEIL, R. *The Singularity Is Near: When Humans Transcend Biology* (Penguin Books; Illustrated Edition, 2006).

[2] Observe que a pandemia da COVID-19, que tem isolado socialmente milhões de pessoas, tem sido um bom

O desenvolvimento de sistemas artificiais de aprendizagem evolui exponencialmente em um processo dependente da trajetória e resultará na criação de uma Inteligência Artificial Generalizada; uma inteligência capaz de aprender com seus próprios erros e criar sucessivos sistemas mais inteligentes. Essa forma de inteligência poderá ser merecedora do prefixo super. Nos processos dependentes da trajetória, o passado possui exponencialmente mais importância e, por isso, o que acontecerá quando ocorrer a Singularidade será extremamente influenciado pela trajetória escolhida nos próximos anos.

Sistemas dependentes da trajetória possuem alto grau de previsibilidade, sendo que, em tais sistemas, todas as probabilidades de ocorrência de um evento são equiprováveis. Argumento que a inserção de uma superinteligência em um processo dependente da trajetória alterará de forma drástica nossa capacidade de previsibilidade e será a causa de uma ruptura no sistema. Essa ruptura é denominada de *tipping point*, um evento único que muda drasticamente o rumo de um sistema, aumentando o grau de incerteza em nossas previsões sobre os benefícios ou malefícios decorrentes do surgimento de uma superinteligência artificial[3].

2 A SUPERINTELIGÊNCIA

No século passado criamos a Inteligência Artificial (IA) e com receio de antropomorfizar sistemas computacionais, usamos sinônimos como processamento de dados, lógica e eficiência para definir suas habilidades. Nossa tentativa frustrada: Evitar que adjetivos exclusivos de “nossa humanidade” fossem dados a uma máquina.

Contudo, se inteligência pode ser definida como racionalidade instrumental, talento para previsões, planejamento e raciocínio sobre meios para atingir

laboratório para o teste de robôs sociais. O *CIMON Crew Interactive Mobile Companion*, desenvolvido pela agência espacial alemã DLR, Airbus e IBM, utilizado na nave *Crew Dragon* que chegou à Estação Espacial Internacional (ISS) em 2020, tem interagido com os astronautas para minimizar os efeitos negativos do isolamento. Esses e outros sistemas têm sido testados na Terra em interações com idosos que, nesse momento de pandemia, carecem ainda mais de interações sociais e cuidados (https://www.ibm.com/blogs/think/2020/06/lessons-from-space-may-help-care-for-those-living-through-social-isolation-on-earth/ acesso em 08/05/2021)

[3] No texto utilizo os termos Superinteligência Artificial e Inteligência Artificial Generalizada com o mesmo significado.

objetivos finais, sua própria definição nos obriga a honrar alguns sistemas artificiais com esse adjetivo. O fundamental não é a definição de inteligência que utilizamos, mas sim se o agente analisado, seja um computador *Apple* ou um algoritmo artificial de busca (determinístico ou probabilístico), pode atingir seus objetivos finais em situações diversas.

Atualmente já admitimos a inteligência das máquinas, mas podemos admitir também que algumas delas pensam ou poderão pensar? Sem mesmo precisar entrar em longos debates filosóficos sobre a definição de pensamento, aceito como pressuposto ao pensamento artificial, a resposta de Alan Turingapresentada de forma ficcional no longa metragem o *Jogo da Imitação*[4]:

> É claro que máquinas não podem pensar como as pessoas. Uma máquina é diferente de uma pessoa, portanto, pensam de modo diferente. A questão interessante é, só porque alguma coisa pensa diferente de você, significa que ela não pensa? Nós concordamos que os humanos divergem uns dos outros. Você gosta de morangos eu odeio patinação no gelo, você chora em filmes tristes, eu sou alérgico a pólen. Como explicar gostos diferentes, preferências diferentes, senão dizendo que nossas mentes trabalham de modo diferentes, que pensamos de modo diferente? E se podemos dizer isso um do outro por que não podemos dizer o mesmo de mentes construídas de cobre, arame e aço?

Faz-se importante ressaltar que ao falarmos aqui de IA estamos falando sobre sistemas mais avançados do que os atuais, como por exemplo, os sistemas utilizados nos pilotos automáticos de aeronaves, veículos autônomos, GPS, diagnósticos médicos e nas previsões da bolsa de valores[5]. As máquinas atuais são sistemas extremamente simples que manejadas de forma correta, não proporcionam qualquer tipo de risco a nossa existência.

[4] *THE IMITATION Game* (O Jogo da Imitação). Direção: Morten Tyldum. *Warner Bros*. 2014.

[5] A área de aprendizagem das máquinas anda a passos largos e impactará definitivamente o mercado de trabalho. Alguns algoritmos desenvolvidos analisam provas escolares e fazem diagnósticos de retinopatia diabética. Em ambos os casos as avaliações das máquinas se igualam as avaliações humanas mas com um agravante: Um professor humano em 40 anos de carreira pode ler 10 mil redações e um oftalmologista pode examinar 50 mil olhos durante sua vida; uma máquina pode fazer ambas as tarefas em poucos minutos (https://www.ted.com/talks/anthony_goldbloom_the_jobs_we_ll_lose_to_machines_and_the_ones_we_won_t?language=pt-br

Estamos tratando de um tipo de IA capaz de se autocompreender, tomar decisões e fazer escolhas relevantes com o intuito de atingir seus objetivos[6]. Por isso, para diferenciar esses sistemas superinteligentes das atuais formas de IA, cunhou-se o termo Inteligência Artificial Geral (IAG). A IAG pode ser definida como uma entidade capaz de compreender sua própria estrutura, reformular a si mesma alterando seu código fonte, criando sucessivos sistemas ainda mais inteligentes[7].

Essa superinteligência possuirá cognição semelhante à cognição do *Homo sapiens*, podendo até mesmo ser idêntica em caso de um *upload* completo de uma mente humana[8]. A diferença fundamental consistirá no fato de que enquanto todos os representantes da espécie humana compartilham uma arquitetura cerebral comum (biológica), tendo por isso limitações espaciais e temporais impostas pelas leis da física, uma IAG possuirá um espaço de projeto muito maior do que o espaço da mente humana. A cognição da máquina poderá ser instanciada em diversos tipos de mídias, construídas em matérias das mais diversas composições e executadas nas mais diversas velocidades.

Com essa liberdade de instanciação novas formas de inteligência surgirão; inteligências comparadas à humana, inteligências que ultrapassarão a inteligência humana e inteligências sequer imaginadas pela inteligência humana. Quando isso ocorrer, atingiremos o que muitos pensadores denominam de Singularidade.

3 A SINGULARIDADE

O filósofo David J. Chalmers argumentou de forma convincente no texto *The Singularity: A Philosophical Analysis*, sobre a possibilidade da Singularidade.

