2º EITCCC

# Anais do II Encontro Internacional Tecnologia, Comunicação e Ciência Cognitiva

Apoio: FAPESP CAPES

Volume 2, Número 1, Ano 2016
ISSN: 2358-4513

# SUPERINTELIGÊNCIA ARTIFICAL:

# UTOPIA OU DISTOPIA TECNOLÓGICA?

**Pablo de Araújo Batista**

**Universidade São Judas Tadeu**

**pablotitan@ig.com.br**

**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*

## Introdução

O evento denominado de 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.

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 as tarefas que atualmente desejamos abandonar ou tornando nossa existência totalmente obsoleta[1]. 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.

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. Nos processos dependentes da trajetória, o passado possui exponencialmente mais importância com o passar do

[1] Alguns possíveis cenários distópicos foram explorados em filmes como *O Exterminador do Futuro* e *Matrix.* Em ambos os longa metragem, as máquinas se rebelam e dominam a Terra exterminando ou escravizando a raça humana.

tempo 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 do surgimento de uma superinteligência artificial.

## 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 um adjetivo exclusivamente humano fosse dado a uma máquina.

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

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 e veículos terrestres, GPS, diagnósticos médicos e nas previsões da bolsa de valores. As máquinas atuais são sistemas extremamente simples que, manejadas de forma correta, não proporcionam qualquer tipo de risco a nossa existência.

Estamos tratando de um tipo de IA capaz de se autocompreender, tomar decisões e fazer escolhas relevantes com o intuito de atingir seus objetivos. 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[2].

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[3]. 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 denominaram Singularidade.

## A Singularidade

O filósofo David J. Chalmers argumentou de forma convincente no texto *The Singularity: A Philosophical Analysis*, sobre a possibilidade da Singularidade. 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. Provavelmente, quando isso ocorrer, nossos cérebros biológicos se tornarão obsoletos.

Nesse contexto, o termo Singularidade é providencial em conformidade com as concepções sobre 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

[2] 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. Em outras palavras, máquinas limitadas em inteligência e socialização podem colocar em perigo a segurança de seus criadores.

[3] 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.

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[4].

A evolução de sistemas artificiais e o aumento em sua complexidade segue uma linha exponencial baseada principalmente na Lei de Moore[5]. 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 na noção de crescimento exponencial, o que nos conduzirá à Singularidade, podemos fazer previsões acertadas sobre o avanço no desenvolvimento de sistemas de IA.

**Analisando a importância do tempo**

O tempo é de crucial importância em nossa análise sobre o advento de uma superinteligência artificial. Concebemos o tempo de forma linear, mas, essa suposta linearidade somente é percebida quando nada de relevante está ocorrendo. Quando eventos significativos ocorrem, a natureza exponencial do tempo se torna evidente. O inventor e futurólogo Ray Kurzweil denomina esses momentos significativos de "joelho da curva", ou seja, os momentos em que a natureza exponencial do tempo explode, para dentro ou para fora, para esquerda ou para direita:

> Está na natureza do crescimento exponencial que os eventos desenrolem de forma extremamente lenta por extremamente longos períodos de tempo, mas, à medida que se passa do "joelho da curva", os acontecimentos passam a se desenrolar em um ritmo cada vez mais furioso (KURZWEIL, 2007, p. 30).

[4] 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 de forma 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 relevantemente limitados.

[5] 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.

Ao analisarmos a importância do tempo, as implicações de caos e ordem são fundamentais:

> *Caos:* Quando o caos aumenta exponencialmente o tempo diminui exponencialmente, ou seja, o intervalo de tempo entre eventos relevantes aumenta com o passar do tempo.

Considere o exemplo do surgimento do universo. Após o *Big Bang* diversos acontecimentos épicos ocorreram em cascata acelerando a passagem do tempo. Porém, após o início da expansão e à medida que o caos aumentava, o tempo passava mais lentamente, aumentando, assim, o intervalo entre acontecimentos importantes.

