# *Prevenção Contra o Risco Existencial*

## A Mais Importante Tarefa Para a Humanidade

*por Nick Bostrom,* 2012*

Tradução de Lucas Machado

(2012) Nick Bostrom
Faculty of Philosophy & Oxford Martin School
University of Oxford
**www.existential-risk.org**
**www.nickbostrom.com**
[Original em **pdf**]

## RESUMO

Riscos existenciais são aqueles que ameaçam o futuro inteiro da humanidade. Muitas teorias do valor implicam que mesmo reduções relativamente pequenas em risco existencial líquido têm um valor estimado enorme. Apesar de sua importância, problemas envolvendo riscos de extinção humana e perigos semelhantes continuam a ser mal compreendidos. Nesse artigo, eu clarifico o conceito de risco existencial e desenvolvo um esquema aprimorado de classificação. Eu discuto a relação entre riscos existenciais e problemas básicos em axiologia, e mostro como a redução de riscos existenciais (via a regra de *maxipok*) podem servir como um princípio forte de orientação de ações para preocupações utilitaristas. Eu também mostro como a noção de risco existencial sugere uma nova maneira de pensar sobre o ideal de sustentabilidade.

**Palavras-chave:** risco existencial, risco catastrófico, futuro da humanidade, extinção humana, sustentabilidade, *maxipok*, ética de população.

# 1. A regra maxipok

## 1.1. Risco existencial e incerteza

Um risco existencial é aquele que traz a ameaça da extinção prematura da vida inteligente originária da Terra ou a destruição drástica e permanente de seu potencial para futuros desenvolvimentos desejáveis [(1)]. Por mais que seja frequentemente difícil avaliar a probabilidade de riscos existenciais, há muitas razões para supor que o risco total confrontando a humanidade nos próximos séculos é significante. Estimativas de 10-20% de risco existencial nesse século são razoavelmente típicas entre aqueles que examinaram a questão, embora essas estimativas inevitavelmente se baseiem em juízos extremamente subjetivos[1]. A estimativa mais razoável pode ser substancialmente maior ou menor. Mas talvez a maior razão para julgar que o total de risco existencial nos próximos séculos será significante é a extrema magnitude dos valores em risco. Mesmo uma pequena probabilidade de risco existencial poderia ser, na prática, altamente significante. [(4, 5, 6, 61)]

A humanidade sobreviveu ao que nós podemos chamar de riscos existenciais naturais por centenas de milhares de anos; logo, a princípio, é improvável que qualquer um deles acabará conosco dentro do próximo século[2]. Essa conclusão encontra suporte quando nós analisamos os riscos específicos vindos da natureza, como impactos de asteroides, erupções supervulcânicas, terremotos, estouros de raios-gama, e assim por diante. Distribuições empíricas de impacto e modelos científicos sugerem que a probabilidade da extinção por causa desses riscos é extremamente pequena em uma escala de tempo de aproximadamente um século.[3]

---

[1] Uma enquete informal entre muitos acadêmicos *experts* em riscos de catástrofes globais deu uma estimativa média de 19% de probabilidade de que a espécie humana seja extinta antes do fim desse século [(2)]. As visões desses participantes não são necessariamente representativas da comunidade expert mais ampla. A influente e britânica Resenha Severa Sobre a Economia de Mudança Climática (*Stern Review on the Economics Climate Change*, 2006) usou uma probabilidade de extinção de 0,1% por ano para calcular uma taxa de desconto efetiva. Isso é o equivalente a assumir um risco de 9,5% de extinção humana nos próximos cem anos [(3: Capítulo 2, Apêndice Técnico, p.47)].

[2] A força dessa consideração é em parte enfraquecida pela possibilidade de efeitos de observação seletiva lançando uma "sombra antrópica" na evidência disponível [(7)].

[3] Cf. (60).

Em contraste, nossa espécie está introduzindo tipos inteiramente novos de riscos existenciais – ameaças das quais não temos registro de já termos sobrevivido. Nossa longevidade como espécie, por conseguinte, não oferece fundamentos prévios para termos um otimismo confiante. A consideração dos cenários específicos de risco existencial leva à suspeita de que a maior parte do risco existencial no futuro previsível consiste de riscos existenciais antropogênicos – isso é, esses que surgem da atividade humana. Em particular, a maior parte dos maiores riscos existenciais parece estar ligada a grandes descobertas científicas futuras que podem expandir radicalmente nossa habilidade para manipular o mundo exterior ou a nossa biologia. Na medida em que nossos poderes se expandem, também aumenta a escala de suas consequências em potencial – intencionadas e não-intencionadas. Por exemplo, parece haver riscos existenciais significantes em algumas das formas avançadas de biotecnologia, nanotecnologia molecular, e inteligência de máquinas que podem ser desenvolvidas nas próximas décadas. O 'grosso' do risco existencial no próximo século pode, então, residir em cenários consideravelmente especulativos para os quais nós não podemos atribuir probabilidades precisas por meio de qualquer método estatístico ou científico rigoroso. Mas o fato de que a probabilidade de algum risco é difícil de quantificar não implica que o risco seja negligenciável.

A probabilidade pode ser entendida em sentidos diferentes. O mais relevante aqui é o sentido epistêmico no qual a probabilidade é construída como (algo parecido com) a credibilidade que um observador idealmente racional deveria atribuir para a materialização do risco baseado na evidência atual.[4] Se não se pode saber atualmente que algo é objetivamente seguro, ele é arriscado ao menos no sentido subjetivo relevante para o processo de decisão. Uma caverna vazia não é segura precisamente no sentido de que você não pode dizer se ela é o lar de um leão faminto. Seria racional que você evitasse a caverna se você julga racionalmente que o malefício estimado por entrar na caverna é maior do que o benefício estimado.

A incerteza e tendência ao erro de nossas avaliações de primeira ordem do risco é algo que, em si mesmo, nós devemos contar em nossas atribuições de probabilidade 'levando tudo em conta' [*all-things-considered*].

[4] A probabilidade, por conseguinte, é indexada ao tempo. Quantidades que dependam de probabilidade, como a gravidade de um risco, podem variar ao longo do tempo na medida em que novas informações se tornam disponíveis.

Esse fator frequentemente predomina em riscos de probabilidades baixas, porém grandes consequências – especialmente aqueles envolvendo fenômenos naturais pouco compreendidos, dinâmicas sociais complexas, ou novas tecnologias, ou que são difíceis de avaliar por outras razões. Suponha que alguma análise científica *A* indica que alguma catástrofe *X* tem uma probabilidade *extremamente* pequena P(*X*) de ocorrer. Então a probabilidade de que *A* tenha alguma falha escondida e crucial pode facilmente ser muito maior do que P(X).[5] Além disso, a probabilidade *condicional* de *X* dado que *A* é crucialmente falha, $P(X \mid \neg A)$, pode ser consideravelmente alta. Nós podemos então descobrir que a maior parte do risco de *X* reside na incerteza de nossa avaliação científica de que P(X) era pequena (fig. 1).[(9)]

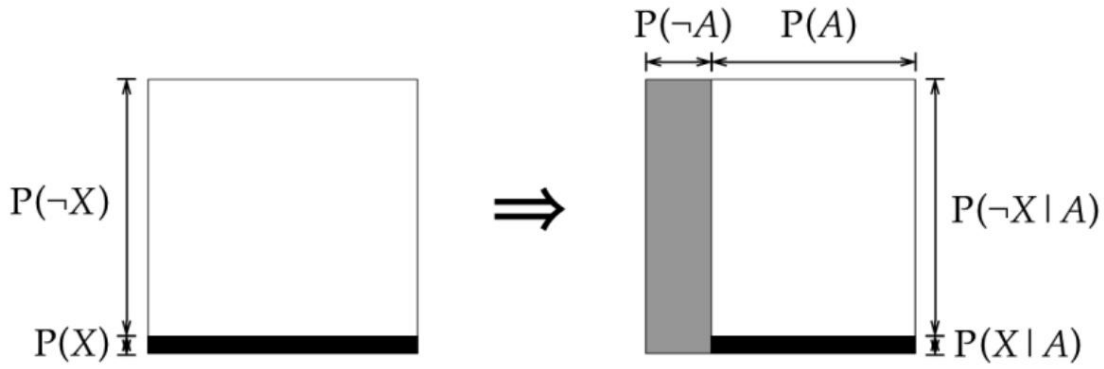


**Figura 1: Incerteza no nível Meta.** Levar em conta a falibilidade de nossa avaliação de primeira ordem pode ampliar a probabilidade de riscos avaliados como extremamente pequenos. Uma análise inicial (lado esquerdo) dá uma pequena probabilidade de desastre (tira preta). Mas a análise poderia estar errada; isso é representado pela área cinza (lado direito). A maior parte do risco 'levando tudo em conta' pode estar na área cinza em vez de na área preta.

## 1.2. Categorias de risco qualitativo

Já que um risco é um prospecto que é negativamente avaliado, a gravidade [*seriouness*] do risco – de fato, o que deve ser visto como arriscado – depende de uma avaliação. Antes que nós possamos determinar a gravidade de um risco, nós devemos especificar o critério de avaliação pelo qual o valor negativo de um cenário particular de perda possível é medido. Há vários tipos de tal critério de avaliação. Por exemplo, poderíamos usar uma função de utilidade que representa as preferências de um agente em particular sobre vários resultados. Isso pode ser apropriado quando a tarefa de alguém é auxiliar uma pessoa em particular que toma decisões. Mas aqui nós vamos considerar a *avaliação normativa*, uma atribuição

[5] Há evidência histórica ampla de que análises científicas aparentemente corretas são, às vezes, crucialmente falhas.

eticamente justificada de valores a vários resultados possíveis. Esse tipo de avaliação é mais relevante quando nós estamos investigando quais deveriam ser as prioridades de nossa sociedade (ou de nós mesmos enquanto indivíduos) na mitigação de riscos.

Há teorias conflitantes na filosofia moral sobre quais avaliações normativas são corretas. Eu não tentarei aqui adjudicar qualquer discordância axiológica de fundamentos. Em vez disso, vamos considerar uma versão simplificada de uma classe importante de teorias normativas. Suponhamos que as vidas de pessoas geralmente têm algum valor positivo significativo, e que esse valor é agregativo (no sentido que o valor de duas vidas similares é o dobro do de uma vida). Vamos assumir também que, mantendo a qualidade e a duração de uma vida constante, seu valor não depende de quando ela ocorre ou de se ela já existe ou virá ainda à existência como um resultado de eventos e escolhas futuros. Essas pressuposições poderiam ser flexibilizadas e complicações poderiam ser introduzidas, mas nós confinaremos nossa discussão ao caso mais simples.

Dentro dessa estrutura, então, nós podemos caracterizar grosseiramente a gravidade de um risco usando três variáveis: o *escopo* (o tamanho da população em risco), a *severidade* (o quanto essa população será prejudicada) e a *probabilidade* (o quão provável é que o desastre ocorra, de acordo com o juízo mais razoável, dada a evidência atual). Usando as duas primeiras dessas variáveis, nós podemos construir um diagrama qualitativo dos diferentes tipos de risco (fig. 2). (A dimensão de probabilidade poderia ser mostrada ao longo do eixo-z).