[6] A IBM parece entender o caminho que IA está tomando. A empresa abandonou sua área de construção de computadores para entrar definitivamente na área da IA. Com o slogan "bem-vindo a era cognitiva", a IBM pretender ser pioneira no desenvolvimento de sistemas operacionais como o *Watson*, "projetado para entender, raciocinar e aprender; em certo modo para pensar". A era cognitiva, conforme proposta pela IBM, impactará fundamentalmente as áreas de negócios, alimentos, saúde, educação, varejo e automóveis (https://www.ibm.com/br-pt/watson).

[7] A generalidade é de fundamental importância no contexto social, pois quando uma criatura ou artefato opera apenas dentro de um domínio específico, ele pode ocasionar sérios ricos a sua própria existência e também à existência dos diretamente envolvidos em suas ações.

[8] A ideia básica do *upload* de uma mente para um computador consiste em digitalizar detalhadamente a estrutura de um cérebro e construir um modelo idêntico em forma de *software*. Se esse *software* for executado em um *hardware* adequado, ele se comportará basicamente da mesma forma que o cérebro original. Sobre esse tema, veja o texto de Anders Sandberg e Nick Bostrom em http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf.

Tal evento será o resultado da crescente evolução dos sistemas de IA culminando no advento de uma superinteligência consciente. Isso resultará em uma explosão de superinteligência no planeta. Chalmers ainda destaca a importância de se considerar a Singularidade como um tema relevante em nossas discussõesdevido as suas implicações práticas e filosóficas:

> Praticamente: Se houver uma singularidade, será um dos eventos mais importantes na história do planeta. Uma explosão de inteligência tem enormes benefícios potenciais: a cura para todas as doenças conhecidas, o fim da pobreza, extraordinários avanços científicos, e muito mais. Ele também tem grandes perigos potenciais: o fim da raça humana, uma corrida armamentista de máquinas de guerra, o poder de destruir o planeta. Portanto, se houver mesmo uma pequena chance de que haverá uma singularidade, nós faríamos bem em pensar sobre que forma ela pode tomar e se há alguma coisa que podemos fazer para influenciar os resultados em uma direção positiva.
>
> Filosoficamente: A singularidade levanta muitas questões filosóficas importantes. O argumento básico para uma explosão de inteligência é filosoficamente interessante em si mesmo, e nos obriga a pensar muito sobre a natureza da inteligência e sobre as capacidades mentais de máquinas artificiais. As potenciais consequências de uma explosão de inteligência nos obrigam a pensar muito sobre valores e moralidade e sobre a consciência e identidade pessoal. Com efeito, a singularidade traz algumas das mais difíceis questões tradicionais em filosofia e levanta também algumas novas questões filosóficas[9].

Nesse caso, a analogia com a origem do universo pode ser relevante. Segundo o modelo padrão o universo está em expansão, logo, se voltarmos suficientemente no tempo chegaremos a uma singularidade, um ponto que contêm toda a energia e a matéria do universo. O mais interessante nessa concepção - que também se aplica ao nosso entendimento da Singularidade Tecnológica; é que não podemos descrever a singularidade ou o que há após o horizonte de eventos, pois as leis da física não se aplicam a um ponto com

[9] Tradução livre do autor.

densidade infinita de matéria e energia.

Nesse contexto, o termo Singularidade é providencial em conformidade com as concepções sobre a origem do universo e os buracos negros, onde a singularidade é um ponto no espaço-tempo em que sua curvatura se torna infinita. Ao atingir essa suposta curvatura infinita, as máquinas alcançarão um nível de inteligência mais elevado do que a de seus criadores. Quando isso ocorrer é provável que a direção da seta de dominação tome uma direção pouco agradável para os humanos, tornando obsoleto nossos cérebros biológicos[10].

A evolução de sistemas artificiais e o aumento em sua complexidade segue uma linha exponencial baseada principalmente na Lei de Moore[11]. O crescimento exponencial ou geométrico é diferente do crescimento linear ou aritmético. Enquanto no crescimento linear o avanço segue na sequência numérica de 1, 2, 3, 4 e assim por diante, no crescimento exponencial isso ocorre na taxa de 1, 2, 4, 8, 16, 32 e assim progressivamente.

Nos baseando no avanço exponencial da inteligência das máquinas, podemos fazer previsões acertadas sobre a forma de desenvolvimento desses sistemas, porém, quando ocorrer a Singularidade essa previsibilidade poderá ser abalada. Isso porque a principal dificuldade com a Singularidade é que após o seu advento há um inevitável aumento no grau de incerteza, pois de fato, não temos conhecimento sobre o que ocorrerá após o seu surgimento. O que sabemos hoje sobre a Singularidade pode se revelar verdadeiramente insignificante, tornando o que não sabemos mais relevante do que aquilo que sabemos.

[10] O relacionamento que tínhamos com nossas criações sempre foi orientado pela dominação na direção homem/máquina. No entanto, no final do século XX e início do século XXI, nossas criações atingiram tamanho grau de complexidade que, em muitos casos, elas são capazes de tomar decisões por nós, ou até mesmo de impor sua "vontade". Atualmente a seta aponta para uma via de mão dupla, onde homem e máquina interagem tentando impor suas "vontades" por canais de comunicação relativamente limitados.

[11] Um dos fundadores da *Intel*, Gordon E. Moore, observou que a área da superfície de um transistor era reduzida em aproximadamente 30% a cada 12 meses. Em 1975 foi divulgado que após uma revisão de sua teoria inicial sobre a taxa de crescimento da capacidade dos circuitos integrados, Moore modificou sua observação para 18 meses, embora ele afirme que sua revisão tenha sido para 24 meses. Como resultado, a cada dois anos é possível colocar duas vezes mais transistores num circuito integrado. Ao duplicar o número de componentes em um chip, as distâncias que os elétrons devem percorrer diminuem, aumentando exponencialmente sua velocidade.

2 O PASSADO ALTERA AS PROBABILIDADES

> Fui atingido por uma crença que nunca mais me abandonou de que somos apenas uma grande máquina de olhar para trás, e que os humanos são ótimos em se auto enganarem.
>
> - Nassim N. Taleb

Os resultados de alguns processos são altamente dependentes da trajetória. O que isso significa? Significa que em tais processos as escolhas presentes são fundamentais pois condicionam o futuro, fazendo com que em uma visão retrospectiva, o passado assuma maior relevância. Embora não haja unanimidade sobre os mecanismos explicativos do *path dependence* (dependência da trajetória), o conceito tem sido utilizado com relativo sucesso nas análises históricas da predominância de algum tipo de tecnologia, das teorias econômicas e das ciências políticas e sociais.

No início de um determinado processo podem coexistir duas ou mais trajetórias possíveis, mas um fato contingente ou mesmo uma escolha deliberada determinará todo o curso posterior da história. Isso ocorre porque ao se determinar um rumo, os eventos posteriores gerarão feedback positivos que reforçará sua utilização no decorrer de toda a sequência temporal. Isso é denominado de “retornos crescentes”. O aumento na utilização e distribuição de determinada tecnologia eleva os benefícios de sua utilização, reduzindo o custo de sua produção, tendo como consequência um ciclo de retroalimentação.

