Editora UFPel

Série Investigação Filosófica

Textos selecionados de

Ferramentas Formais para a Filosofia

Guilherme A. Cardoso
Sérgio R. N. Miranda
(Organizadores)

DISSERTATIO
FILOSOFIA

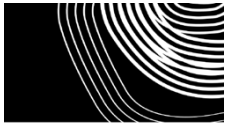
Textos Selecionados de Ferramentas Formais para a Filosofia

Série Investigação Filosófica

TEXTOS SELECIONADOS DE FERRAMENTAS FORMAIS PARA A FILOSOFIA

Guilherme A. Cardoso
Sérgio R. N. Miranda
(Organizadores)

Pelotas, 2022.



Série Investigação Filosófica

A Série Investigação Filosófica, uma iniciativa do ***Núcleo de Ensino e Pesquisa em Filosofia*** do Departamento de Filosofia da UFPel e do ***Grupo de Pesquisa Investigação Filosófica*** do Departamento de Filosofia da UNIFAP, sob o selo editorial do NEPFil online e da Editora da Universidade Federal de Pelotas, tem por objetivo precípuo a publicação da tradução para a língua portuguesa de textos selecionados a partir de diversas plataformas internacionalmente reconhecidas, tal como a *Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/), por exemplo. O objetivo geral da série é disponibilizar materiais bibliográficos relevantes tanto para a utilização enquanto material didático quanto para a própria investigação filosófica.

ORGANIZADORES DO VOLUME
Guilherme A. Cardoso (UFOP)
Sérgio Ricardo N. Miranda (UFOP)

TRADUTORES E REVISORES
Arthur de Castro Machado (UFMG)
Débora de Oliveira Silva (UNICAMP)
Guilherme A. Cardoso (UFOP)
Hulian Ferreira de Araújo (UFMG)
Sérgio R. N. Miranda (UFOP))
Wladimir Vieira (UFF)

CRÉDITO DA IMAGEM DE CAPA
CHIRICO, Giorgio de. La nostalgia dell'infinito (1911-1913). Museu de Arte Moderna de Nova York (Domínio Público).

GRUPO DE PESQUISA INVESTIGAÇÃO FILOSÓFICA (UNIFAP/CNPq)

O Grupo de Pesquisa Investigação Filosófica (DPG/CNPq) foi constituído por pesquisadores que se interessam pela investigação filosófica nas mais diversas áreas de interesse filosófico. O grupo foi fundado em 2010, como grupo independente, e se oficializou como grupo de pesquisa da Universidade Federal do Amapá em 2019.

MEMBROS PERMANENTES DO GRUPO

Aluízio de Araújo Couto Júnior
Bruno Aislã Gonçalves dos Santos
Cesar Augusto Mathias de Alencar
Daniel Schiochett
Daniela Moura Soares
Everton Miguel Puhl Maciel
Guilherme da Costa Assunção Cecílio
Kherian Galvão Cesar Gracher
Luiz Helvécio Marques Segundo
Paulo Roberto Moraes de Mendonça
Pedro Merlussi
Rafael César Pitt
Rafael Martins
Renata Ramos da Silva
Rodrigo Alexandre de Figueiredo
Rodrigo Reis Lastra Cid
Sagid Salles
Tiago Luís Teixeira de Oliveira

Universidade Federal de Pelotas
Departamento de Filosofia
Núcleo de Ensino e Pesquisa em Filosofia
Editora da Universidade Federal de Pelotas

Universidade Federal do Amapá
Departamento de Filosofia
Grupo de Pesquisa Investigação Filosófica

NEPFil online
Rua Alberto Rosa, 154 – CEP 96010-770 – Pelotas/RS

Primeira publicação em 2022 por NEPFil online e Editora da UFPel.

Dados Internacionais de Catalogação

N123 Textos selecionados de ferramentas formais para a filosofia.
[recurso eletrônico] Organizadores: Guilherme Araújo Cardoso, Sérgio Ricardo Neves de Miranda – Pelotas: NEPFIL Online, 2022.