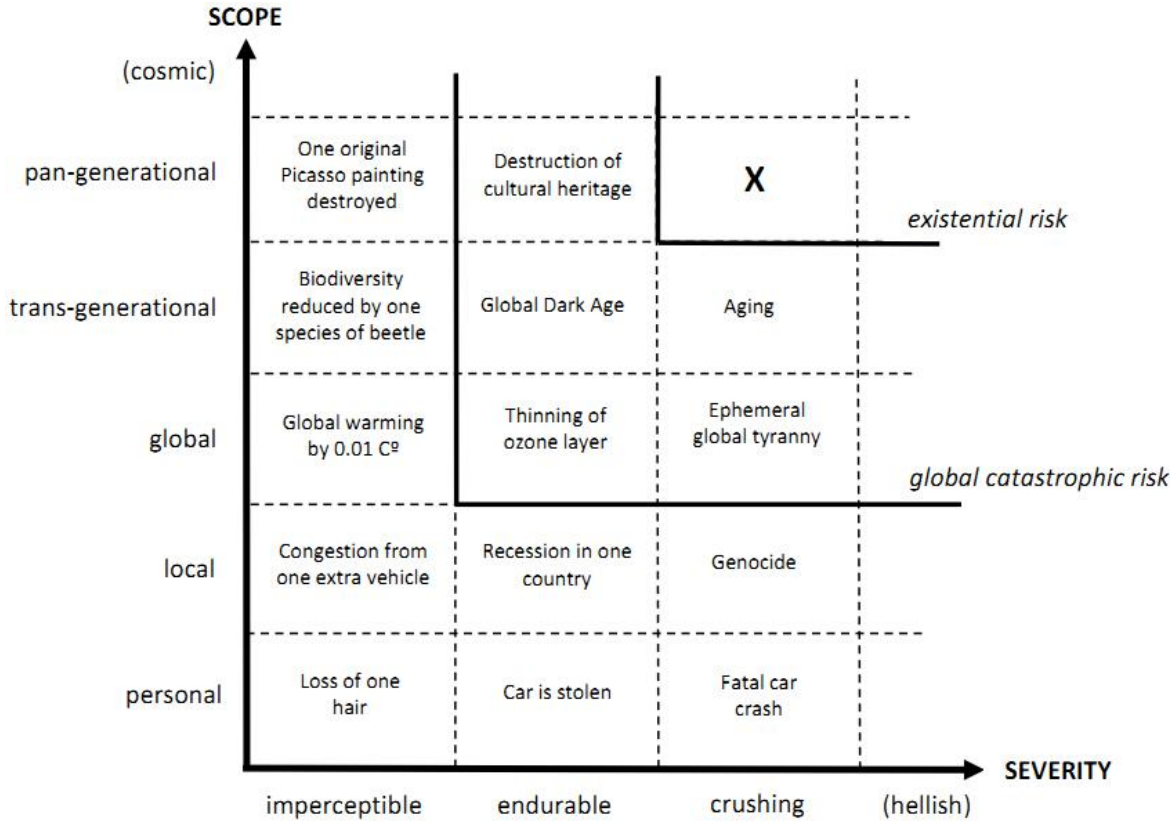


**Figura 2: Categorias de risco qualitativas.** O escopo de um risco pode ser *pessoal* (afetando apenas uma pessoa), *local* (afetando alguma região geográfica ou um grupo distinto), *global* (afetando toda a população humana ou uma grande parte dela, *trans-geracional* (afetando a humanidade por várias gerações) ou *pan-geracional* (afetando a humanidade por todas, ou quase todas as gerações futuras. A severidade do risco pode ser classificada como *imperceptível* (mal se pode notá-la), *duradoura* (causando dano significativo mas não arruinando completamente a qualidade de vida), ou *esmagadora* (causando a morte ou a redução drástica e permanente de qualidade de vida).

A área marcada com "X" na figura 2 representa riscos existenciais. Essa é a categoria de riscos que têm (pelo menos) severidade esmagadora e (pelo menos) escopo pan-geracional[6]. Como foi notado, um risco existencial é aquele que ameaça causar a extinção da vida inteligente originada na Terra ou a falha permanente e drástica dessa vida realizar o seu poten-

[6] Como indicado na figura, os eixos podem ser estendidos para abarcar riscos conceitualmente possíveis que sejam ainda mais extremos. Em particular, riscos pan-geracionais podem conter uma subclasse de riscos tão destrutivos que sua realização não apenas afetaria ou anteciparia gerações humanas futuras como também destruiria o potencial da parte do universo que reside em nosso cone de luz do futuro para produzir seres inteligentes ou autoconscientes (escopo *cósmico*). Além disso, de acordo com algumas teorias do valor, pode haver estados de ser que são muito piores do que a não-existência ou a morte (por exemplo, doenças horríveis e incuráveis), então, a princípio, se poderia estender o eixo-x também (severidade *infernal*). Nós não exploraremos essas possibilidades conceituais nesse artigo.

cial para desenvolvimento desejável. Em outras palavras, um risco existencial compromete todo o futuro da humanidade.

## 1.3. Magnitude da perda esperada em catástrofes existenciais

Mantendo a probabilidade constante, riscos se tornam mais graves na medida em que nós nos movemos para a região da direita superior da figura 2. Para qualquer probabilidade fixa, riscos existenciais são mais graves que outras categorias de risco. Mas *o quão* mais graves pode não ser intuitivamente óbvio. Poder-se-ia pensar que nós poderíamos compreender o quão ruim uma catástrofe existencial seria, considerando alguns dos piores desastres históricos em que nós podemos pensar – tal como as duas guerras mundiais, a pandemia de gripe espanhola, ou o Holocausto – e, em seguida, imaginando algo um pouco pior. Mas, se nós olharmos para as estatísticas globais de população ao longo do tempo, nós descobrimos que esses eventos horríveis do século passado falham em ser registrados (fig. 3).

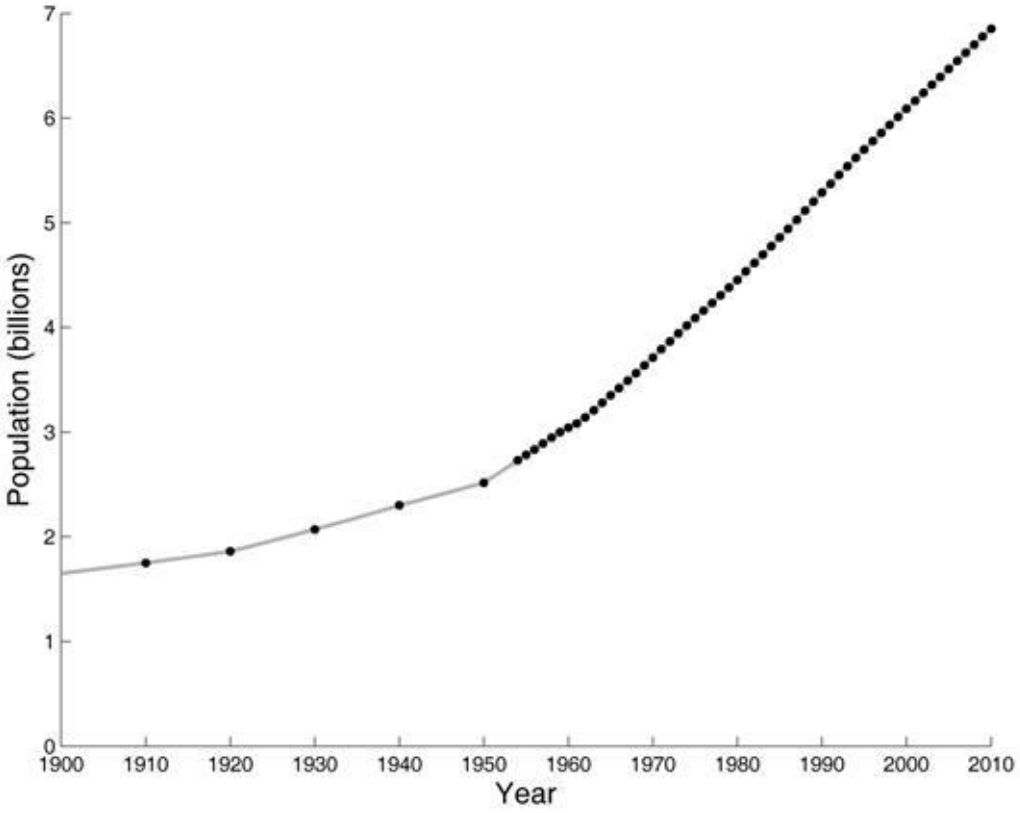


**Figura 3: População mundial ao longo do ultimo século.** Calamidades como a pandemia de gripe espanhola, as duas guerras mundiais ou o Holocausto mal são registradas. (Se se encarar esse gráfico por algum tempo, talvez se possa perceber uma leve redução temporária na taxa de crescimento de população mundial durante esses eventos).

Mas mesmo esse reflexo falha em mostrar a gravidade do risco existencial. O que faz catástrofes existenciais especialmente ruins não é que elas apareceriam robustamente em um gráfico como o da figura 3, causando uma enorme queda na população mundial ou na qualidade média de vida. Na verdade, sua significância reside primariamente no fato de que elas destruiriam o futuro. O filósofo Derek Parfit argumentou algo semelhante com o seguinte experimento:

> Eu acredito que se nós destruirmos a humanidade, tal como nós podemos atualmente, o resultado será muito pior do que a maior parte das pessoas pensa. Compare esses três resultados:
>
> (1) Paz.
>
> (2) Uma guerra nuclear que mata 99% da população mundial existente.
>
> (3) Uma guerra nuclear que mata 100%.
>
> (2) seria pior do que (1), e (3) seria pior que (2). Qual é a maior dessas duas diferenças? A maior parte das pessoas acredita que a maior diferença está entre (1) e (2). Eu acredito que a diferença entre (2) e (3) é *extremamente* maior. ... A Terra permanecerá habitável por pelo menos mais um bilhão de anos. A civilização começou apenas alguns milhares de anos atrás. Se nós não destruirmos a humanidade, esses milhares de anos podem ser apenas uma minúscula fração do todo da história da civilização humana. A diferença entre (2) e (3) pode ser, portanto, a diferença entre essa minúscula fração e o resto dessa história. Se nós compararmos essa história possível a um dia, o que aconteceu até agora é apenas uma fração de segundo. (10: p.453-454)

Para calcular a perda associada com uma catástrofe essencial, nós precisamos considerar quanto valor viria a existir em sua ausência. Acontece que o potencial total para a vida inteligente originada na Terra é literalmente astronômico. Nós obtemos um número grande mesmo se nós restringirmos nossas considerações para seres humanos biológicos vivendo na Terra. Se nós supusermos, como Parfit, que nosso planeta permanecerá habitável por pelo menos um bilhão de anos, e assumirmos que pelo menos um bilhão de pessoas poderia viver sustentavelmente, então, existe o potencial para pelo pelos $10^{16}$ vidas humanas. Essas vidas também poderiam ser consideravelmente melhores do que a média contemporânea de vida humana, que é tão frequentemente abatida pela doença, pobreza, injustiça e várias limitações biológicas que poderiam ser parcialmente superadas por meio de um progresso moral e tecnológico contínuo.

Entretanto, a quantidade relevante aqui não é quantas pessoas poderiam viver na Terra, mas sim quantos descendentes nós poderíamos ter

no total. Um limite inferior no número de anos de vida de humanos biológicos no universo de futuro acessível (baseado em estimativas cosmológicas atuais) é de $10^{34}$ anos[7]. Outra estimativa, que assume que as mentes futuras serão implementadas em *hardwares* de computadores em vez de *wetwares* biológicos neurais, produz um limite inferior de $10^{54}$ anos de vida subjetivos de emulações de cérebros humanos (ou $10^{71}$ operações computacionais básicas)[8] [(4)]. Se nós fizermos uma pressuposição menos conservadora de que civilizações futuras poderão pressionar os limites absolutos da física conhecida (usando alguma tecnologia ainda não imaginada), nós obtemos estimativas radicalmente mais altas da quantidade de computação e armazenamento de memória que seria alcançável e do número de anos subjetivos de experiência que seriam assim realizados[9].

Mesmo se nós usarmos a estimativa mais conservadora, que ignora completamente a possibilidade de colonização do espaço e mentes de *software*, nós descobrimos que a perda esperada de uma catástrofe existencial é maior do que o valor de $10^{18}$ vidas humanas. Isso implica que o valor estimado de reduzir um risco existencial em apenas *um milionésimo de um ponto de porcentagem* é pelo menos dez vezes o valor de um bilhão de vidas humanas. A estimativa mais tecnologicamente abrangente de $10^{54}$ anos de vida subjetivos de emulações de cérebros humanos (ou $10^{52}$ vidas de duração normal) reforçam o mesmo ponto ainda mais duramente. Mesmo se nós dermos a esse suposto limite menor sobre o potencial de *output* cumulativo de uma civilização tecnologicamente madura uma chance de mero 1% de estar correta, nós descobrimos que o valor estimado de reduzir um risco existencial em um mero *um bilionésimo de*

---

[7] Isso é baseado em um universo em aceleração com uma distância máxima de co-movimento alcançável de 4.74 Gpc, uma densidade de matéria bariônica de 4.55 x $10^{-28}$ kg/m$^3$, uma proporção luminosa de estrelas de ~100, e 1 planeta a cada 1000 estrelas sendo habitável por 1 bilhão de humanos por 1 bilhão de anos [(11,12)]. Obviamente, os valores dos últimos três parâmetros são discutíveis, mas o tamanho astronômico da conclusão é muito pouco afetado por uma mudança de algumas ordens de magnitude.