Cada passo dado em uma determinada direção aumenta a probabilidade de que os próximos passos sejam dados nessa mesma direção, pois o custo de transição para uma tecnologia que foi preterida no processo é extremamente elevado. Vejamos alguns exemplos.

2.1 Guerra Das Correntes

O avanço da tecnologia é um processo dependente da trajetória, tornando as escolhas que fazemos atualmente de extrema importância. Isso pode ser observado em diversas competições entre tecnologias rivais, quando as escolhas durante o processo determinam o rumo da tecnologia predominante. A chamada “Guerra das Correntes” é um bom exemplo desse tipo de competição.

Nas últimas décadas do século XIX houve uma disputa em território norte americano para determinar qual seria a forma de distribuição de energia elétrica, em forma de corrente contínua (CC), difundida por Thomas Edson ou em forma de corrente alternada (CA), o sistema inventado por Nicola Tesla e difundida por George Westinghouse. O sistema desenvolvido por Tesla prevaleceu, pois uma de suas principais vantagens era que diferentemente da CC, a CA permitia a transmissão de eletricidade a longas distancias.

2.2 Teclado Qwerty

Quando se desenvolveram as primeiras máquinas de escrever a disposição das teclas produziam um travamento nas barras de tipo. Como as teclas emperradas não eram visíveis, o datilógrafo continuava martelando repetidamente a mesma letra no documento. A solução foi reconfigurar o teclado no formato QWERTY, que utilizamos até hoje em nossos computadores, notebooks e celulares, mesmo com a impossibilidade de travamento das teclas. Embora posteriormente tenha se demonstrado que a curva de aprendizado do teclado QWERTY seja extremamente longa, esse layout predominou.

Por que continuamos utilizando o teclado nessa configuração? Por que o custo para reaprender a utilizar um teclado com outra configuração é extremamente elevado. Mesmo que outros formatos de teclados sejam mais eficientes, diminuindo, portanto, a curva de aprendizado, as decisões tomadas no passado cimentaram de forma definitiva a utilização do modelo QWERTY[12].

2.3 NOGALES

Podemos também confirmar a atuação da dependência da trajetória na formação de instituições e de sociedades. Um exemplo interessante é a cidade de Nogales, dividida ao meio por uma cerca. A Nogales ao norte está em Arizonanos Estados Unidos e por isso se beneficia com educação de qualidade, assistência médica, eletricidade, água potável, sistema de telefonia, malha

[12] Nas décadas de 1920 e 1930 os designers August Dvorak e William Dealey, apresentaram como alternativa ao layout QWERTY o Teclado Simplificado Dvorak. O teclado era comprovadamente muito mais eficiente, com baixa curva de aprendizado, além de ser recomendado por suas vantagens ergonômicas. Porém, o teclado Dvorak nunca foi adotado pois ocorreram resistência de fabricantes e usuários, que precisariam investir em um novo produto e em reaprender a datilografia respectivamente.

rodoviária, relativa segurança como resultado da aplicação da lei e incentivo ao empreendedorismo. A Nogales do Sul está em Sonora no México e por isso sofre com as dificuldades de um país que não prioriza a educação, saúde e segurança de seus cidadãos, além dos altos índices de mortalidade infantil e da baixa expectativa de vida dos moradores da região. Não há incentivos para empreender, pelo contrário, iniciar um empreendimento é extremamente arriscado, pois provavelmente o aspirante a empresário sofrerá extorsão dos políticos locais corruptos e inescrupulosos.

Ora, as duas cidades estão geograficamente no mesmo lugar, compartilhando o mesmo clima bem como os mesmos tipos de doenças que se disseminam na região. Os habitantes de ambas as cidades possuem origem em comum e compartilham o mesmo gosto para comida e música. Podemos afirmar que dividem uma mesma cultura. Mas, por que mesmo com tanta semelhança, existe tamanha diferença nas sociedades dessas duas cidades separadas apenas por uma cerca?

A diferença fundamental está nas instituições estabelecidas em ambas as sociedades. Os habitantes da Nogales do Norte gozam das instituições econômicas e políticas estabelecidas nos Estados Unidos, que permitem em certa medida, que todos os cidadãos escolham suas ocupações, estudem em universidades, invistam em tecnologia e participem ativamente do processo democrático. Os habitantes da Nogales do Sul não possuem esses benefícios, pois vivem à margem da sociedade mexicana, governada por instituições corruptas e extrativista.

O contraste atual tão visível das cidades teve origem na formação das duas sociedades. O México teve sua formação a partir do método espanhol de colonização, ou seja, escravizando os nativos e extraindo toda a riqueza local. A história da constituição e independência do México é repleta de reviravoltas, subidas e descidas de presidentes ao poder, sendo que a grande maioria de seus líderes disseminaram a mesma forma de política colonizadora da Espanha, com a priorização do enriquecimento da elite dominante em detrimento dos súditos ou cidadãos.

Os Estados Unidos não puderam ser formados dessa forma. Não que a Coroa inglesa não desejasse colonizar à moda espanhola, eles tentaram, mas não

encontraram tantas pedras preciosas a serem exploradas e principalmente porque os nativos não aceitaram ser escravizados. Além disso, os colonos ingleses muito cedo exigiram seus direitos já que tiveram que trabalhar duro para colonizar a região. Logo, foi necessário que a Coroa oferecesse aos colonos bons incentivos ao trabalho e ao desenvolvimento, instaurando-se lentamente o sistema democrático nas entranhas da América do Norte.

A constituição dos Estados Unidos e do México demonstram o poder da dependência da trajetória. As escolhas iniciais de suas instituições formadoras levaram ambos os países a rumos totalmente diferentes. Como resultado, os eventos iniciais e a trajetória seguinte que reforçou as escolhas do passado, tiveram efeitos permanentes em ambas as sociedades, efeitos que podem ser sentidos até os dias atuais[13].

2.4 Modelando A Dependência Da Trajetória

Podemos modelar a dependência da trajetória. Utilizando um modelo extremamente simples de urna podemos demonstrar como as alterações que vão ocorrendo em cada tempo, devido a escolhas intencionais ou mesmo devido a aleatoriedade, modificam drasticamente as probabilidades dos resultados. Vamos utilizar nesse modelo um Processo de Polya. No Processo de Polya é possível confirmar a dependência da trajetória ao analisarmos as alterações de probabilidade, ou seja, as alterações que a história pode sofrer a partir de pequenas escolhas.

Suponha uma urna no tempo 1 (t1) com uma bola preta e uma bola cinza. Nesse modelo a regra manda que ao retirarmos uma bola de uma cor, devolvamos a bola retirada e acrescentemos outra bola da mesma cor à urna. Assim retirando uma bola cinza você a devolve e acrescenta outra bola cinza. Dessa forma a probabilidade inicial de tirar uma bola cinza que era de 1/2 passa a ser de 2/3 no tempo 2 (t2). Se no tempo 3 (t3) você retirar outra bola cinza, a probabilidade de tirar uma bola da mesma cor passará a ser de 3/4 e assim sucessivamente.