> *Ordem:* O inverso também é verdade. A medida que a ordem aumenta exponencialmente, o tempo acelera exponencialmente, ou seja, o intervalo entre eventos relevantes diminui com o passar do tempo.

A evolução é um processo ordenado, pois é um sistema aberto altamente suscetível às influências externas, evitando o aumento da entropia ou do caos conforme previsto pela Segunda Lei da Termodinâmica. Sendo a tecnologia uma consequência inevitável do processo evolutivo, é a ordem, e não o caos, que está presente em seu advento. Portanto, o crescimento da complexidade da tecnologia sendo parte de um processo evolucionário e ordenado, ocorre de forma exponencial.

**O passado altera as probabilidades**

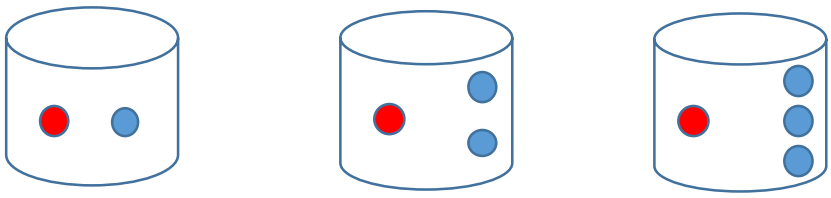
Os resultados de alguns processos são altamente dependentes da trajetória. O que isso significa? Significa que em tais processos as escolhas presentes e futuras são fundamentais, fazendo com que em uma visão retrospectiva, o passado assuma maior relevância. Se pudéssemos antecipar uma visão do futuro, poderíamos determinar com facilidade todo o processo de dependência da trajetória e identificar as escolhas que nos levaram inequivocamente à Singularidade.

O avanço da tecnologia é um processo dependente da trajetória (história), 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 que será a 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 continua defendida por Thomas Edson ou em forma de corrente alternada defendida por George Westinghouse e Nicola Tesla.

Podemos utilizar um modelo extremamente simples de urna para 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 vermelha e uma bola azul. Nesse modelo a regra manda que ao retirarmos uma bola de uma cor, devolvamos e acrescentemos outra bola da mesma cor. Assim retirando uma bola azul você a devolve e acrescenta outra bola azul. Dessa forma a probabilidade inicial de tirar uma bola azul que era de 1/2 passa a ser de 2/3 no tempo 2 (t2). Se no tempo 3 (t3) você retirar outra bola azul, a probabilidade de tirar uma bola da mesma cor passará a ser de 3/4 e assim sucessivamente.



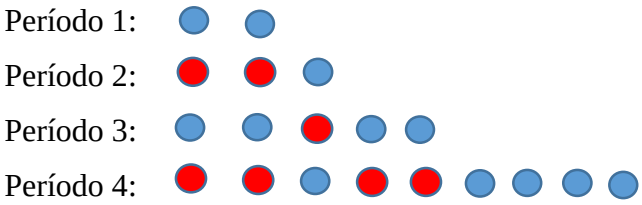
(t1) (t2) (t3)

A probabilidade se modifica a cada tempo e essa mudança é notavelmente dependente da trajetória. Um processo extremamente simples como esse demonstra que qualquer coisa pode acontecer e todo acontecimento é igualmente provável, assim sendo, qualquer probabilidade de bolas vermelhas e azuis é um equilíbrio a longo prazo. Isso significa que podemos ter como resultado 4%, 50% ou mesmo 90% de bolas azuis. A probabilidade de qualquer ocorrência será a mesma. O que também podemos verificar é que qualquer sequência de eventos de bolas vermelhas e bolas azuis 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 vermelha ou bola azul) 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, 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 azul e uma vermelha. Entretanto, nesse caso, ao retirar uma bola azul em t1, devolveremos a bola e adicionaremos uma bola da mesma cor, mas em t2 ao retirarmos uma bola vermelha devolveremos a vermelha, adicionaremos uma vermelha e também uma azul pela azul adicionada no tempo anterior. Se em t3 retirarmos uma bola azul, devemos devolvê-la adicionando uma bola da mesma cor e adicionar uma bola vermelha para a vermelha do período anterior e duas azuis para as azuis que adicionei nos dois períodos anteriores. Se em t4 retiramos uma bola vermelha devemos devolvê-la e adicionar outra bola vermelha, adicionar uma bola azul pelo período anterior, adicionar duas bolas vermelhas para o período 2 e quatro bolas azuis para os períodos anteriores.