427p. - (Série Investigação Filosófica).

Modo de acesso: Internet
<wp.ufpel.edu.br/nepfil>
ISBN: 978-85-60696-06-2

1. Filosofia. 2. Matemática I. Cardoso, Guilherme Araújo. II. Miranda, Sérgio Ricardo Neves.

COD 100

Para maiores informações, visite o site wp.ufpel.edu.br/nepfil

Sumário

Sobre a série Investigação Filosófica

A *Série Investigação Filosófica* é uma série de livros de traduções de verbetes da Enciclopédia de Filosofia da Stanford (*Stanford Encyclopedia of Philosophy*), que intenciona servir tanto como material didático para os professores das diferentes subáreas e níveis da Filosofia quanto como material de estudo para a pesquisa e para concursos da área. Nós, professores, sabemos o quão difícil é encontrar bom material em português para indicarmos. E há uma certa deficiência na graduação brasileira de filosofia, principalmente em localizações menos favorecidas, com relação ao conhecimento de outras línguas, como o inglês e o francês. Tentamos, então, suprir essa deficiência, ao introduzirmos essas traduções ao público de língua portuguesa, sem nenhuma finalidade comercial e meramente pela glória da filosofia.

Essas traduções foram todas realizadas por filósofos ou por estudantes de filosofia supervisionados e revisadas por especialistas na área. Todas as traduções de verbetes da Stanford foram autorizadas pelo querido Prof. Dr. Edward Zalta, editor da Enciclopédia de Filosofia da Stanford; por isso o agradecemos imensamente. Sua disposição para ajudar brinda os países de língua portuguesa com um material filosófico de excelência, que será para sempre disponibilizado gratuitamente no site da Editora da Universidade Federal de Pelotas (Editora UFPel), dado o nosso maior princípio se fundar na ideia de conhecimento livre e a nossa maior intenção ser o desenvolvimento da filosofia em língua portuguesa e do seu ensino. Aproveitamos o ensejo para agradecer

também ao editor da Editora UFPel, na figura do Prof. Dr. Juliano do Carmo, que apoiou nosso projeto desde o início. Agradecemos também a todos os organizadores, tradutores e revisores, que participam de nosso projeto. Sem sua dedicação voluntária, nosso trabalho não teria sido possível. Esperamos, com o início desta coleção, abrir as portas para o crescimento desse projeto de tradução e trabalharmos em conjunto pelo crescimento da filosofia em português.

Prof. Dr. Rodrigo Reis Lastra Cid
Prof. Dr. Juliano Santos do Carmo (NEFIL/UFPEL)
Editores da Série Investigação Filosófica

Introdução

A atividade filosófica sempre foi marcada pelo uso do vocabulário, conceitos e teorias de outras áreas do conhecimento. Devido à sofisticação técnica e ao alto grau de especialização da produção intelectual e científica dos dias atuais, tornou-se fundamental ter à disposição bons textos introdutórios que ofereçam panoramas de diferentes áreas nas quais são desenvolvidas ferramentas úteis para o trabalho filosófico. Os textos publicados neste volume cumprem essa função, apresentando áreas do conhecimento nas quais foram desenvolvidas ferramentas que são efetivamente utilizadas por autores contemporâneos para tratar de problemas filosóficos. O uso dessas ferramentas é bastante amplo, como se pode facilmente observar consultando a produção filosófica atual em disciplinas como a epistemologia, metafísica, lógica, filosofia da matemática, filosofia da religião, ética e política, entre outras. Além disso, as teorias aqui apresentadas têm enorme apelo intelectual intrínseco e muitas possuem profundas questões conceituais em aberto, o que as tornam ainda mais atrativas para o filósofo.