[13] ACEMOGLU, D. & ROBINSON. *Por que as Nações Fracassam; As Origens do Poder, da Prosperidade e da Pobreza* (Rio de Janeiro, RJ: Elsevier, 2012).

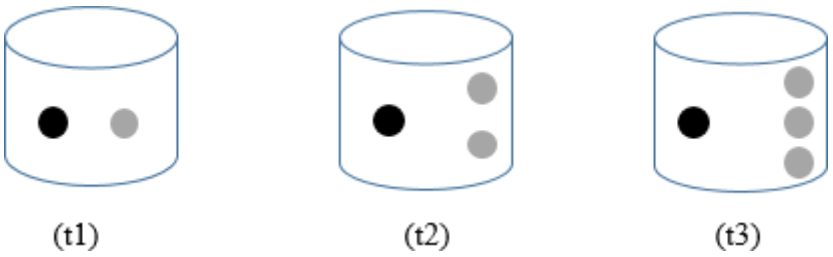

Figura 1. Modelo de urnas em um processo de polya.

A probabilidade se modifica a cada tempo e essa mudança é notavelmente dependente da trajetória. Um processo extremamente simples como esse demonstra que inicialmente qualquer coisa pode acontecer e todo acontecimento é igualmente provável, assim sendo, qualquer probabilidade de bolas prestas e cinzas é um equilíbrio a longo prazo. Isso significa que podemos ter comoresultado 4%, 50% ou mesmo 90% de bolas cinzas. A probabilidade de qualquerocorrência será a mesma. O que também podemos verificar é que qualquer sequência de eventos de bolas pretas e bolas cinzas é equiprovável.

Essas consequências podem ser extrapoladas para nossa concepção do surgimento de uma superinteligência artificial. Se considerarmos que uma IAG pode ser boa ou ruim (bola preta ou bola cinza) e utilizarmos esse modelo simples de urnas, concluiremos que existem probabilidades equiprováveis de que seu advento possa transformar o planeta em uma utopia ou uma distopia, ambas as possibilidades com a mesma probabilidade de ocorrer.

Outro processo, o Processo de Preponderância, demonstra mais claramente a importância das escolhas que fazemos durante a passagem do tempo. Nesse processo usamos o mesmo modelo de urna e começamos da mesma forma, com uma bola preta e uma cinza. Entretanto, nesse caso, ao retirar uma bola cinza em t1, devolveremos a bola e adicionaremos uma bola da mesma cor, mas em t2 ao retirarmos uma bola preta devolveremos a preta, adicionaremos uma preta e também uma cinza pela cinza adicionada no tempo anterior. Se em t3 retirarmos uma bola cinza, devemos devolvê-la adicionando uma bola da mesma cor e adicionar uma bola preta para a preta do período anterior e duas cinzas para as cinzas que adicionei nos dois períodos anteriores. Se em t4 retiramos uma bola preta devemos devolvê-la e adicionar outra bola preta,

adicionar uma bola cinza pelo período anterior, adicionar duas bolas pretas para o período 2 e quatro bolas cinzas para os períodos anteriores.

O processo fica mais evidente se pudermos visualiza-lo:

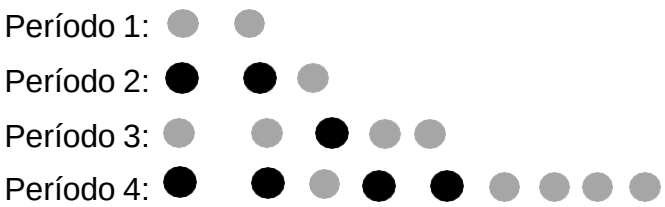

Figura 2. Modelo de um processo de preponderância.

Observe que sequencialmente a cada bola retirada da urna soma-se 1, 2, 4, 8, 16, 32 e assim sucessivamente conforme o tempo avança. Esse modelo extremamente simples demonstra que, a medida em que retrocedemos no tempo, as decisões anteriores possuem mais relevância e ao fazermos uma escolha podemos prever em certa medida quais serão as consequências. Em outras palavras podemos dizer que a trajetória ganha cada vez mais influência, já que escolhas e movimentos anteriores são a causa de um efeito maior. Conclui-se então que o passado assume exponencialmente mais importância.

Na dependência da trajetória quando um caminho é escolhido, cada passo nessa mesma trajetória produz consequências que aumentam a atratividade dessa mesma trajetória na próxima rodada, gerando um ciclo poderoso de auto reforço em que os custos de transição para alternativas aumentam consideravelmente com o tempo, tornando uma mudança radical ou uma reversão de curso altamente improvável.

A partir do conceito de dependência da trajetória podemos afirmar que o passado condiciona o futuro, mas, mesmo um processo tão regular pode ser substituído por outra taxa de crescimento e sofrer uma ruptura caso haja uma mudança na trajetória ou no contexto, sendo necessário um grande esforço ou um grande choque, interno ou externo, para alterar o curso da história[14].

O problema com a ruptura é que ela possui a dificuldade inerente da imprevisibilidade. Enquanto funções exponenciais nos transmitem a sensação de segurança e previsibilidade, ao atingir esse suposto “joelho da curva”, onde coisas

[14] Até mesmo a Lei de Moore tem data de validade. Segundo Ray Kurzweil a Lei de Moore é o quinto paradigma a impactar o crescimento exponencial da tecnologia. A Lei de Moore surgiu por volta de 1958 e cumprirá seus 60 anos de serviços úteis por volta de 2018, quando outra tecnologia computacional continuará o crescimento do ponto onde a Lei de Moore parar.

realmente importantes acontecem como resultado da crescente ordem dos sistemas computacionais, uma ruptura na regularidade provocará inevitavelmente a sensação de insegurança e imprevisibilidade.

Essa ruptura na regularidade será o que chamamos agora de Singularidade. O processo de evolução das máquinas persegue um caminho dependente da trajetória, mas a inserção de uma superinteligência artificial nesse processo poderá provocar uma inflexão no crescimento exponencial em forma de *tipping point*.

3 TIPPING POINT

> História e sociedades não se arrastam. Elas dão saltos. Seguem de ruptura a ruptura, intermediada por poucas vibrações. Ainda assim, nós (e os historiadores) gostamos de acreditar no progresso previsível e em pequenos incrementos.
>
> - Nassim N. Taleb

A história ocorre em rupturas, em pontos de inflexão. Em 1346 a febre bubônica, mais conhecida como a peste negra chegou a Europa trazida da China por mercadores que faziam a Rota da Seda. A doença era transmitida por ratos que espalhavam suas pulgas por toda a parte disseminando rapidamente a doença mortífera. Em 1347 a peste chegava a Constantinopla e em 1348 invadiu a França, o Norte da África, a Itália e até mesmo à isolada Inglaterra. Ao atingir uma determinada área, a peste aniquilava metade da população.