O processo fica mais evidente se pudermos visualiza-lo:




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. 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.

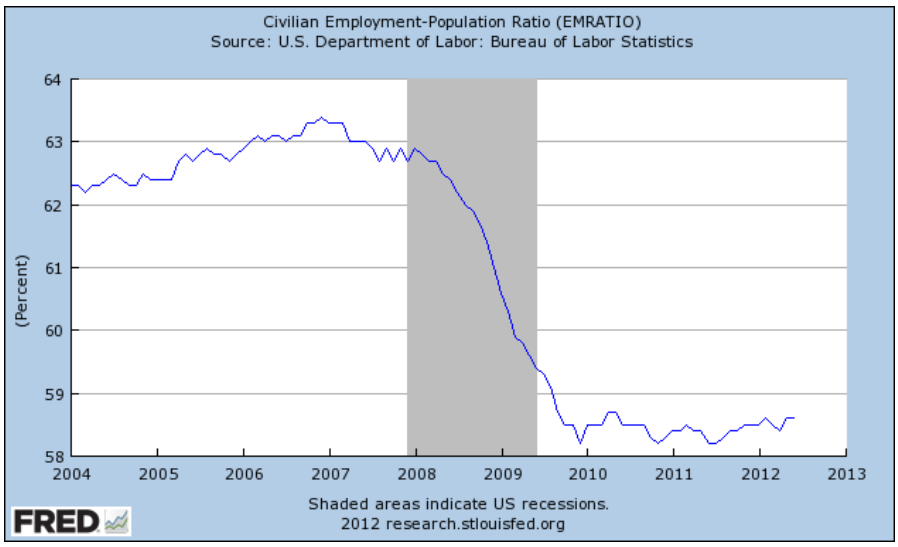
O que o processo de preponderância também demonstra é que, embora não saibamos inicialmente o que ocorrerá, ao fazermos uma escolha podemos prever quais serão as consequências. No entanto, mesmo um processo tão regular pode ser substituído por outra taxa de crescimento e sofrer uma ruptura caso haja uma pequena mudança na

trajetória ou no contexto[6]. 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 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*.

### *Tipping point*

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[7]:




[6] 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 1918, quando outra tecnologia computacional continuará o crescimento do ponto onde a Lei de Moore parar.

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

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

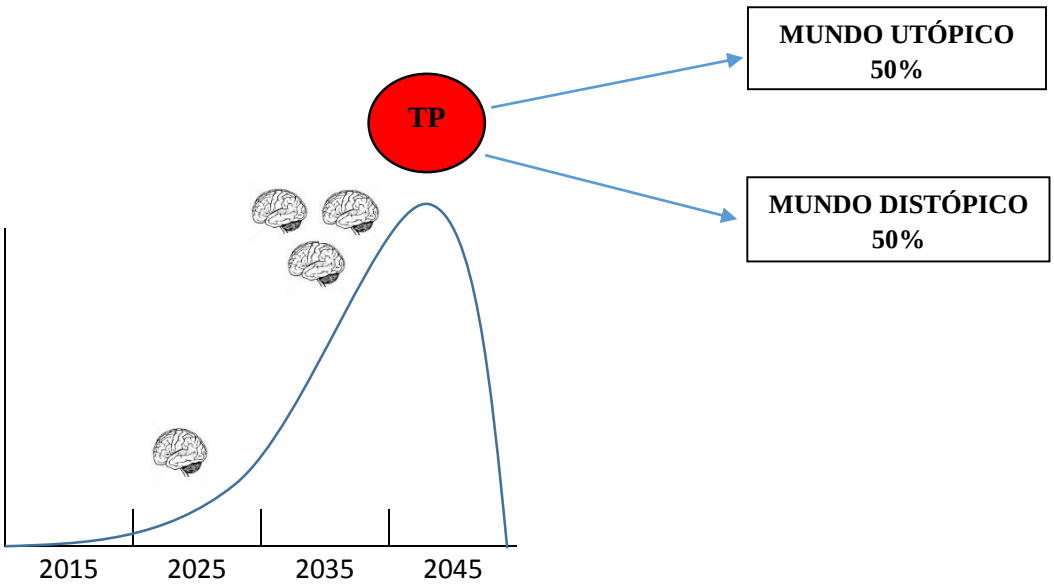
Ocorrendo a Singularidade em forma de um TP, as consequências não poderão ser previstas com precisão devido ao alto grau de incerteza inerentes a esse tipo de ruptura. A imprevisibilidade é o principal problema dos TP ao projetarmos sistemas altamente inteligentes. Será preciso desenvolver uma forma de analisar e prever o comportamento de tais agentes.