[8] Isso usa uma estimativa do falecido futurista Robert Bradbury de que uma estrela pode fornecer energia para $10^{42}$ operações por segundo usando computadores eficientes construídos com nanotecnologia. Além disso, ela assume (assim como as estimativas cosmológicas da nota de rodapé anterior) que o cérebro humano tem um poder de processamento de $10^{17}$ operações por segundo e que estrelas duram, em média, 5 bilhões de anos. Nenhuma formação de novas estrelas é presumida. Veja também (13).

[9] Por exemplo, se toda a massa-energia no universo acessível for poupada até que a temperatura da radiação cósmica de fundo pare de diminuir (devido à temperatura de horizonte constante de $10^{-29}$ K) e for, então, usada para computação, isso permitiria até $10^{121}$ computações termodinamicamente irreversíveis[(14)]. Ver também (15).

*um bilionésimo de um ponto de porcentagem* vale centenas de bilhões de vezes mais do que um bilhão de vidas humanas.

Seria possível argumentar consequentemente que mesmo a menor redução do risco existencial tem um valor maior do que o da provisão definida de qualquer bem "ordinário", como o benefício direto de salvar 1 bilhão de vidas. E, além disso, o valor absoluto do efeito *indireto* de salvar 1 bilhão de vidas na quantidade total cumulativa de risco existencial – positivo ou negativo – é quase que certamente maior do que o valor positivo de qualquer benefício direto de tal ação[10].

## 1.4. Maxipok

Essas considerações sugerem que a perda no valor estimado resultante de uma catástrofe existencial é tão gigantesca que o objetivo de reduzir riscos existenciais deveria ser uma consideração predominante quando nós agimos por uma preocupação impessoal com a humanidade como um todo. Pode ser útil adotar a seguinte regra geral [*rule of thumb*] para tal ação moral impessoal:

> *Maxipok*
>
> Maximize a probabilidade de um "resultado OK", em que um resultado OK é qualquer resultado que evite uma catástrofe existencial.

Na melhor das hipóteses, o *maxipok* é uma regra geral [*rule of thumb*] ou uma sugestão *prima facie*. Ela não é um princípio de validade absoluta, já que claramente há outras finalidades morais além da prevenção de uma catástrofe existencial. A utilidade do princípio é a de um auxílio para a priorização. Altruísmo irrestrito não é tão comum que nós possamos desperdiçá-lo em uma pletora de projetos que nos fazem nos sentirmos bem e têm uma eficácia sub-otimizada. Se beneficiar a humanidade aumentando a segurança existencial alcança um bem estimado em uma escala de tantas ordens de magnitude maior do que de contribuições alternativas, nós faríamos bem em nos focarmos nessa filantropia mais eficiente.

---

[10] Nós devemos enfatizar, no entanto, que há questões importantes não resolvidas no consequencialismo agregativo – em particular na relação com valores infinitos e chances extremamente pequenas.[16, 17] Não discutiremos essas questões aqui, mas, na seção 5, nós discutiremos o estatuto normativo do conceito de risco existencial a partir de outras perspectivas.

Note que o *maxipok* difere da máxima popular ("Escolhe a ação que tem o melhor pior resultado")[11]. Já que nós não podemos eliminar completamente o risco existencial – a qualquer momento, nós podemos ser lançados na lixeira da história cósmica pela frente de uma fase de transição de vácuo ativada em alguma galáxia remota um bilhão de anos atrás – o uso da máxima nesse contexto presente implicaria em escolher a ação que tem o melhor benefício assumindo-se a extinção iminente. A máxima implica, portanto, que nós deveríamos começar a festejar como se não houvesse amanhã. Essa implicação, apesar de tentadora, é implausível.

## 2. Classificação do Risco Existencial

Para trazer a atenção ao espectro completo do risco existencial, nós podemos distinguir quatro classes desse risco: *extinção humana, estagnação permanente, realização falha* e *ruína subsequente*. Nós podemos definilos da seguinte forma:

| **CLASSES DE RISCO EXISTENCIAL** | |
|---|---|
| **Extinção Humana** | A Humanidade é extinta prematuramente, quer dizer, antes de atingir maturidade tecnológica[12]. |
| **Estagnação permanente** | A Humanidade sobrevive mas nunca atinge a maturidade tecnológica.<br><br>Subclasses: *colapso sem recuperação*, *saturação* [plateauing], *colapso recorrente* |
| **Realização falha** | A Humanidade chega à maturidade tecnológica mas de uma forma que é terrível e irremediavelmente falha.<br><br>Subclasses: *realização não consumada*, *realização efêmera* |
| **Ruína subsequente** | A Humanidade chega à maturidade tecnológica de uma maneira que dá bons prospectos futuros, mas desenvolvimentos subsequentes causam a ruína permanente desses prospectos. |

[11] Seguindo John Rawls, o termo "máxima" [*maximin*] é usado em um sentido diferente do da economia de bem-estar, para denotar o princípio de que (dadas certas restrições) nós devemos optar pelo estado que maximiza a expectativa das classes menos prejudiciais [*worst-off classes*][(18)]. Essa versão do princípio não é necessariamente afetada pelas considerações do texto.

[12] Nós podemos nos referir a isso mais precisamente como extinção humana "precoce" ou "prematura". Note que a humanidade pode ser extinta sem instanciar essa categoria se a humanidade alcançar se potencial de capacidades e só depois for extinta.

Por “humanidade” nós queremos dizer aqui a vida inteligente originada na Terra, e por “maturidade tecnológica” nós queremos dizer a obtenção de capacidades que fornecem um nível de produtividade econômica e controle da natureza próximo do máximo que poderia ser obtido.

## 2.1. Extinção Humana

Por mais que seja concebível que, no um bilhão de anos em que a Terra pode permanecer habitável antes de ser sobreaquecida pelo Sol em expansão, uma nova espécie inteligente evolua em nosso planeta para preencher o nicho esvaziado por uma humanidade extinta, de modo algum é certo que isso ocorrerá. A probabilidade de uma vida inteligente recrudescente é reduzida se a catástrofe que causou a extinção da espécie humana também exterminou os grandes primatas e outros parentes próximos, como ocorreria em muitos (mesmo que não todos) cenários de extinção humana. Além disso, mesmo se outra espécie inteligente evoluísse para tomar nosso lugar, não haveria garantias de que a espécie que nos sucedesse evoluiria para tomar nosso lugar, não há garantias de que a espécie sucessora instanciaria suficientemente qualidades que nós temos razões para valorizar. A inteligência pode ser necessária para a realização de nosso potencial futuro de desenvolvimento desejável, mas não é suficiente. Todos cenários envolvendo a extinção prematura da humanidade serão contados como catástrofes existenciais, mesmo que alguns desses cenários possam, de acordo com algumas teorias do valor, ser relativamente benignos. Não é parte da *definição* de catástrofe existencial que ela é ruim quando se leva tudo em conta, por mais que essa seja uma suposição razoável na maior parte dos casos.

Acima nós definimos “humanidade” como vida inteligente originada na Terra, em vez de uma espécie biológica em particular definida como *Homo sapiens*[13]. A razão para nos concentrarmos no risco existencial nesse conceito mais amplo é que não há razão para supor que o conceito de espécies biológicas dá conta do que nós temos motivos para valorizar. Se nossa espécie evoluísse, ou usasse tecnologia para a auto-modificação, de tal modo que ela não mais satisfizesse o critério biológico para identidade de espécie (tal como inter-reprodutibilidade) com o *Homo sapiens* contem-

[13] Nós podemos entender aqui “inteligente” como capaz de desenvolver linguagem, ciência, tecnologia, e cultura cumulativa.

porâneo, isso não tem que ser, em qualquer sentido da palavra, uma catástrofe. Dependendo de em que somos mudados, essa mudança pode ser muito desejável. De fato, o impedimento permanente de qualquer possibilidade desse tipo de mudança da natureza biológica humana pode constituir, em si mesmo, um risco existencial.

A maior parte da discussão sobre riscos existenciais até hoje se concentrou inteiramente na primeira das quatro classes, "extinção humana". A presente estrutura chama a atenção para os outros três modos de falha para humanidade. Como a extinção, esses outros modos de falha envolveriam esmagamento pan-geracional. Eles são, por conseguinte, de gravidade comparável, implicando perdas igualmente grandes de valor estimado.

## 2.2. Estagnação Permanente

Estagnação permanente é instanciada como se a humanidade sobrevivesse, mas nunca alcançasse maturidade tecnológica – quer dizer, a obtenção de capacidades que nos fornecessem um nível de produtividade econômica e controle sobre a natureza que se aproximasse do máximo que pode ser obtido (na completude do tempo e na ausência de impedimentos catastróficos). Por exemplo, uma civilização tecnologicamente madura poderia (presumivelmente) começar uma colonização em larga escala do espaço pelo uso de "sondas Von Neumann" automáticas e auto-replicantes.[19, 20, 21] Ela também seria capaz de mudar e aprimorar a biologia humana – digamos, pelo uso de biotecnologia avançada ou de nanotecnologia molecular.[22, 23] Além disso, ela poderia construir *hardwares* computacionais extremamente poderosos e usá-los para criar emulações completas de cérebros e tipos inteiramente artificiais de mentes sencientes e superinteligentes.[24] Ela pode ter muitas outras capacidades, algumas das quais talvez não seja completamente imagináveis do nosso ponto de vista atual[14].

A destruição permanente da oportunidade da humanidade para obter maturidade tecnológica é *prima facie* uma perda enorme, porque as capacidades de uma civilização tecnologicamente madura poderiam ser usadas para produzir resultados que seriam plausivelmente de grande va-

[14] Não é necessário que uma civilização tecnologicamente madura *implemente de fato* todas essas tecnologias; basta que elas estejam *disponíveis* a ela, no sentido de que essa civilização poderia fácil e rapidamente desenvolvê-las e implementá-las caso decidisse fazê-lo. Então, uma civilização de máquinas superinteligentes suficientemente poderosa que pudesse inventar e implementar rapidamente essas e outras tecnologias já poderia contar como tecnologicamente madura.

lor, como números astronômicos de vidas extremamente longas e satisfatórias. Mais especificamente, a tecnologia madura permitiria um uso muito mais eficiente de recursos naturais básicos (como matéria, energia, espaço, tempo e negentropia [*negentropy*] para a criação de valor do que é possível com tecnologias menos avançadas. E a tecnologia madura permitiria a coleta (por meio da colonização do espaço) de muito mais recursos do que é possível com a tecnologia cujo alcance é limitado à Terra e sua vizinhança imediata.