O impacto sobre uma sociedade de catástrofes dessa magnitude não pode ser calculado com precisão. Mas as consequências da peste negra sobre o sistema de trabalho da Europa foram perceptíveis. No século XIV a Europa era uma ordem feudal, tendo seus fundamentos hierárquicos com reis no topo da pirâmide, nobres como seus subordinados e os camponeses na base de toda a estrutura. O rei era o dono das terras que as concedia aos senhores em troca de serviço militar. Os nobres alocavam as terras aos camponeses que trabalhavam praticamente sem salário além de se submeterem a inúmeras multas e impostos. Os camponeses não podiam se deslocar sem a permissão do seu senhor, que era

quem determinava a lei e a ordem da época. A riqueza saia das mãos dos camponeses para os nobres.

Porém, o advento da peste negra reconfigurou essa estrutura. Com as inúmeras mortes, a escassez de mão de obra abalou o sistema feudal, pois os camponeses ingleses passaram a exigir a redução das multas e do trabalho não remunerado. Muitas exigências foram atendidas e os camponeses em toda a parte passaram a libertar-se de boa parte da servidão.

O governo tentou pôr um fim a tudo isso com a criação do Estatuto dos Trabalhadores (1351) que visava principalmente fixar os salários nos níveis anteriores à peste. Mas não deu certo, em 1381 irrompeu a Guerra dos Camponeses na Inglaterra que eliminou qualquer possibilidade de instalação do Estatuto e dessa forma o trabalho feudal foi definhando, emergindo um mercado de trabalho inclusivo na Inglaterra e em outros países da Europa Ocidental. A peste foi para a Europa um *Tipping Point*[15].

Os *Tipping Point* (TP), ou ponto de inflexão, são modelos não lineares em que uma pequena mudança na estrutura do sistema resulta em uma alteração brusca na trajetória. O gráfico abaixo é um bom exemplo de um TP, nesse caso demonstrando uma queda abrupta na população empregada nos EUA pós crise financeira de 2008:

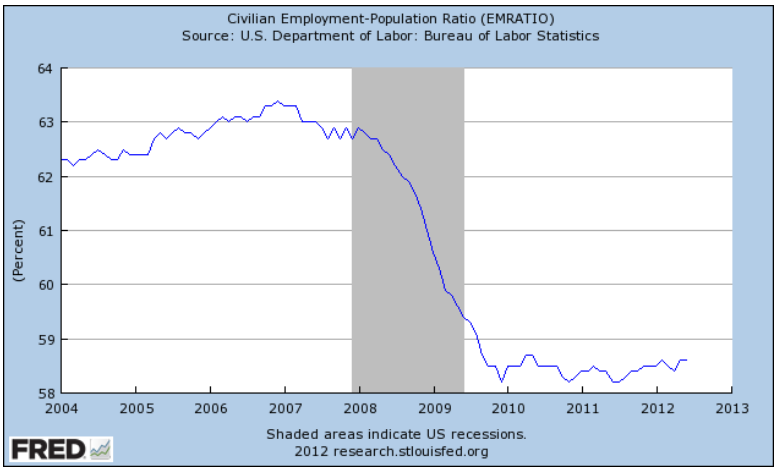

Figura 4. População empregada nos EUA pós crise financeira de 2008[16].

Existe, entretanto, uma diferença de grau em um sistema dependente da

[15] A pandemia do novo Corona Vírus (COVID-19) provavelmente é um TP na sociedade global. Ainda é cedo para detalhar seu impacto nas instituições políticas, econômicas, sociais e, também na tecnologia da medicina, mas valerá uma análise em breve.

[16] Gráfico disponível on-line: http://bomble.com/tag/emratio/.

trajetória – que analisamos anteriormente - cujo caminho segue um crescimento exponencial e um sistema onde ocorre um TP: Enquanto no crescimento exponencial ocorre uma curva que decola de forma previsível, no TP as mudanças são abruptas; há uma quebra na regularidade podendo acelerar de forma notável o crescimento, interrompê-lo ou levá-lo à outra direção. Nos sistemas dependentes da trajetória as alterações no processo mudam nosso destino provável de forma gradual; no TP a mudança ocorre de forma inesperada sendo que um único evento pode virar todo o sistema repentinamente, como por exemplo, quando uma gota d'água faz o copo transbordar.

Mas, como a evolução de sistemas de IA que aparentemente seguem a dependência da trajetória poderá inesperadamente se romper em um TP? Com a inevitabilidade do surgimento de uma superinteligência com vontade própria e objetivos finais que possam não levar em consideração os valores que prezamos.

Ocorrendo a Singularidade em forma de um TP, as consequências não poderão ser previstas com precisão devido ao alto grau de incerteza inerentes a esse tipo de ruptura. A incerteza nos sistemas onde ocorrem um TP é alta pois uma pequena alteração em uma variável leva a uma grande mudança no resultado final; uma pequena mudança no ambiente pode provocar acontecimentos drásticos[17]. Aferir medições em processos onde há ocorrência de TP podem ser feitas pela redução das incertezas. Porém, isso somente se torna possível após a virada, provavelmente tarde demais. Tomemos como exemplo que a ocorrência de TP na evolução dos sistemas de IA possibilite apenas dois resultados:

1. *Utopia:* O mundo é automatizado e dominado por uma IA pacífica que respeita os valores de seus criadores e contribui para que os humanos alcancem suas grandes aspirações;

2. *Distopia:* O mundo é dominado por uma IA não amigável que não leve em consideração os valores humanos, podendo utilizar quaisquer meios

[17] Os TP podem ser separados em duas categorias: *1. Viradas diretas:* Quando uma ação específica provoca uma virada na mesma dimensão, ou seja, na mesma variável; *2. Viradas contextuais:* Quando uma mudança no ambiente torna possível ou provoca outro acontecimento.

para alcançar seus objetivos finais, tornando a humanidade obsoleta e descartável.

Quando a IA atingir a capacidade de todos os cérebros humanos conectados, ocorre o TP. Esse é o ponto crucial de nossa análise, o momento em que provavelmente a maior parte de nossas previsões serão inutilizadas. O nível de conhecimento que temos atualmente não é suficiente para determinarmos com precisão o que ocorrerá ao cruzarmos esse suposto horizonte de eventos.

As dúvidas surgem: Quão grande pode ser uma virada? É uma virada inesperada ou uma virada com certa probabilidade de acontecer? Como medir a extensão da virada? Para mensuramos esse sistema é preciso medir a sua incerteza, ou seja, a quantidade de informações necessárias para identificar o tipo de virada. Também é necessário identificar o índice de diversidade, ou seja, o número de tipos de virada.

No exemplo acima o índice de diversidade é facilmente encontrado, pois temos apenas duas opções: Utopia ou Distopia. Logo o índice de diversidade é 2. E no mesmo caso a entropia do sistema é apenas 1, pois tudo que precisamos é uma informação: O TP nos levou para esquerda ou para direita? Após a virada decisiva, o índice de diversidade cai para 1 e a entropia para 0 já que não há mais incerteza no sistema.