O primeiro texto do volume é **A Teoria dos Conjuntos**. Escrito pelo lógico e matemático catalão Joan Bagaria, o verbete vai desde o início da teoria com Cantor até as apresentações da técnica de *forcing*, utilizada para estabelecer a independência da hipótese do contínuo em relação à teoria dos conjuntos padrão (chamada de ZFC), e do desenvolvimento do programa de Gödel de construir teorias fortes o bastante para estabelecer hipóteses que não são decidíveis em ZFC. Entre esses extremos, Bagaria apresenta a hierarquia

cumulativa de conjuntos, as teorias dos ordinais e cardinais transfinitos, a questão da fundamentação da matemática e o fenômeno da incompletude, o universo construtivo de Gödel e a própria hipótese do contínuo. Da sua metade em diante, o texto de Bagaria é muito exigente. Para o leitor iniciante, recomendamos, antes de mais nada, a leitura dos dois complementos publicados na sequência do texto principal: o primeiro é **A Teoria Básica dos Conjuntos**, no qual Bagaria apresenta as noções elementares da teoria, como as de conjunto, relações, função, números cardinal e ordinal, entre outras; o segundo é **A Teoria dos Conjuntos de Zermelo e Frankel (ZF)**, no qual são apresentadas as versões formalizadas dos axiomas dessa teoria. A leitura desses dois complementos é suficiente para o estudante de cursos introdutórios de lógica e para acompanhar boa parte das discussões filosóficas nas quais aparecem noções da teoria de conjuntos.

O segundo texto trata dos dois **Teoremas de Gödel**, os mais famosos e importantes resultados da lógica no século XX. O primeiro teorema diz que há afirmações em qualquer sistema formal consistente, no qual uma certa quantidade de aritmética elementar possa ser executada, que não podem ser nem demonstradas e nem refutadas nesse sistema. O segundo afirma que um sistema formal consistente, no qual uma certa quantidade de aritmética elementar possa ser executada, não pode demonstrar a sua própria consistência. Primeiramente, o filósofo finlandês Panu Raatikainen explica no seu artigo os conceitos básicos para a compreensão desses teoremas, como “sistema formal”, “consistência” e “completude”, e apresenta as teorias aritméticas de Robinson e de Peano. Desenvolve, então, os elementos básicos para a prova do primeiro teorema, entre os quais a numeração de Gödel e o lema da diagonalização. O mesmo procedimento é adotado na apresentação da prova do segundo teorema, destacando-se, na preparação da prova, a discussão sobre as condições de derivabilidade de Löb. Alguns resultados relacionados às provas dos teoremas de Gödel são discutidos na sequência, entre os quais vale a pena destacar o teorema de Tarski sobre a indefinibilidade da verdade. Raatikainen descreve, ainda, a história da recepção dos teoremas de Gödel e discute na última seção algumas das implicações filosóficas desses teoremas. A tradução é publicada com dois complementos muito úteis, versando o primeiro deles sobre uma técnica

engenhosa conhecida como **Numeração de Gödel**, e o segundo sobre um importante lema utilizado nas provas dos teoremas de Gödel, **O Lema da Diagonalização**.

O terceiro texto deste volume foi escrito pela filósofa da computação Liesbeth De Mol e trata das **Máquinas de Turing**. Inicialmente, De Mol apresenta o *Entscheidungsproblem*, o problema de decidir para toda sentença da linguagem da lógica de primeira ordem se ela é derivável nessa lógica ou não, e detalha os instrumentos para lidar com esse problema, as máquinas de Turing, nas versões de Turing e de Post. O *Entscheidungsproblem* é retomado na seção seguinte, juntamente com a discussão do Problema da Parada e de variações das máquinas de Turing. De Mol apresenta então algumas questões filosóficas relacionadas às máquinas de Turing e discute modelos alternativos de computabilidade. O texto finaliza com a apresentação da relevância das máquinas de Turing em discussões relacionadas à filosofia da computação e da relevância da Turing para o desenvolvimento do computador moderno e para teorias da computação.