Nós podemos distinguir vários tipos de cenários de estagnação permanente: *colapso sem recuperação* – muitas das nossas capacidades econômicas e tecnológicas atuais são perdidas e nunca recuperadas; *saturação* [plateauing] – o progresso satura em um nível talvez um pouco maior do que o atual mas abaixo da maturidade tecnológica; e *colapso recorrente* – um ciclo sem fim de colapsos seguidos de sua recuperação[15].(25)

A plausibilidade relativa desses cenários depende de um número de fatores. Pode-se esperar que mesmo que a civilização global passasse por um colapso completo, talvez depois de uma guerra global termonuclear, ela seria, eventualmente, reconstruída. Para ter um cenário plausível de colapso permanente, seria necessário, portanto, uma explicação de por que não ocorreria uma recuperação[16]. No que diz respeito à saturação, tendências modernas de mudanças sociais e tecnológicas rápidas fazem essa ameaça parecer menos iminente; contudo, cenários poderiam ser concebidos em que, por exemplo, um regime global estável impede a continuidade de mudanças tecnológicas[17]. Quanto a cenários de colapso recorrente, eles parecem requerer a postulação de um tipo especial de causa: um que (*a*) é forte o bastante para trazer o colapso total da civilização global mas (*b*) não é forte o bastante para causar a extinção humana e (*c*) pode plausivelmente ocorrer novamente cada vez que a civilização é reconstruída até um certo ponto, apesar da variação aleatória nas condições

---

[15] Não estritamente *sem* fim, é claro, mas apenas uma sequência de ciclos que ocorre por um longo período de tempo e termina com a extinção humana, sem que a maturidade tecnológica tenha sido obtida.

[16] Um cenário de colapso sem recuperação pode postular que algum recurso crítico para recuperação foi permanentemente destruído, ou que o conjunto de genes humanos degenerou irreversivelmente, ou talvez que uma descoberta seja feita que permite a pequenos grupos causar uma destruição tão imensa que eles podem arruinar a civilização, e o conhecimento dessa descoberta não pode ser erradicado.

[17] Técnicas aprimoradas de governo, como vigilância onisciente e manipulação neuroquímica, podem cimentar a posse de poder de um regime ao ponto de fazer sua queda impossível.

iniciais e de quaisquer tentativas de civilizações sucessivas de aprenderem com o fracasso de suas predecessoras. A probabilidade de permanecer em uma trajetória de colapso recorrente diminui com o número de ciclos postulado. Quanto maior o horizonte de tempo considerado (e isso também se aplica à saturação), maior a probabilidade de que o padrão será rompido, resultando ou em uma ruptura na direção superior, rumo à maturidade tecnológica, ou na direção inferior, rumo a um colapso sem recuperação e, talvez, à extinção (fig. 4)[18].

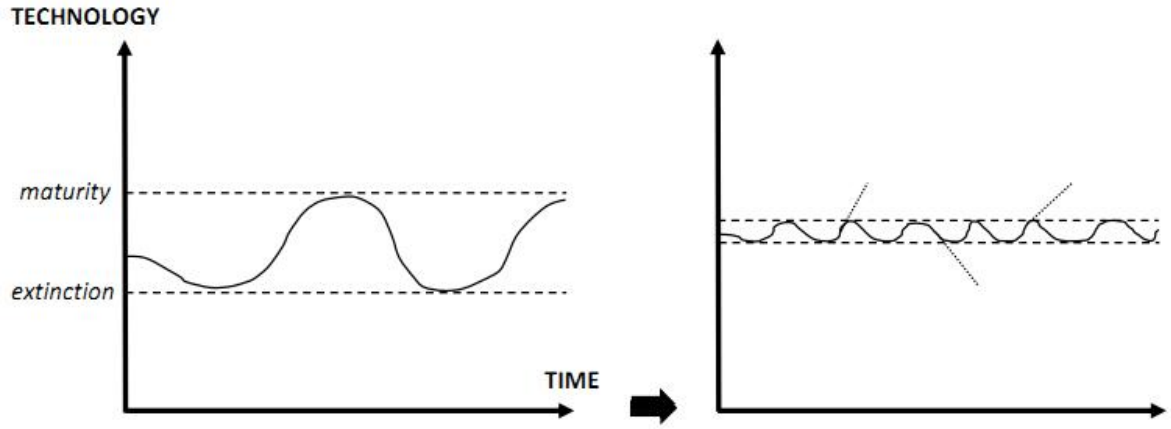


**Figura 4: Colapso ocorrendo indefinidamente?** A condição humana moderna representa um espectro estreito em um espaço de possibilidades. Quanto maior a escala de tempo considerada, menor a probabilidade que o nível de desenvolvimento tecnológico da humanidade permanecerá confinado ao intervalo definido no limite inferior de qualquer que seja a capacidade tecnológica necessária para a sobrevivência e, no limite superior, para a maturidade tecnológica.

## 2.3. Realização falha

Uma realização falha ocorre se a humanidade alcançar a maturidade tecnológica de um modo que seja terrível e irremediavelmente falho. Por "irremediável" nós queremos dizer que não é concebível que ela seja subsequentemente corrigida. Por "terrível" nós queremos dizer que ela permite a realização de apenas uma pequena parte do valor que poderia ter sido realizado. Classificar um cenário como uma instância de uma realização falha requer um julgamento de valor. Nós retornaremos a essa questão normativa na próxima seção.

[18] Outra dificuldade da hipótese do colapso recorrente é explicar o fato de que nós estamos no primeiro ciclo tecnológico aqui na terra. Se é comum que haja muitos ciclos de colapso e recuperação, como que nós podemos ainda estar no ciclo #1? Essa consideração de gênero antrópico pode sugerir que a extinção ou transformação é mais provável do que nós suporíamos ingenuamente.

Nós podemos distinguir duas versões da realização falha: *realização não consumada* e *realização efêmera*.

Na realização não consumada, a humanidade desenvolve a tecnologia madura mas falha em fazer bom uso dela, de tal modo que a quantidade de valor realizado é apenas uma pequena fração do que poderia ter sido obtido. Um exemplo desse tipo é um cenário em que uma inteligência maquinal substitui a inteligência biológica, mas as máquinas são construídas de tal forma que falta a elas consciência (no sentido de experiência fenomênica).(26) O futuro pode ser, nesse caso, muito rico e capaz, mas, em um sentido relevante, inabitado: não haveria (discutivelmente) seres moralmente relevantes para apreciar a riqueza. Mesmo se a consciência não desaparecesse inteiramente, poderia haver muito menos dela do que haveria em um uso mais otimizado dos recursos. Alternativamente, pode haver uma vasta quantidade de experiências, mas de qualidade muito menor do que deveria ser o caso: mentes que são muito menos felizes do que poderiam ser. Ou poderia haver vastos números de mentes muito felizes mas algum outro ingrediente crucial de um futuro maximamente valoroso estar faltando.

Na realização efêmera, a humanidade desenvolve a tecnologia madura que é, inicialmente, bem utilizada. Mas a maturidade tecnológica é obtida de tal maneira que o estado inicialmente excelente é insustentável e está condenado a degenerar. Há um lampejo de valor, seguido de um perpetuo crepúsculo ou escuridão. Uma maneira pela qual a realização efêmera poderia ocorrer seria se houvesse fraturas no estado inicial da maturidade intelectual que estivessem destinadas a levar a humanidade a se dividir em facções adversárias. Pode ser impossível reintegrar a humanidade depois de tal divisão ter ocorrido, e o processo de adquirir a maturidade tecnológica pode ter representado a última e melhor oportunidade para a humanidade formar um *singleton*(57). A ausência de coordenação global e vários processos podem degradar o potencial a longo prazo da humanidade. Um desses processos é a guerra entre grandes potências, apesar de talvez ser improvável que essas guerras sejam sem fim (em vez de terminarem eventualmente de maneira definitiva por um acordo ou conquista)[19]. Outro processo erosivo envolve formas indesejáveis de competição evolutiva e econômica em uma ampla ecologia de máquinas inteligentes(58). Ainda outro processo desse tipo seria uma raça colonizadora do

[19] Mesmo a ameaça de uma guerra que nunca ocorra pode resultar em muito desperdício, em termos de gastos com armas e oportunidades de colaboração perdidas.

espaço em que os replicadores podem desperdiçar recursos cósmicos para derrotar a competição [(59)].

## 2.4. Ruína subsequente

Em nome da completude, nós registramos uma quarta classe de riscos existenciais: a ruína subsequente. Em cenários desse tipo, a humanidade alcança a maturidade tecnológica com uma configuração inicial "boa" (no sentido de não ser terrível e irremediavelmente falha), mas desenvolvimentos subsequentes levam, de todo modo, à ruína permanente de nossos prospectos.

De uma perspectiva prática, nós não precisamos nos preocupar com a ruína subsequente. O que acontece depois que a humanidade alcançar a maturidade tecnológica não é algo que nós podemos afetar agora, *a não ser* garantindo que a humanidade a alcance e de uma maneira que ofereça os melhores prospectos possíveis para os desenvolvimentos subsequentes – quer dizer, evitando as outras três classes de risco existencial. Ainda assim, o conceito de ruína subsequente é relevante para nós de várias maneiras. Por exemplo, para estimar quanto valor estimado é obtido ao reduzir outros riscos existenciais em uma certa quantidade, nós precisamos estimar a condicional de valor estimado em evitar as primeiras três classes de riscos existenciais, o que requer estimar a probabilidade da ruína subsequente.

A probabilidade da ruína subsequente pode ser baixa – e é talvez um condicional extremamente baixo para conseguir a configuração certa. Uma razão para isso é que uma vez que nós tenhamos construindo muitas colônias espaciais auto-sustentáveis, qualquer desastre confinado a um único planeta não pode eliminar toda a humanidade. Outra razão é que uma vez de que a maturidade tecnológica tiver sido seguramente alcançada, há menos tecnologias potencialmente perigosas restantes para serem descobertas. Uma terceira razão é que uma civilização tecnologicamente madura seria superinteligente (ou teria acesso ao conselho de entidades artificiais superinteligentes) e seria mais capaz de prever o perigo e elaborar planos para minimizar o risco existencial. Por mais que previsão não reduza o risco se nenhuma ação efetiva estiver disponível, uma civilização com tecnologia madura pode tomar ação contra um grande espectro de riscos existenciais. Além disso, se ocorrer que adquirir a maturidade tecnológica sem adquirir a "singletonidade" condene a civilização à degeneração irre-

versível, então se a realização falha for evitada, nós podemos assumir que nossa civilização tecnologicamente madura pode resolver problemas de coordenação global, o que aumenta a habilidade de tomar ação para prevenir a ruína subsequente.

A fonte principal de risco de ruína subsequente pode muito bem ser um encontro com adversários inteligentes externos, como extraterrestres inteligentes ou simuladores. Note, entretanto, que cenários nos quais a humanidade eventualmente é extinta como resultado de limites físicos duros, como a morte térmica do universo, não contam como ruína subsequente, desde que antes de seu fim a humanidade tenha conseguido realizar uma parte razoavelmente grande de seu potencial para desenvolvimento desejável. Esses cenários não são catástrofes existenciais, mas sim sucessos existenciais.

# 3. Capacidade e Valor

Mais algumas considerações ajudarão a clarificar as ligações entre capacidade, valor e risco existencial.

## 3.1. Convertibilidade de Recursos em valor

Como o futuro da humanidade é potencialmente astronomicamente longo, as perdas integrais associadas com ineficácias persistentes são muito grandes. É por isso que os cenários de realização falha e ruína subsequente constituem catástrofes existenciais mesmo que eles não envolvam necessariamente extinção[20]. Pode até mesmo valer a pena cair em um estado de assistência [*welfare*] temporário para garantir uma realização a longo prazo um pouco mais eficiente do potencial da humanidade.

Para evitar a realização falha, é mais importante se concentrar em maximizar a eficiência a longo prazo do que maximizar o *output* inicial de valor no período imediatamente após a chegada à maturidade tecnológica. Isso é porque a quantidade de estrutura-valor que pode ser produzida em um tempo dado depende não apenas do nível de tecnologia, mas tam-

[20] Isso também é uma das razões pelas quais estagnação permanente é um risco existencial, por mais que a estagnação permanente também possa excluir a sobrevivência para além do tempo em que a Terra se tornar inabitável, talvez em torno de daqui um bilhão de anos, devido à crescente luminosidade solar.[(27)]

bém dos recursos físicos e outras formas de capital disponíveis no momento. Em linguajar econômico, a fronteira de possibilidades de produção da humanidade (representando as várias combinações possíveis de *outputs* que poderiam ser produzidos na economia global) depende não apenas da função de produção global (ou da "função de meta-produção"), mas também da quantidade total de todos os fatores de produção (trabalho, terra, bens capitais físicos, etc.) que estão disponíveis em algum ponto no tempo. Com tecnologia madura, a maior parte dos fatores de produção são intercambiáveis e reduzidos, em última instância, a recursos físicos básicos, mas a quantidade de energia livre disponível para uma civilização impõe duros limites ao que ela pode produzir. Já que a velocidade de colonização é limitada pela velocidade da luz, uma civilização que obtenha maturidade tecnológica começará com um dote modesto de recursos físicos (um único planeta e talvez algumas partes próximas de seu sistema solar), e levará um longo tempo – bilhões de anos – antes que uma civilização começando pudesse chegar até mesmo a 1% de sua base máxima de recursos obteníveis[21]. Por conseguinte, é a eficiência do uso em tempos tardios, em vez de imediatamente após a chegada à maturidade tecnológica, que importa mais para quanto valor é, em última instância, realizado.