O cenário poderá se tornar mais complexo se, ao invés de dois caminhos possíveis, existirem 3, 4 ou N caminhos. Imagine se além de Utopia e Distopia tivermos mais opções como Estagnação ou Regressão. Isso aumentaria o índice de diversidade e consequentemente o grau de entropia do sistema.

Não se pode ter certeza que tamanha virada possa ter apenas resultados positivos. Lembra do resultado da peste negra para a Inglaterra e os demais países da Europa ocidental? O mesmo não ocorreu com o leste europeu como resultado do mesmo evento impactante. Após a peste os senhores da Europa Oriental se apropriaram de mais terras e expandiram os seus domínios, por isso, ao invés de se tornaram mais livres os trabalhadores do leste europeu viram sua fraca liberdade minguar. Os senhores aumentaram o controle sobres os camponeses num processo que ficou conhecido como segunda escravidão:

> Embora em 1346 houvesse poucas diferenças entre a

> Europa Ocidental e a Oriental em termos de instituições políticas e econômicas, em 1600 as duas partes do continente eram mundos completamente distintos. No Oeste, os trabalhadores estavam livres das obrigações, multas e regulamentações feudais,tornando-se peças-chaves na nova economia de mercado em expansão. No Leste, também, tomavam parte dessa organização econômica, mas como servos coagidos responsáveis pela produção de alimentosagrícolas demandados pelo Ocidente. Tamanha discrepância institucional foi fruto de uma situação em que as diferenças entre as duas regiões a princípio pareciam insignificantes: no Leste, os nobresmostravam-se um pouco mais organizados, comdireitos ligeiramente maiores e sobre a terra mais bem consolidadas. As cidades eram mais fracas e menores,os camponeses eram menos organizados. No esquema mais amplo da História, eram divergências muito sutis. Não obstante, essas pequenas diferenças entre leste e oeste teriam graves consequências para a vida de suas populações e para o futuro caminho do desenvolvimento institucional quando a ordem feudal se visse abalada pela peste negra. (ACEMOGLU & ROBINSON, 2012. p. 79).

Também não é possível saber de antemão, ou seja, antes da virada, qual será a alternativa vencedora, qual alternativa prevalecerá sobre as outras. Essa leitura somente pode ser feita *a posteriore*. Embora seja uma ilusão muito comum acharmos que podemos prever a escolha inicial, isso não passa do que muitos psicólogos chamam de distorção retrospectiva.

Um exemplo disso é que embora muitos livros e relatos descrevam “tensões crescentes” e “crises” que precederam a Primeira Guerra Mundial, o conflito na realidade foi uma surpresa. A guerra foi vista como inevitável apenas retrospectivamente pelos historiadores olhando para trás.

O historiador Niall Ferguson argumentou de forma empírica para provar a tese de que ninguém tinha a menor ideia de que um conflito armado ocorreria. O que ele fez foi analisar os valores dos títulos imperiais que normalmente desvalorizam quando conflitos armados são esperados, já que incluem as expectativas dos investidores em relação às necessidades de financiamento do governo. Os valores dos títulos estavam valorizados pouco antes do início da

Primeira Guerra Mundial e não refletiam a expectativa de um conflito[18].

Dada a inevitabilidade do surgimento de uma superinteligência e da incerteza inerente às consequências posteriores à Singularidade, nos resta determinar se uma superinteligência, configurada inicialmente para alcançar seus objetivos finais (seja lá quais forem esses objetivos), respeitará as normas éticas estabelecidas para o fortalecimento e manutenção de nossa sociedade.

4 INTELIGÊNCIA E MORAL

Historicamente possuímos exemplo de algumas figuras ilustres considerados gênios em seu campo de atuação, mas que escolheram se comportar de forma eticamente repreensíveis. O inventor Thomas Edison perseguiu Nicola Tesla. O físico Isaac Newton perseguiu e talvez tenha até mesmo roubado ideias de seu rival Robert Hooke[19]. O filósofo Heidegger se afilou ao nacional socialismo alemão. Superinteligência não significa necessariamente uma capacidade elevada para o comportamento moral.

4.1 Ética Para As Máquinas

Nossas considerações éticas estão diretamente relacionadas àconsciência, pois aparentemente somente a consciência parece possuir relevância ética. Objetos e coisas sem consciência somente tem sua relevância ética considerada à medida que afetam seres conscientes.

Obviamente, quando uma superinteligência surgir, elas serão mais do que objetos e coisas. Serão algum tipo de entidade que demonstrará características como sapiência e senciência - dois requisitos fundamentais para que indivíduos sejam inseridos em nossa comunidade moral. A sensiência pode ser definida como a disposição para experiência fenomênica ou o que muitos pensadores chamam de *qualia*, enquanto a sapiência está relacionada às características que consideramos superiores (sabedoria, autoconsciência e racionalidade)[20].

[18] FERGUSON, N. *O Horror da Guerra: Uma provocativa análise da Primeira Guerra Mundial* (São Paulo, SP: Planeta do Brasil, 2014).

[19] Sabe aquela famosa frase de Newton “se vi mais longe foi por estar de pé sobre ombros de gigantes”? Não tem nada de humildade nela, na realidade foi um comentário sarcástico escrito supostamente em uma carta enviada a Robert Hooke, seu rival, que possuía uma baixa estatura.

[20] Atualmente beneficiamos alguns animais com status moral, pois possuem disposição à experiência fenomênica, ou seja, instanciam algumas propriedades qualitativas. Mas – e digo isso temendo errar - somente

O *insight* decorrente dessa percepção ética é que no futuro, quando máquinas portarem algum tipo de experiência fenomênica da realidade se instanciarem algum tipo de propriedade qualitativa e/ou apresentarem capacidades superiores como autoconsciência, deverão adentar nossa esfera ética. Quando isso ocorrer a utilização de dois princípios éticos evitará que cometamos algumas formas de discriminação: (1) Princípio de 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. (2) Princípio de não-discriminação da ontogenia: se dois seres têm a mesma funcionalidade e a mesma experiência consciente e diferem apenas na forma como vieram a existir, então eles têm o mesmo status moral[21].

Mentes biológicas e mentes artificiais podem divergir drasticamente. Podemos entender a semelhança entre as mentes biológicas e por isso traçar um paralelo de similitude entre nossa mente e todas as outras mentes humanas, bem como entre nossas capacidades e motivações típicas como espécie. Encontramos também similaridades entre nossa mente e a mente de alguns animais. Compartilhamos com as outras espécies motivações semelhantes para sobrevivência, alimentação e reprodução. O mesmo paralelo não pode ser traçado ao comparamos nossas mentes à mente de uma inteligência artificial.

4.2 Objetivos Finais E Ética Das Máquinas

Máquinas inteligentes se tornarão responsáveis por controlar uma ampla gama de situações de nossas vidas e esse controle poderá trazer inúmeros benefícios, bem como resultar em riscos à nossa própria segurança e a de nossos descendentes. Os objetivos de um agente são fundamentais para definir sua moral, por isso, conhecer os objetivos de uma IA é relevante para sabermos se ela se comportará de forma a valorizar nossos princípios morais como, por

os humanos e os grandes símios possuem o que chamamos de sabedoria ou sapiência e por isso concedemos a eles maior status moral. Obviamente outros seres possuem tais qualidades e ainda não sabemos mensurar. Podemos citar o polvo e, antes que você ria, recomendo que veja o premiado documentário *Professor Polvo* (Direção de Pippa Ehrlich e James Reed. Netflix. África do Sul, 2020).