A incerteza nos sistemas onde ocorrem um TP é alta pois uma pequena mudança em uma variável leva a uma grande mudança no resultado final; uma pequena mudança no ambiente pode provocar acontecimentos drásticos[8]. Aferir medições em processos onde há ocorrência de TP podem ser feitas pela redução das incertezas. Porém, isso somente se tornará 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;

[8] 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.

2. *Distopia:* O mundo é dominado por uma IA não amigável que não leve em consideração os valores humanos, podendo utilizar quaisquer meios para alcançar seus objetivos finais, tornando a humanidade obsoleta e descartável.




Observe que ao 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.

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.

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 nossas normas éticas estabelecidas para o fortalecimento e manutenção de nossa sociedade.

**Inteligência e moral**

Superinteligência não significa necessariamente uma capacidade elevada para o comportamento moral. Historicamente possuímos exemplo de 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[9]. O filósofo Heidegger se afilou ao nacional socialismo alemão.

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. 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.

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 exemplo, não causar sofrimento desnecessário a uma criatura consciente. Nossas considerações éticas estão diretamente relacionadas à consciência, pois aparentemente somente a consciência parece possuir

[9] 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.

relevância ética. Objetos e coisas sem consciência somente tem sua relevância ética considerada à medida que afetam seres conscientes[10].

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. Mentes biológicas e mentes artificiais podem divergir drasticamente. Para que uma IA 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).

[10] 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). Atualmente beneficiamos alguns animais com status moral, pois possuem disposição à experiência fenomênica, ou seja, instanciam algumas propriedades qualitativas. Mas somente os humanos e os grandes símios possuem o que chamamos de sabedoria ou sapiência e por isso concedemos a eles maior status moral. 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.

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.

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[11]:

> Deve ser enfatizado que a existência de razões instrumentais convergentes, mesmo se elas se aplicarem a e forem reconhecidas por um agente especifico, 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).

[11] 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.

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

***Don't be evil***

Provavelmente já caminhamos para provocar, mesmo que despropositadamente, uma distopia tecnológica. Empresas como o *Google* 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 processamento de linguagem natural, sistemas de buscas, 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 ocupacional desenvolvesse sua forma de inteligência (Levy, 2012. p. 89).

O *Google* é de fato uma forma de inteligência. Um sistema de inteligência artificial que aprende com o usuário e que desenvolve habilidades próprias[12]. 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[13].

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 contrariará o lema interno da empresa: *don't be evil*.

## 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 auto preservar e aprimoramento. Sistemas com tais objetivos terão razões instrumentais para agir de forma

[12] 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).

[13] 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).

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, devido a um *tipping point* (Singularidade), 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 escolhas atuais, 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 implica em novos projetos orientados para a transparência e previsibilidade.

**BIBLIOGRAFIA**


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.

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. 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