Além disso, pode ocorrer que a maneira ideal de usar a maior parte do dote cósmico que a humanidade pudesse eventualmente garantir seria adiar o consumo por tanto tempo quanto possível. Ao conservar nossa energia livre acumulada até que o universo esteja mais velho e mais frio, nós talvez sejamos capazes de realizar algumas computações mais eficientemente [22]. Isso reforça o argumento de que seria um engano colocar muito peso na quantidade de valor gerada um pouco após a maturidade tecnológica, quando estivermos decidindo se algum cenário deve contar como realização falha (ou ruína subsequente). É muito mais importante acertar na configuração, no sentido de colocar a humanidade em um caminho que irá eventualmente agregar a maior parte dos recursos cósmicos obteníveis e colocamos em um uso quase otimizado. Importa menos se haverá um pequeno atraso antes disso acontecer – e um atraso de mesmo alguns milhões de anos é "curto" nesse contexto.(4)

---

[21] Uma qualificação potencialmente significativa é que o tempo para alcançar a base máxima de recursos obteníveis poderia ser menor se opositores inteligentes (como civilizações extraterrestres) emergirem e impossibilitarem a expansão cósmica.

[22] Há um custo mínimo de entropia associado com o apagamento de um bit de informação, um custo que diminui com a temperatura.

Mesmo para agentes individuais, a passagem do tempo sideral pode se tornar menos significante depois da maturidade tecnológica. Agentes que existam como processos computacionais em *hardware* computacional distribuído podem ter extensões de vida em potencial ilimitadas. O mesmo vale para agentes corporais em uma era em que tecnologias de reparo físico sejam avançadas o suficiente. A quantidade de vida disponível para esses agentes é proporcional à quantidade de recursos físicos que eles controlam. (Uma mente de *software* pode experienciar uma certa quantidade de tempo subjetivo rodando em um computador lento por um longo período de tempo sideral ou, de forma equivalente, rodando por um curto período de tempo sideral em um computador rápido.) Mesmo do ponto de vista moral que concerne a como as pessoas são afetadas, portanto, quando avaliamos se uma realização falha ocorreu, nós devemos nos concentrar não tanto em quanto valor foi criado logo depois da obtenção da maturidade tecnológica, mas sim se as condições criadas são tais que darão um bom prospecto de realizar uma grande parte integral de valor ao longo do resto de tempo de vida de nosso universo.

## 3.2. Algumas outras perspectivas éticas

Até agora nós consideramos os riscos existenciais de um ponto de vista utilitário (combinado com alguns pressupostos simplificadores). Nós podemos considerar brevemente como essa questão pode aparecer quando vista pelas lentes de outras perspectivas éticas.

Por exemplo, o filósofo Robert Adams esboça uma visão diferente sobre essas questões:

> *Eu acredito que uma base melhor para a teoria ética pode ser encontrada em uma direção inteiramente diferente – em um comprometimento com o futuro da humanidade como um vasto projeto, ou uma rede de projetos que se sobrepõem, que é, de um modo geral, compartilhada pela raça humana. A aspiração por uma sociedade melhor – mais justa, mais recompensadora e mais pacífica – é uma parte desse projeto. Também é parte dele a busca sem fim por conhecimento científico e entendimento filosófico, e o desenvolvimento da arte e de outras tradições culturais. Isso inclui as tradições culturais às quais nós pertencemos, em toda a sua história acidental e diversidade étnica. Isso também inclui o interesse nas vidas de nossos filhos e netos, e a esperança que eles serão capazes, por sua vez, de ter as vidas de seus filhos e netos como projetos. Na medida em que uma política ou prática parecer provavelmente favorável ou não favorável à realização desses projetos no futuro próximo ou dis-*

> *tante, nós temos razões para persegui-las ou evita-las. ... Continuidade é tão importante para o nosso comprometimento com o projeto do futuro da humanidade quanto o nosso comprometimento aos projetos de nossos futuros pessoais. Assim como a forma de toda a minha vida, e sua conexão com o meu presente e passado, tem um interesse que vai além do total ou da média de qualidade de vida de uma população em algum dado momento, considerado isoladamente de como se chegou a ele.*
>
> *Nós devemos, eu penso, alguma lealdade a esse projeto de futuro humano. Nós também devemos a ele um respeito que nós deveríamos mesmo se nós não fôssemos da raça humana, mas seres de outro planeta que tivessem algum entendimento sobre ela* (28: p.472-473).

Já que uma catástrofe existencial colocaria ou um fim ao projeto do futuro da humanidade ou restringiria drasticamente seu escopo de desenvolvimento, nos pareceria haver uma forte razão *prima facie* para evitá-los, na perspectiva de Adams.

Nós também notamos que uma catástrofe existencial implicaria a frustração de muitas fortes preferências, sugerindo que de uma perspectiva de satisfação de preferências, isso seria algo ruim. De forma similar, uma visão ética enfatizando políticas públicas deveria ser determinada por meio de deliberação democrática informada por todos os investidores [*stakeholders*] que favoreceriam a mitigação do risco existencial se nós supusermos, como é plausível, que a maior parte da população mundial viria a favorecer essas políticas após deliberação racional (mesmo se pessoas do futuro hipotético não estiverem inclusas como investidores). Nós também talvez tenhamos deveres custodiais de preservar a herança da humanidade passada para nós por nossos ancestrais e transmiti-la de modo seguro para nossos descendentes[23]. Nós não queremos ser o elo falho na corrente de gerações, e nós não devemos apagar ou abandonar o grande épico da civilização humana em que a humanidade esteve trabalhando por milhares de anos, quando é claro que a narrativa está muito longe de ter chegado ao seu término natural. Além disso, muitas perspectivas teológicas deploram catástrofes existenciais naturais, especialmente aquelas induzidas por atividades humanas. Se Deus criou o mundo e a espécie humana, seria de se imaginar que não seria do Seu agrado se nós decidíssemos es-

[23] Nós podemos ter responsabilidades com seres não-humanos, como animais terrestres (e, possivelmente, extraterrestres). Por mais que nós não estejamos atualmente fazendo muito para ajuda-los, nós temos a oportunidade de fazê-lo no futuro. Se oferecer ajuda para animais não humanos em sofrimento no ambiente natural é um valor importante, então, obter a maturidade tecnológica de uma maneira que falhe em produzir essa ajuda poderia contar como uma realização falha. Cf. (29, 30).

magar a Sua obra-prima (ou se, por nossa negligência ou húbris, nós permitíssemos que ela fosse irreparavelmente danificada)[24].

Nós também podemos considerar a questão de um ponto de vista menos teórico e tentar formar uma avaliação, em vez de considerar casos análogos sobre os quais nós temos uma intuição moral bem definida. Assim, por exemplo, se nós nos sentirmos confiantes de que cometer um pequeno genocídio é errado, e de que cometer um grande genocídio não é menos errado, nós podemos conjecturar que cometer omnicídio também é errado[25]. E se nós acreditamos que ter alguma razão moral para impedir catástrofes naturais que matariam um pequeno número de pessoas, e uma razão moral mais forte para impedir catástrofes naturais que matariam um número maior de pessoas, nós podemos conjecturar que nós temos uma razão moral ainda maior para impedir catástrofes que matariam toda a população humana.

Muitas perspectivas normativas diferentes concorrem, portanto, em seu suporte à mitigação do risco existencial, apesar de que o nível do mal [*badness*] envolvido em uma catástrofe existencial e a prioridade que a mitigação do risco existencial deve ter em nossa economia moral pode variar substancialmente entre diferentes teorias morais[26]. Note, entretanto, que não é, de modo algum, uma verdade *conceitual* que catástrofes existenciais são ruins ou que reduzir o risco existencial é certo. Há situações possíveis em que a ocorrência de um tipo de catástrofe existencial é benéfica – por exemplo, porque ela impede outra catástrofe existencial que, de outro modo, certamente teria ocorrido e teria sido pior.

---

[24] Poderia haver, de uma perspectiva teológica, uma categoria especial de riscos existenciais com um estatuto moral diferente: catástrofes ou apocalipses trazidos por um agente divino, talvez como punição por nossos pecados. Um crente pode julgar tal evento como, no saldo, bom. Entretanto, parece implausível que meros mortais sejam capazes de impedir Deus se Ele realmente quisesse nos esmagar, então, quaisquer contramedidas físicas que nós implementássemos contra riscos existenciais seriam presumivelmente efetivas apenas contra riscos existenciais naturais ou antropogênicos, e nós talvez não tenhamos nenhuma razão para nos contermos em nossa mitigação de riscos naturalistas por medo de frustrar os desígnios de Deus.

[25] Apesar de que o omnicídio seria pelo menos imparcial, em contraste com o genocídio, que é, frequentemente, racista ou nacionalista.

[26] Por exemplo, James Lenman argumentou que é em larga medida uma questão indiferente quando a humanidade será extinta, ao menos se não acontecer em breve.(31)

## 3.3. Risco existencial e incerteza normativa

Enquanto as duas primeiras classes de risco existencial (extinção humana e estagnação permanente) são especificadas por critérios puramente descritivos, as duas últimas (realização falha e ruína subsequente) são definidas normativamente. Isso significa que o conceito de risco existencial é, em parte, uma noção avaliativa[27].

Onde questões normativas estão envolvidas, essas questões podem se tornar contenciosas. A ética de populações, por exemplo, está carregada de problemas sobre como lidar com vários parâmetros (como tamanho da população, bem-estar médio, limites para o que conta como uma vida que vale a pena ser vivida, desigualdade, e escolhas de pessoas diferentes vs. escolhas de pessoas iguais). A avaliação de alguns cenários que envolvem transformações fundamentais na natureza humana também será provavelmente contestada.[32, 33, 34, 35] Mas nem todas questões normativas são controversas. Se concordará geralmente, por exemplo, que um futuro no qual uma pequena população humana sua para conseguir sair de uma existência miserável em um ecossistema arruinado na presença de grandes mas não utilizadas capacidades tecnológicas contaria como uma realização terrivelmente falha do potencial da humanidade e constituiria uma catástrofe existencial se não fosse revertida.

Haverá alguns tipos de riscos existenciais putativos para os quais a principal incerteza é normativa, e outros em que a principal incerteza é positiva. Com respeito à incerteza positiva ou descritiva, nós vimos anteriormente que se não se sabe que algo é objetivamente seguro, ele é perigoso, pelo menos no sentido subjetivo relevante para tomada de decisões. Nós podemos fazer um movimento paralelo quanto à incerteza normativa. Suponha que algum evento *X* reduziria a biodiversidade. Suponha (em nome da ilustração) que é sabido que *X* não teria nenhuma outra consequência significativa e que a biodiversidade reduzida não afetaria os humanos ou quaisquer outros seres moralmente consideráveis. Agora, nós podemos estar incertos quanto se a diversidade biológica tem um valor final (é valorosa "em si mesma"). Logo, nós podemos estar incertos sobre se *X* seria ruim ou não. Mas nós podemos dizer que se nós não temos cer-

[27] Nesse sentido, o conceito de risco existencial é similar a conceitos como "democracia" e "mercado de trabalho eficiente". Um buraco negro, ou uma jarra de seixos estéreis, não é uma democracia nem um mercado de trabalho eficiente, e nós podemos saber disso sem ter que fazer nenhum julgamento normativo; mas pode haver outros objetos que não podem ser classificados como instâncias ou não-instâncias desses conceitos sem tomarem lugar (pelo menos implicitamente) em alguma questão normativa.

teza se *X* seria realmente ruim (mas nós *temos* certeza de que *X* não seria bom), então *X* é ruim pelo menos no sentido subjetivo relevante para tomada de decisões. Quer dizer, nós temos razões para preferir que *X* não ocorra e talvez para tomar atitudes para impedir *X*.