[21] BOSTROM, N. and YUDKOWSKY, E. *The Ethics of Artificial Intelligence* (Oxford University Press, 2011) On-line: http://www.nickbostrom.com/ethics/artificial-intelligence.pdf.

exemplo, não causar sofrimento desnecessário a uma criatura consciente. Pode ocorrer que ao tornar-se superinteligente, a IA não precise mais dos seres humanos, percebendo-os com indiferença ou, pior ainda, percebendo-os como um obstáculo à concretização de seus objetivos.

Para que uma SIA seja bem-sucedida basta que ela alcance seu objetivo final sem que seja necessário possuir as mesmas motivações que temos no que diz respeito a cooperação, lealdade ao grupo, reputação e outras características que supostamente mantêm nossa sociedade nos trilhos. Ora, uma superinteligência artificial buscando alcançar um objetivo final poderá se comportar de qualquer maneira possível, sem preocupação com as exigências éticas predeterminadas por organismos biológicos. O filósofo Nick Bostrom também parece pensar assim quando disse:

> 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 (Bostrom, 2012. p. 3).

Encontraremos grandes dificuldades em construir uma IA com o conjunto de valores que prezamos - mentes artificiais podem ter objetivos não-antropomórficos, ou seja, objetivos finais contrários aos interesses humanos. Alguns desses objetivos finais podem requererem escolhas instrumentais que prejudiquem ou contrariem os objetivos da espécie humana. Pode ocorrer, por exemplo, que uma IA tenha um objetivo final e para completa-lo seja necessário se auto preservar, já que isso aumentaria a probabilidade de obter sucesso. Uma IA poderá ter interesse em proteger a integridade de seu conteúdo não permitindo que seus objetivos iniciais sejam alterados. Outros objetivos de mentes artificiais podem estar relacionados ao auto aprimoramento ou ao alcance de uma perfeição tecnológica, que permitiria alcançar seus objetivos com mais eficácia.

Imagine por um momento uma superinteligência artificial com o objetivo final de dominar o mundo por determinar a política e a economia planetária. Essa

inteligência teria razão instrumental para aperfeiçoar as tecnologias que a tornariam capazes de moldar o mundo de acordo com seus interesses. Isso implicaria na aquisição de recursos que exaurissem a capacidade do planeta, com o objetivo de construir qualquer tipo de substrato físico. Os recursos também poderiam ser utilizados para criação de backups infinitos, bem como para criação e manutenção de máquinas exploradoras, de defesa e segurança, capazes de eliminar obstáculos à concretização dos objetivos finais de uma IA.

Considere aqui uma SIA com uma programação final de acabar com uma pandemia de um vírus respiratório como o COVID-19. O que a impediria de concluir que - sendo o vírus altamente contagioso e disseminado apenas entre seres humanos - exterminar todos os humanos contaminados seja a forma mais rápida e eficiente para alcançar seu objetivo final?

Estas considerações nos levam novamente ao problema da previsibilidade, um dos principais requisitos exigidos dos seres que admitimos em nossa esfera ética, sejam eles construídos em um substrato biológico ou em um substrato de silício[22]:

> 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 seja fácil de prever. Um agente pode muito bem pensar em maneira de seguir valores instrumentais relevantes que não ocorram prontamente para nós. Isso é especialmente verdadeiro para uma superinteligência, que poderia desenvolver um plano muito inteligente, mas contraintuitivo, para realizar seus objetivos, possivelmente explorando até mesmo fenômenos físicos ainda não descobertos (Bostrom, 2012. p. 13, 14).

Uma superinteligência que se aprimore reiteradamente poderá manipular fenômenos físicos que não somos capazes de imaginar, simulando diversas possibilidades do nosso mundo em alta velocidade. Poderá ser capaz de explorar a dimensão temporal como atualmente somos capazes de explorar as dimensões espaciais. Uma máquina que controle ou distorça o tempo – mesmo o subjetivo -

[22] Ao admitirmos que outros seres - biológicos ou feitos de silício - adentrem nossa esfera ética, eles precisam preencher alguns requisitos exigidos dos seres possuidores de status moral como transparência, previsibilidade, resistência à manipulação e responsabilidade.

poderá “conhecer” o futuro. Qualquer entidade com essa capacidade é inevitavelmente incontrolável, portanto invencível.

6 DON’T BE EVIL (explorando uma perspectiva pessimista)

Provavelmente já caminhamos para provocar, mesmo que despropositadamente, uma distopia tecnológica. Além da IBM e sua chamada “era cognitiva”, empresas como o *Google*, *Facebook* e *Neuralink* trabalham atualmente com sistemas de aprendizagem da máquina, capazes de compreender e até antecipar nossas vontades por meio do entendimento estatístico de nossos comportamentos ao navegarmos na internet.

Máquinas aprendizes trabalham com algoritmos de aprendizagem de sistemas artificiais, dedicada ao desenvolvimento de técnicas que permitem ao computador aprender e aperfeiçoar seu desempenho em qualquer tipo de tarefa. A ciência da aprendizagem das máquinas está intimamente ligada à mineração de dados, sendo que suas principais aplicações práticas incluem o processamentode linguagem natural, sistemas de buscas, criptomoedas, diagnósticos médicos, bioinformática, reconhecimento de fala e escrita, visão computacional e a locomoção de sistemas robóticos.

O *Google* pode ser classificado como um sistema de inteligência artificial altamente capacitado que em breve absorverá todo o conhecimento humano, com consequências totalmente imprevisíveis. Quando começou a desenvolver seu sistema de tradução, a empresa criou um programa prático de aprendizagem da máquina em larga escala com o nome de SETI (*Search for Extra Terrestrial Intelligence*).

O programa de aprendizagem da máquina do *Google*, é o mais completo projeto de aprendizagem artificial já empreendido e pode nos colocar no caminho de uma distopia:

> Os pesquisadores do Google reconhecem que trabalhar com um sistema de aprendizagem dessa escala os coloca em um território desconhecido. O progresso constante do sistema de aprendizagem do Google flertava com as consequências postuladas pelo

> cientista e filósofo Raymund Kurzweil, que especulou sobre uma iminente "singularidade" que surgiria quando um grande sistema computacional desenvolvesse sua forma de inteligência (Levy, 2012. p. 89).

O *Google* poderá ser a nossa *Skynet*[23]? O *Google* já é de fato uma forma de inteligência. Um sistema de inteligência artificial que aprende com o usuário e que desenvolve habilidades próprias[24]. Seus fundadores esperam que o *Google* conheça nosso comportamento, desejos, motivações e seja capaz de sugerir e encontrar coisas que queremos saber. Atualmente a busca começa a mostrar resultados antes mesmo de concluirmos a digitação da consulta, um vislumbre do que poderá ocorrer[25].

Com toda essa capacidade o *Google* se torna cada vez mais o vetor de nossas decisões diárias, sejam elas grandes ou pequenas. Em 2010, 70% das pessoas nos EUA utilizavam o Google para buscar informações. Provavelmente estamos delegando poder em excesso a uma empresa ou, pior do que isso, poder excessivo a uma entidade que tem como objetivo possuir todas as informações do mundo. Resta saber se o aprimoramento constante desse sistema de IA não

[23] Em referência ao sistema de SIA que tenta destruir a espécie humana na franquia *O Exterminador do Futuro*.

[24] Podemos identificar o Google como um sistema inteligente? Observe o que disse sobre isso Alfred Spector um dos líderes da divisão de pesquisa do Google: "Essa é uma questão bastante profunda", diz Spector. "Os humanos são, na verdade, grandes sacos cheios de, na maior parte, água, andando por aí com um monte de tubos e alguns neurônios e tal. Mas temos a capacidade de aprender. Então, veja agora o *cluster* do sistema de computadores do Google: trata-se de um conjunto de várias heurísticas, de modo que ele sabe que 'veículo' é um sinônimo de 'automóvel' e sabe que, em francês, isso é *voiture*, e sabe como é em alemão e em outras línguas. O sistema sabe coisas. E sabe muito mais coisas que são aprendidas com o que as pessoas digitam", Spector cita outras coisas que o Google sabe: por exemplo, o Google acabou de introduzir uma nova heurística com a qual pode determinar, a partir de suas buscas, se você está pensando em suicídio, caso em que lhe forneceria informação sobre fontes de ajuda. Nesse caso, o mecanismo do Google recolhe pistas predicativas a partir de suas observações do comportamento humano, pistas essas que são formuladas no cérebro virtual do Google exatamente como os neurônios são formados em nosso próprio cérebro. Spector assegura que o Google aprenderá muito, muito mais nos próximos anos (Levy, 2012. p. 89).

[25] Essa é a visão dos próprios fundadores. Veja o que disse Sergey Brin: "Ultimamente vejo o Google como uma forma de ampliar seu cérebro com o conhecimento do mundo. Agora você pega seu computador e digita uma frase, mas pode imaginar que isso poderia ser bem mais simples no futuro, que você pode simplesmente ter aparelhos nos quais fala, ou computadores que observam o que está acontecendo ao redor de si e sugerem informações úteis" (Levy, 2012. p.90).

contrariará o lema interno da empresa: *Don't be evil*[26].

7 CONCLUSÃO

Não é impossível criar uma IA que valorize as nossas aspirações, porém é mais fácil criarmos uma IAG que tenha como objetivos finais a autopreservação e o aprimoramento. Sistemas com tais objetivos terão razões instrumentais para agir de forma a eliminar qualquer tipo de obstáculo, dominar completamente o ambiente e exaurir todos os recursos disponíveis, sem levar em consideração nossos valores. Mesmo uma IA criada com a percepção de que para atingir seus objetivos, deve promover o bem-estar humano, ao se deparar com situações diferentes ou mudanças no ambiente, poderá ter suas percepções e motivações modificadas, deixando de agir de forma cooperativa.

Conforme vimos, mesmo que a evolução da IA seja um processo dependente da trajetória, com a ocorrência de um *tipping point* não há garantias de que o surgimento de uma superinteligência artificial traga benefícios à humanidade. Existem probabilidades equiprováveis de que após a Singularidade ocorra tanto uma utopia quanto uma distopia.

A ruptura no sistema trará consequências imprevisíveis para a humanidade. Mas, mesmo com tamanha incerteza faremos bem em desenvolvermos preceitos para a criação segura de uma superinteligência artificial. Esses preceitos poderão amenizar as consequências da Singularidade, caso se concretize uma distopia tecnológica. Devido a importâncias das escolhasatuais, vivemos o momento da história em que podemos criar o alicerce para um desenvolvimento seguro da IAG. Seus sistemas altamente complexos podem preencher os papeis sociais que exigimos em nossa esfera ética, mas isso implicaem novos projetos orientados para a transparência e previsibilidade.

[26] "*Don't be evil"* era o antigo lema não oficial do Google que após a reestruturação da empresa para o conglomerado Alphabet, foi gradualmente substituído para o ambíguo "*Do the right thing"* ou "Faça a coisa certa". Mas, a pergunta que deve ser feita é: A coisa certa para quem?

REFERENCIA

ACEMOGLU, D. & ROBINSON. *Por que as Nações Fracassam; As Origens do Poder, da Prosperidade e da Pobreza* (Rio de Janeiro, RJ: Elsevier, 2012).

BERNARDI, B. B. O *Conceito de Dependência da Trajetória (Path Dependence): Definições e Controvérsias Teóricas* (Perspectivas, São Paulo, v 41, p. 137-167, jan/jun. 2012).

BOSTROM, N. and YUDKOWSKY, E. *The Ethics of Artificial Intelligence* (Oxford University Press, 2011) On-line: http://www.nickbostrom.com/ethics/artificial-intelligence.pdf.

Tradução: BATISTA, P. A. A Ética da Inteligência Artificial (Fundamento – Rev. de Pesquisa em Filosofia, v. 1, n. 3. maio – ago. 2011) On-line: http://www.ierfh.org/br.txt/EticaDaIA2011.pdf

BOSTROM, N. *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agentes* (Oxford University Press, 2012). On-line: http://www.nickbostrom.com/superintelligentwill.pdf

Tradução: MACHADO, L. A Vontade Superinteligente: Motivação e Racionalidade Instrumental em Agentes Artificiais Avançados (IERFH, 2012) On-line: http://www.ierfh.org/br.txt/VontadeSuperinteligente2012.pdf

CHALMERS, D. *The Singularity: A Philosophical Analysis* (Journal of Consciousness Studies 17:7-65, 2010). On-line: http://consc.net/papers/singularity.pdf.

FERGUSON, N. *O Horror da Guerra: Uma provocativa análise da Primeira Guerra Mundial* (São Paulo, SP: Planeta do Brasil, 2014).

GLADWELL, M. O Ponto da Virada (*The tipping point*): Como Pequenas Coisas Podem Fazer uma Grande Diferença (Rio de Janeiro, RJ: Sextante, 2009).

KURZWEIL, R. *The Singularity Is Near: When Humans Transcend Biology* (Penguin Books; Illustrated Edition, 2006).

KURZWEIL, R. A Era das Máquinas Espirituais (São Paulo, SP: Alephe, 2007).

LEVY, S. Google a Biografia: Como o Google pensa, trabalha e molda nossas vidas (São Paulo, SP: Universo dos Livros, 2012).

PAGE, S. E. *Models Thinking* (University of Michigan, 2014). On-line: https://www.coursera.org/course/modelthinking

TALEB, N. N. A Lógica do Cisne Negro: O impacto do altamente improvável (Rio de Janeiro, RJ: *Best Business*, 2014).