Exatamente *como* se deve levar em conta incertezas morais fundamentais é uma questão aberta, mas *que* se deve fazê-lo é claro.[36] Nós podemos, então, incluir como riscos existenciais situações em que nós sabemos o que vai acontecer e podemos razoavelmente julgar que o que vai acontecer *pode* ser existencialmente ruim – mesmo quando não haveria, de fato, nada de ruim no resultado.

Nós podemos destacar uma consequência disso: Suponha que um gênio completamente confiável oferecesse à humanidade qualquer desejo que ela tivesse para o futuro. Então – mesmo se nós pudéssemos todos concordar com um tal futuro – nós ainda estaríamos encarando mais um sério risco existencial em potencial: a saber, o de escolher imprudentemente e selecionar um futuro terrivelmente falho apesar de parecer, no momento de nossa escolha, o mais desejável de todos os futuros possíveis.

## 3.4. Mantendo nossas opções vivas

Essas reflexões sobre incerteza moral sugerem uma maneira alternativa e complementar de ver riscos existenciais; elas também sugerem uma nova maneira de pensar sobre o ideal de sustentabilidade. Permita-me elaborar.

Nosso entendimento atual de axiologia pode ser confuso. Nós podemos não saber – pelo menos não com detalhes concretos – que resultados contariam como uma grande vitória para a humanidade; nós talvez não sejamos ainda nem capazes de imaginar os melhores fins para nossa jornada. Se nós estamos de fato profundamente incertos sobre nossos objetivos finais, então, nós deveríamos reconhecer que há um grande *valor de opção* em preservar – e idealmente aprimorar – a nossa habilidade de reconhecer valor e guiar o futuro de acordo com ela. Garantir que haverá uma versão futura da humanidade com grandes poderes e uma propensão para usá-los sabiamente é plausivelmente a melhor maneira disponível para nós para aumentar a probabilidade de que o futuro conterá muito valor. Para fazer isso, nós precisamos impedir quaisquer catástrofes existenciais

Por isso, nós querermos alcançar um estado em que nós tenhamos (*a*) uma inteligência muito maior, conhecimento, e um juízo mais sensato [*sound*] do que nós temos atualmente; (*b*) uma habilidade muito maior para resolver problemas de coordenação global; (*c*) capacidades tecnológicas muito maiores e recursos físico; e tal que (*d*) nossos valores e preferências não sejam corrompidos no processo de chegar lá (mas sim, se possível, aperfeiçoados). Os fatores *b* e *c* expandem o conjunto de opções disponíveis para a humanidade. O fator *a* aumenta a habilidade da humanidade para prever os resultados das opções disponíveis e entender o que cada resultado implicaria em termos de realização de valores humanos. O fator *d*, finalmente, faria mais provável que a humanidade *queira* realizar valores humanos.

Como, a partir de nossa situação atual, nós poderíamos alcançar da melhor maneira esses fins não é óbvio (fig. 5). Por mais que nós precisemos, em última instância, de mais tecnologia, compreensão [*insight*], e coordenação, não é evidente que o caminho mais curto para o objetivo seja o melhor.

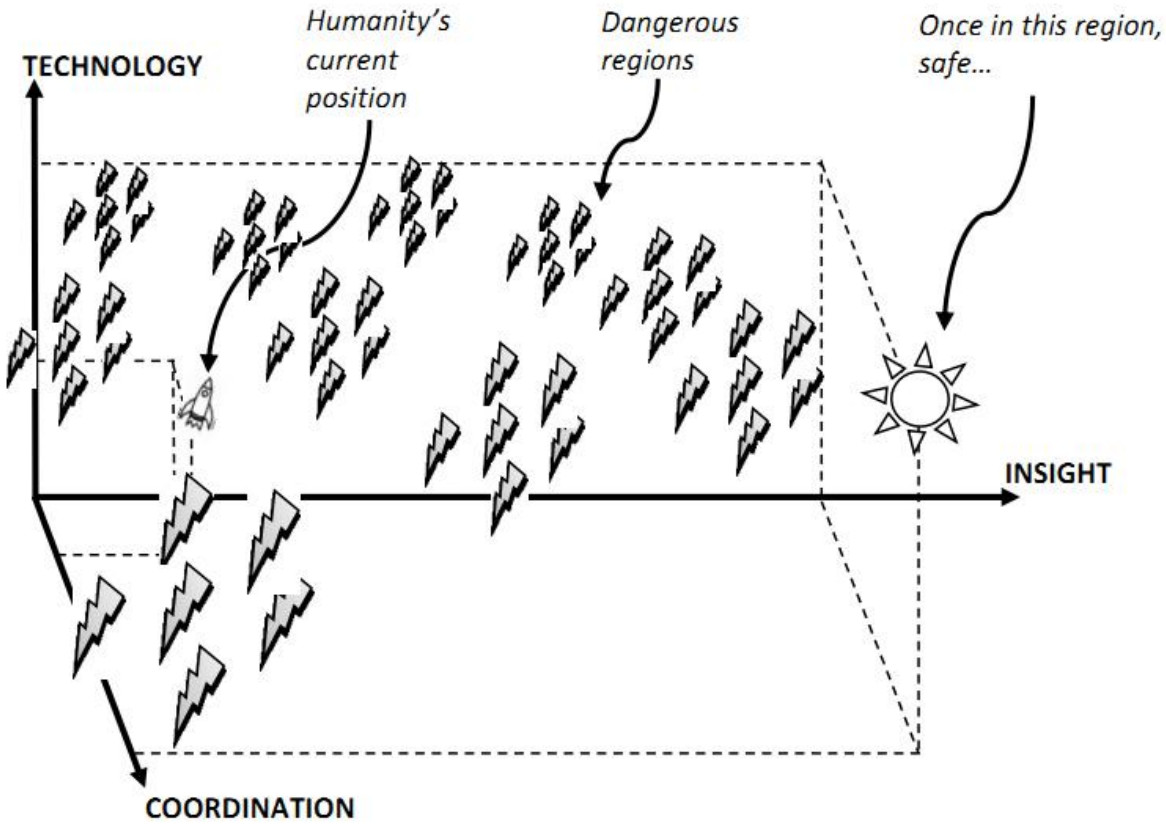


**Figura 5: O desafio de encontrar um caminho seguro.** Uma situação ideal pode ser aquela em que nós temos um nível muito alto de tecnologia, coordenação global excelente, e uma grande compreensão de como essas capacidades podem ser utilizadas. Disso não se segue que

obter qualquer quantidade adicional de tecnologia, coordenação, ou compreensão é sempre bom para nós. Talvez seja essencial que nosso crescimento ao longo de diferentes dimensões seja formado em torno de algum esquema em particular para que nosso desenvolvimento siga uma trajetória pelo espaço de estados que eventualmente alcance a região desejada.

Poderia ocorrer, por exemplo, que obter certas capacidades tecnológicas *antes* de obter compreensão e coordenação o suficiente invariavelmente signifique a ruína de uma civilização. Podemos prontamente imaginar uma classe de cenários de catástrofe existencial em que alguma tecnologia é descoberta que dá um poder imensamente destrutivo para um número grande de indivíduos. Se não houver defesa efetiva contra esse poder destrutivo, e nenhuma maneira de impedir que os indivíduos tenham acesso a ela, então a civilização não pode durar, já que uma população suficientemente grande está fadada a ter alguns indivíduos que usarão qualquer poder destrutivo disponível a eles. A descoberta da bomba atômica poderia ter resultado em algo assim, exceto pelo fato afortunado de que a construção de armas nucleares requer um ingrediente especial – material físsil no nível necessário para armas – que é raro e caro para se manufaturar. Mesmo assim, se nós tirarmos continuamente amostras da urna de descobertas tecnológicas possíveis antes de implementar meios efetivos de coordenação global, vigilância, e/ou restrição de informação potencialmente perigosa, então, nós arriscamos eventualmente tirar uma bola preta: uma intervenção fácil de fazer que causa danos extremamente disseminados contra os quais a defesa é inconcebível [28].

Por conseguinte, nós talvez não devamos buscar nos aproximar *diretamente* de algum estado que seja "sustentável", no sentido de que nós poderíamos permanecer nele por algum tempo. Em vez disso, nós deveríamos nos concentrar em entrar em uma trajetória que ofereça uma alta probabilidade de evitar catástrofes existenciais. Em outras palavras, nosso foco deveria ser maximizar as chances de que nós obteremos um dia a maturidade tecnológica de uma maneira que não seja terrível e irremediavelmente falha. Sob a condição dessa obtenção, nós temos uma boa chance de realizar nosso potencial axiológico astronômico.

Para ilustrar esse ponto, considere a seguinte analogia. Quando um foguete fica na plataforma de lançamento, ele está em um estado consideravelmente sustentável. Ele poderia permanecer na posição atual por um

[28] É claro, conseguir coordenação global suficientemente forte para monitorar continuamente a população do mundo inteiro ou censurar indefinidamente qualquer informação considerada prejudicial por alguma autoridade criaria (pelo menos na ausência de prevenções adequadas) criar seus próprios e muito significativos riscos existenciais, como riscos de estagnação permanente ou realização falha em um regime totalitário repressivo.

longo tempo, por mais que, eventualmente, fosse destruído pelo vento e pelo clima. Outro local sustentável para o foguete é o espaço, onde ele pode viajar sem peso por um longo tempo. Mas, quando o foguete está no ar, ele está em um estado insustentável e transitório: seus propulsores estão queimando e ele logo ficará sem gasolina. Retornar o foguete para um estado sustentável é desejável, mas isso não significa que *qualquer* maneira de fazer seu estado mais sustentável é desejável. Por exemplo, reduzir seu consumo de energia para que ele apenas mal consiga ficar estacionário talvez faça seu estado mais sustentável no sentido de ele poder ficar em um espaço por mais tempo; contudo, quando seu combustível terminar, o foguete irá se chocar com o chão. A melhor política para um foguete no ar é, em vez disso, manter propulsão o bastante para escapar do campo gravitacional da Terra: uma estratégia que envolve entrar em um estado *menos* sustentável (consumindo combustível mais rápido) para conseguir mais tarde o estado sustentável mais desejável. Quer dizer, em vez de tentar se aproximar de um *estado* sustentável, ele deve seguir uma *trajetória* sustentável.

A condição humana atual é, da mesma forma, um estado transitório. Como o foguete de nossa analogia, a humanidade precisa seguir uma trajetória sustentável, uma que minimize o risco de catástrofe existencial [29]. Mas, diferentemente do problema de determinar a melhor taxa de consumo de combustível, o problema de como minimizar riscos existenciais não tem nenhuma solução conhecida.

# 4. PERSPECTIVA

Nós vimos que reduzir riscos existenciais emerge como uma prioridade dominante em muitas teorias morais agregativas consequencialistas (e como uma preocupação muito importante em muitas outras teorias morais). O conceito de risco existencial pode, por conseguinte, ajudar os motivados moral ou altruisticamente a identificar ações que tenham o maior valor estimado. Em particular, dadas certas pressuposições, o problema

[29] Idealmente, se faria isso enquanto se obtivesse os meios para cometer eutanásia coletiva, no caso razoavelmente improvável de que, depois de uma deliberação coletiva longa e cuidados, nós decidíssemos que um fim rápido é preferível à existência contínua. Isso, entretanto, só pode ser uma capacidade benéfica se nós tivermos primeiramente obtido sabedoria o suficiente para não exercê-la incorretamente. Nós devemos enfatizar a necessidade de deliberação filosófica contínua e estímulo das condições que ajudariam a encontrar eventualmente a verdade sobre questões normativas centrais – assim como evitar erros irreversíveis nesse meio tempo.

de fazer a decisão certa é simplificado para o de seguir o princípio *maxi-pok*.

## 4.1. Barreiras para o pensamento e a ação

Tendo em vista esse resultado, que sugere que pode haver um valor muito alto em estudar riscos existenciais e analisar estratégias de mitigação em potencial, é chocante quão pouca atenção acadêmica essas questões receberam comparadas com outros tópicos que são menos importantes (fig. 5)[30].

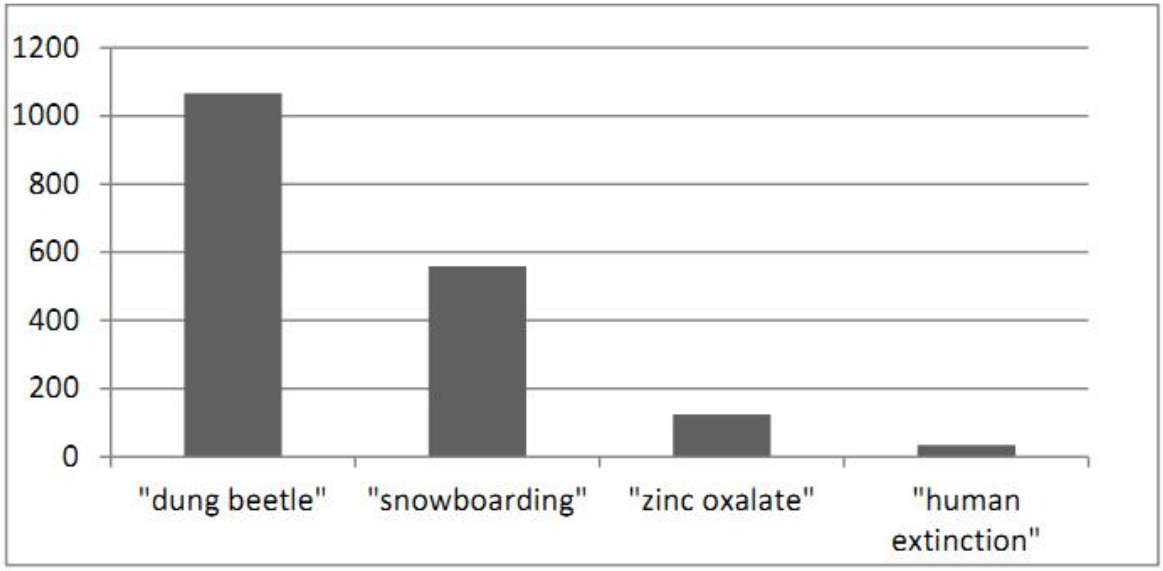


**Figure 5: Priorização acadêmica.** Número de artigos acadêmicos sobre vários tópicos (Listados em Scopus; estimado, 21 de agosto de 2010)

Muitos fatores conspiram contra o estudo e a mitigação de riscos existenciais. A pesquisa talvez seja inibida pela natureza multidisciplinar do problema, mas também por problemas epistemológicos mais profundos. Os maiores riscos existenciais não são passíveis de serem submetidos a metodologias científicas de "ligar e tocar" [*plug-and-play*]. Além disso, há problemas não resolvidos de fundamento, particularmente concernindo à teoria de seleção de observação e ética de populações, que são cruciais

[30] Abordagens acadêmicas do risco existencial em si, ou mesmo do risco de extinção humana, são raros (Cf. 1, 5, 38, 44). Entretanto, uma grande parte da literatura acadêmica trata de riscos existenciais individuais ou outras questões específicas relevantes para muitos riscos existenciais (alguns dos quais são citados nesse artigo). Além disso, alguns trabalhos recentes oferecem uma visão geral sobre riscos de catástrofe global, mas sem restringir o foco a riscos existenciais (Cf. 6, 8, 45, 46, 47, 48).

para a avaliação de riscos existenciais; e a essas dificuldades teóricas se soma fatores psicológicos que tornam difícil pensar claramente sobre questões como o fim da humanidade[31].

Se mais recursos fossem disponibilizados para pesquisa de riscos existenciais, haveria o risco de que eles fluiriam, com preponderância excessiva, para os riscos relativamente menores que é mais fácil para alguma comunidade disciplinar estabelecida estudar com métodos familiares, às custas de área de risco mais importantes – máquinas superinteligentes, nanotecnologia molecular avançada, totalitarismo, riscos relacionados à hipótese da simulação, ou avanços futuros em biologia sintética – que requereriam uma mudança mais inconveniente no foco de pesquisa. Outro desvio plausível é que a pesquisa seria dirigida a riscos de catástrofe global que envolvem poucos ou nenhum risco existencial.

A mitigação de riscos existenciais é dificultada pela falta de entendimento, mas também por um déficit de motivação. A mitigação de riscos existenciais é um bem público normal (quer dizer, não-exclusivo e desprovido de rivalidade), e a teoria econômica sugere que tais bens tendem a ser sub-fornecidos pelo mercado, já que cada produtor de segurança existencial (mesmo se o produtor por uma grande nação) poderia capturar apenas uma pequena porção do valor.[51, 52] De fato, a situação é pior do que com vários outros bens públicos globais, já que a redução de risco existencial é um bem público fortemente *transgeracional* (de fato, pan-geracional): mesmo um estado mundial poderia capturar apenas uma pequena fração de seus benefícios – aquela referente às pessoas atualmente existentes. Os quatrilhões de pessoas felizes que podem vir a existir no futuro, se nós evitarmos a catástrofe existencial, estariam dispostas a pagar à geração presente quantias astronômicas em troca de um pequeno aumento nos esforços para preservar o futuro da humanidade, mas essa troca mutuamente benéfica é infelizmente impedida por óbvias dificuldades transacionais.

Motivações morais também podem falhar em medir a magnitude do que está em jogo. O escopo da insensibilidade de nossos sentimentos morais tem muitas chances de ser especialmente pronunciado quando números muito grandes estão envolvidos:

---

[31] Questões relevantes relacionadas à observação seletiva incluem, entre outras, o argumento Carter-Leslie do dia do julgamento final, o argumento da simulação, e argumentos do "grande filtro"; veja (7, 37, 38, 39, 40, 41, 42, 43). Para algumas questões relevantes na filosofia moral, veja, por exemplo, (16, 17). Para uma revisão da literatura de vieses cognitivos na medida em que se relacionam com riscos catastróficos, veja (49).

> *Números substancialmente grandes, como 500 milhões de mortes, e especialmente cenários qualitativamente diferentes como a extinção de toda a espécie humana, parecem ativar um modo diferente de pensar – entrando em um "magistério diferente". Pessoas que nunca sonhariam em machucar uma criança escutam falar de um risco existencial e dizem, "Bom, talvez a espécie humana não mereça realmente sobreviver."* (49: p.114)

Riscos existenciais requerem uma abordagem proativa. A abordagem reativa – observar o que acontece, limitar danos, e então implementar mecanismos melhorados para reduzir a probabilidade de uma ocorrência repetida – não funciona quando não há oportunidade de aprender com a falha. Em vez disso, nós precisamos antecipar perigos emergentes, mobilizar suporte para ação contra danos hipotéticos futuros, e acertar em nossas precauções na primeira vez. Essa é uma grande ordem. Poucas instituições são capazes de operar consistentemente em tal nível de racionalidade efetiva, e tentativas de *imitar* esse comportamento proativo dentro de instituições menos perfeitas podem facilmente ter efeitos colaterais. Negociação especulativa de riscos poderia ser explorada para racionalizar ações agressivas de interesse próprio, expansão de burocracias de segurança caras e potencialmente opressivas, ou restrições de liberdades civis que mantém as sociedades livres e sãs. O resultado de falsas aproximações do ideal racional poderia facilmente ser um aumento líquido do risco existencial[32].

Desafios multidisciplinares ou epistemológicos, distrações acadêmicas e desvios, vieses cognitivos, problemas parasitários, letargia moral e insensibilidade de escopo, incompetência institucional, e exploração política de ameaças não quantificáveis são algumas das barreiras para a mitigação efetiva. A essas, nós podemos acrescentar a dificuldade de alcançar os níveis requeridos de cooperação global. Enquanto se pode combater alguns riscos existenciais unilateralmente – qualquer estado com uma indústria espacial poderia construir uma defesa global contra impactos de asteroides – outros riscos requerem um empreendimento conjunto de estados. A administração do clima global pode requerer a adesão da vasta maioria das nações industrializadas e em processo de industrialização. Evitar corridas armamentistas e o abandono de direções perigosas de pes-

[32] Um caminho possível para lidar com esse problema envolve tentar manter a quantidade total de preocupação com riscos basicamente constante, ao mesmo tempo em que se aloca um grande proporção do pote de "fichas de medo" ou "chips de preocupação" para o risco existencial. Por conseguinte, pode se advogar que na medida em que nós nos preocupamos mais com o risco existencial, nós devemos simultaneamente nos preocupar menos com riscos menores, como algumas milhares de pessoas morrendo em um ataque terrorista ou desastre natural.

quisa tecnológica pode requerer que *todos* os estados se juntem ao esforço, já que um único desertor poderia anular quaisquer benefícios da colaboração. Alguns perigos futuros podem até mesmo requerer que cada estado monitore e regule cada grupo significativo ou indivíduo dentro de seu território[33].

## 4.2. Razões para otimismo?

Uma coleção formidável de obstáculos enevoa o prospecto de uma resposta clara e efetiva aos riscos existenciais confrontando a humanidade. A fim de que a causa não seja dada como perdida, nós gostaríamos de também tomar nota de algumas considerações encorajadoras.

Nós podemos notar, primeiramente, que muitos dos conceitos-chave e ideias são muito novos[34]. Antes que os fundamentos conceituais e teóricos estivessem em seu lugar, suporte para esforços para pesquisa e mitigação de riscos existenciais não puderam ser construídos. Em muitos casos, as ideias metodológicas, tecnológicas e científicas subjacentes necessárias para estudar riscos existenciais de uma maneira significativa só foram recentemente disponibilizadas. O começo tardio começa a explicar o estado ainda primitivo da arte.

É discutivelmente apenas desde a detonação da primeira bomba atômica em 1945, e da consequente escalada nuclear durante a Guerra Fria, que quaisquer riscos existenciais naturais (quer dizer, não-sobrenaturais) surgiram – ao menos se nós contarmos apenas riscos sobre os quais seres humanos têm alguma influência[35]. A maior parte dos riscos

---

[33] Tal controle interno dentro de estados se tornará mais viável com os avanços da tecnologia de vigilância. Como foi observado, impedir que estados com essas capacidades se tornem opressivos irá oferecer os seus próprios desafios.

[34] Incluindo a própria noção de risco existencial.[(1)]

[35] Poder-se-ia argumentar que pandemias e encontros próximos com meteoros, que ocorreram repetidamente na história humana e levaram a fortes previsões de fim do mundo, deveriam contar como grandes riscos existenciais. Dada a quantidade limitada de informação disponível, poderia não ter sido desarrazoado para observadores contemporâneos atribuírem uma probabilidade significante do fim estar próximo. Cenários religiosos de fim do mundo poderiam também ser considerados; talvez não fosse desarrazoado acreditar, baseando-se na evidência então disponível, que esses riscos eram reais e, mais do que isso, que eles poderiam ser mitigados por ações como penitência, reza, oferendas sacrificiais, perseguição de bruxas ou infiéis, e assim por diante. O primeiro risco existencial distintamente científico pode ter

existenciais realmente grandes parecem ainda estar muitos anos à nossa frente. Até recentemente, portanto, pode ter havido relativamente pouca necessidade de pensar sobre riscos existenciais em geral e poucas oportunidade para a mitigação mesmo se algo desse gênero tivesse ocorrido.

A conscientização pública dos impactos globais de atividades humanas parece ter aumentado. Sistemas, processos e riscos são estudados hoje de uma perspectiva global por muitos acadêmicos – cientistas ambientais, economistas, epidemiologistas, demográficos, e outros. Problemas como mudança climática, terrorismo de fronteira, e crises financeiras internacionais dirigem a atenção para a interdependência global e ameaças para o sistema global. A ideia de risco em geral parece ter aumentado em proeminência[36]. Dados esses avanços no conhecimento, métodos e atitudes, as condições para garantir para os riscos existenciais o escrutínio que eles merecem são propícias de um modo sem precedentes.

As oportunidades também podem se proliferar. Como foi notado, alguns projetos de mitigação podem ser tomados unilateralmente, e pode-se esperar mais projetos desse gênero na medida em que o mundo se tornar mais rico. Outros projetos de mitigação requerem coordenação mais ampla; em muitos casos, coordenação global. Aqui, também, algumas das tendências parecem indicar que isso se tornará mais exequível ao longo do tempo. Há uma tendência histórica a longo prazo na direção de aumentar o escopo de integração política – de caçadores-coletores para bandos para chefaturas [*chiefdoms*], cidades estados, estados nacionais, e agora organizações multinacionais, alianças regionais, várias estruturas governamentais internacionais, e outros aspectos da globalização.(56) A extrapolação dessa tendência pode parecer indicar a criação eventual de um *singleton*.(57) Também é possível que alguns movimentos globais que surgiram na última metade do século – em particular, o movimento pela paz, o movimento ambientalista, e vários outros movimentos globais de justiça e direitos humanos – irão cada vez mais tomar para si preocupações mais generalizadas sobre riscos existenciais[37].

---

surgido com o desenvolvimento da bomba atômica. R. Oppenheimer, o líder científico do Projeto Manhattan, ordenou um estudo antes do teste Trinity para determinar se uma detonação nuclear causaria uma cadeia auto-propagadora de reações nucleares na atmosfera terrestre. O relatório resultante representa a primeira avaliação quantitativa da extinção humana.(54)

[36] Alguns sociólogos chegaram ao ponto de fixar o risco como tema central de nossa era; ver, por exemplo, (55).

[37] Muitos ativistas da paz que se opuseram à corrida armamentista nuclear durante a Guerra Fria se preocupavam explicitamente com um Apocalipse nuclear que poderia supostamente

Além disso, na medida em que a mitigação de riscos existenciais é realmente uma causa merecedora, é de se esperar que melhoras gerais na habilidade da sociedade de reconhecer e agir baseada em verdades importantes irá funilar diferencialmente recursos para a mitigação de riscos existenciais. Melhoras gerais desse tipo podem vir de muitas fontes, incluindo desenvolvimentos em técnicas educacionais e ferramentas de colaboração online, inovações institucionais como mercados de predição, avanços na ciência e filosofia, disseminação da cultura racional, e melhoramentos cognitivos biológicos.

Finalmente, é possível que a causa receberá em algum momento um impulso pela ocorrência de uma catástrofe (não-existencial) maior que delineie a precariedade da condição humana atual. Essa seria, é desnecessário dizer, a pior maneira possível para nossas mentes se concentrarem – mas, ainda assim, em uma estrutura temporal de várias décadas, deve ter atribuída a ela uma probabilidade de ocorrência não negligenciável[38].

terminar com toda a vida humana. Mais recentemente, ambientalistas soaram o alarme sobre o aquecimento global usando uma linguagem apocalíptica semelhante. Não está claro, no entanto, até que ponto a possibilidade percebida de um resultado que extinguisse a espécie foi uma das maiores forças de motivação nesses casos. Talvez a quantidade de preocupação fosse exatamente a mesma, mesmo diante de uma garantia infalível de que nenhuma catástrofe levaria à extinção humana.

[38] Eu sou grato pelos comentários e discussão com Seth Baum, Nick Beckstead, Milan Cirkovic, Sara Lippincott, Gaverick Matheny, Toby Ord, Derek Parfit, Rebecca Roache, Anders Sandberg, Carl Shulman, e outros árbitros anônimos.

## Referências


1. Bostrom N. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 2002; 9(1).
2. Sandberg A, Bostrom N. Global Catastrophic Risks Survey. Oxford: Future of Humanity Institute *Technical Report*, 2008/1, 2008. Available at: http://www.global-catastrophic-risks.com/docs/2008-1.pdf.Accessed on March 9, 2011.
3. U.K. Treasury. *Stern Review on the Economics of Climate Change*, 2006. Available at: http://www.hm-treasury.gov.uk/media/8A3/83/Chapter_2_A_-_Technical_Annex.pdf. Accessed on March 9, 2011.
4. Bostrom N. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 2003; 15(3): 308-314.
5. Matheny JG. Reducing the risk of human extinction. *Risk Analysis*, 2007; 27(5): 1335-1344.
6. Posner RA. *Catastrophe*. Oxford: Oxford University Press, 2004.
7. Cirkovic MM, Sandberg A, Bostrom N. Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis*, 2010; 30(10): 1495-1506.
8. Bostrom N, Cirkovic MM. (eds). *Global Catastrophic Risks*. Oxford: Oxford University Press, 2008.
9. Ord T, Hillerbrand R, Sandberg A. Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, 2010; 13: 191-205.
10. Parfit D. *Reasons and Persons*. Oxford: Clarendon Press, 1984.
11. Gott JR, Juric M, Schlegel D, Hoyle F, Vogeley M, Tegmark M, Bahcall N, Brinkmann J. A map of the universe. *Astrophysical Journal*, 2005; 624(2): 463-483.
12. Heyl JS. The long-term future of space travel. *Physical Review D,* 2005; 72: 1-4.
13. Cirkovic MM. Forecast for the next eon: Applied cosmology and the long-term fate of intelligent beings. *Foundations of Physics*, 2004; 34(2): 239-261.
14. Krauss LM, Starkman GD. Life, the universe, and nothing: Life and death in an ever-expanding universe. *Astrophysical Journal*, 2000; 531(1): 22-30.
15. Cirkovic MM, Radujkov M. On the maximal quantity of processed information in the physical eschatological context. *Serbian Astronomy Journal*, 2001; 163: 53-56.
16. Bostrom N. Infinite ethics, 2003, revised version 2009. Available at: http://www.nickbostrom.com/ethics/infinite.pdf. Accessed on March 15, 2011.
17. Bostrom N. Pascal's mugging. *Analysis*, 2009; 69(3): 443-445.
18. Rawls J. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971, revised edition 1999.
19. Tipler F. Extraterrestrial intelligent beings do not exist. *Royal Astronomical Society Quarterly Journal*, 1980; 21: 267-281.
20. Freitas RA. A self-reproducing interstellar probe. *Journal of the British Interplanetary Society*, 1980; 33: 251-264.

21. Moravec H. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press, 1988.

22. Freitas RA. *Nanomedicine Volume I: Basic Capabilities*. Austin, TX: Landes Bioscience, 1999.

23. Freitas RA. *Nanomedicine Volume IIA: Biocompatibility*. Austin, TX: Landes Bioscience, 2003.

24. Sandberg A, Bostrom N. Whole Brain Emulation: A Roadmap. Oxford: Future of Humanity Institute *Technical Report*, 2008, #2008-3. Available at: http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf. Accessed on March 22, 2011.

25. Bostrom N. The future of humanity. In: Olsen J-KB, Selinger E, Riis S (eds). *New Waves in Philosophy of Technology*. New York: Palgrave Macmillan; 2009. pp. 186-216.

26. Bostrom N. The future of human evolution. In: Tandy C (ed). *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*. Palo Alto, CA: Ria University Press; 2004. pp. 339-371.

27. Schroder K-P, Smith R. Distant future of the Sun and Earth revisited. *Monthly Notices of the Royal Astronomical Society*, 2008; 368(1): 155-163.

28. Adams RM. Should ethics be more impersonal? A critical notice of Derek Parfit, Reasons and Persons. *Philosophical Review*, 1989; 98(4): 439-484.

29. Pearce D. The Hedonistic Imperative, 2004. Available at: http://www.hedweb.com/welcome.htm. Accessed on March 23, 2011.

30. McMahan J. The Meat Eaters. *New York Times*, September 19, 2010. Available at: http://opinionator.blogs.nytimes.com/2010/09/19/the-meat-eaters/. Accessed on March 23, 2011.

31. Lenman J. On becoming extinct. *Pacific Philosophical Quarterly*, 2002; 83(3): 253-269.

32. Savulescu J, Bostrom N (eds). *Enhancing Humans*. Oxford: Oxford University Press, 2009.

33. Glover J. *What Sort of People Should There Be?* Harmondsworth: Penguin, 1984.

34. Kass L. *Life, Liberty and the Defense of Dignity*. San Francisco, CA: Encounter, 2002.

35. Fukuyama F. *Our Posthuman Future: Consequences of the Biotechnology Revolution*. London: Profile, 2002.

36. Bostrom N. Moral uncertainty—towards a solution? *Overcoming Bias*, 2009. Available at: http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html. Accessed on March 23, 2011.

37. Bostrom N. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge, 2002.

38. Leslie J. *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge, 1996.

39. Carter B. The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society*, 1983; A 310: 347-363.

40. Bostrom N. Are you living in a computer simulation? *Philosophical Quarterly*, 2003; 53(211): 243-255.

41. Bostrom N. Where are they? Why I hope the search for extraterrestrial life finds nothing. *MIT Technology Review*, 2008; May/June: 72-77.

42. Hanson R. The Great Filter—Are We Almost Past It? September 15, 1998. Available at: http://hanson.gmu.edu/greatfilter.html. Accessed on March 24, 2011.

43. Tegmark M, Bostrom N. How unlikely is a doomsday catastrophe? *Nature*, 2005; 438: 754.

44. Wells W. *Apocalypse When? Calculating How Long the Human Race Will Survive*. Chichester: Praxis, 2009.

45. Sunstein C. *Worst-Case Scenarios*. Cambridge, MA: Harvard University Press, 2009.

46. Homer-Dixon T. *The Upside of Down: Catastrophe, Creativity and the Renewal of Civilization*. London: Souvenir Press, 2007.

47. Diamond J. *Collapse: How Societies Choose to Fail or Survive*. London: Penguin, 2006.

48. World Economic Forum. *Global Risks, 2011*. Available at: http://riskreport.weforum.org/. Accessed on March 24, 2011.

49. Yudkowsky E. Cognitive biases potentially affecting judgment of global risks. In: Bostrom N, Cirkovic MM (eds). *Global Catastrophic Risks*. Oxford: Oxford University Press, 2008. pp. 91-119.

50. Hughes J. Millennial tendencies in responses to apocalyptic threats. In: Bostrom N, Cirkovic MM (eds). *Global Catastrophic Risks*. Oxford: Oxford University Press, 2008. pp. 73-90.

51. Kaul I. *Global Public Goods*. Oxford: Oxford University Press, 1999.

52. Feldman A. *Welfare Economics and Social Choice Theory*. Boston: Martinus Nijhoff Publishing, 1980.

53. Brin D. *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* Cambridge, MA: Perseus Books, 1998.

54. Manhattan Project. LA-602: Ignition of the Atmosphere with Nuclear Bombs, 1946. Available at: http://www.fas.org/sgp/othergov/doe/lanl/docs1/00329010.pdf. Accessed on March 24, 2011.

55. Beck U. *The World Risk Society*. Cambridge: Polity, 1999.

56. Wright R. *Nonzero: The Logic of Human Destiny*. New York: Pantheon Books, 1999.

57. Bostrom N. What is a singleton? *Linguistic and Philosophical Investigations*, 2006; 5(2): 48-54.

58. Hanson R. If uploads come first? *Extropy*, 1994; 6(2): 10-15.

59. Hanson R. Burning the cosmic commons: evolutionary strategies for interstellar colonization, 1998, Available athttp://hanson.gmu.edu/filluniv.pdf. Accessed on April 3, 2011.

60. Smil, V. *Global Catastrophes and Trends: The Next Fifty Years*. Cambridge, MA: The MIT Press, 2008.

61. Weitzman, M. L. The Extreme Uncertainty of Extreme Climate Change: An Overview and Some Implications. Harvard University, mimeo, Oct 2009.

## Notas

* Texto traduzido por Lucas Machado. Revisado por Lauro Edison.
O original pode ser lido em **http://www.existential-risk.org/concept.pdf**