O verbete seguinte trata de **Diagramas**. Escrito pela filósofa Sun-Joo Shin, pelo filósofo John Mumma e pelo cientista da computação Oliver Lemmon, o texto apresenta, na sua primeira seção, os diagramas como sistemas representacionais, em que os autores discutem os diagramas de Euler, Venn e Peirce. Segue-se a seção na qual há a apresentação dos diagramas enquanto sistemas formais, em que é mostrado o desenvolvimento da elaboração de Peirce dos diagramas a partir da perspectiva elaborada por Shin. A seção seguinte trata das limitações da representação e do pensamento diagramático e da sua eficácia. Os autores discutem, então, o papel dos diagramas na geometria, particularmente na geometria euclidiana, e apresentam dois sistemas diagramáticos formais. O texto termina com reflexões sobre a relação entre diagramas e cognição, diferentes sistemas diagramáticos, diagramas e representações mentais.

O **Teorema de Bayes** é o teorema matemático mais importante para os filósofos que trabalham em áreas nas quais as noções de "evidência", "confirmação" e "probabilidade" têm um papel central (áreas como lógica indutiva, epistemologia, filosofia da ciência, filosofia da estatística e filosofia da religião, entre outras). Escrito pelo filósofo e estatístico James Joyce, o texto apresenta na

primeira seção a definição de probabilidade condicional e duas formulações do teorema de Bayes, explicando os termos envolvidos em ambas as formulações e oferecendo exemplos de fácil compreensão. Em seguida, Joyce oferece formulações especiais do teorema, inicialmente na forma de razões, mas também na forma de diferenças. A primeira é em termos de Razão de Probabilidade, formulação que envolve o grau com que uma hipótese ultrapassa uma tautologia como previsão da evidência dada, a segunda em termos de Razão de Chances, formulação que envolve o grau com que uma hipótese ultrapassa a sua negação como previsora da evidência dada, e a terceira em termos do Fator de Bayes, formulação que envolve o grau com que uma hipótese ultrapassa outra hipótese concorrente como previsora da evidência dada. Segue-se uma discussão das ideias centrais ocorrentes nas teorias bayesianas da confirmação: a relatividade da confirmação, o proporcionalismo da evidência e a confirmação incremental. O texto termina com uma discussão sobre o papel do teorema de Bayes nos modelos de revisão de crença por meio de condicionalização. O texto de Joyce é publicado com um complemento essencial para a leitura do texto principal: **Exemplos, Tabelas e Esboços de Provas** apresenta os exemplos mencionados no texto principal, ilustrações da diferença entre a razão de probabilidade e a razão de chance, entre a razão de probabilidade e a diferença de probabilidade, entre razão de chances e diferença de probabilidade, além de introduzir alguns esboços de provas.

Finalmente, o último texto deste volume trata da **Teoria dos Jogos**. O texto escrito pelo filósofo da economia Don Ross é amplo e muito instrutivo para quem se dedica a investigar questões filosóficas envolvendo as interações humanas. Inicialmente, há uma boa exposição da motivação inicial da teoria e uma exposição detalhada de alguns dos seus elementos principais. Ross apresenta matrizes e árvores de jogos, o famoso dilema do prisioneiro e os conceitos de solução e equilíbrio. A seção seguinte trata de incerteza e risco em jogos, seguidas por seções sobre jogos de lance único, jogos repetidos, jogos de coordenação, raciocínio em equipe, jogos condicionais e comprometimento. Uma seção de enorme interesse para o filósofo discute a teoria evolutiva dos jogos, usada para estabelecer condições sob as quais a linguagem humana e conceitos como “justiça” e “propriedade privada” poderiam emergir. O texto termina com

uma apreciação sobre as questões ainda em aberto e as direções de investigação mais promissoras na teoria dos jogos. Esse texto é especialmente indicado para aqueles que desejam realizar trabalhos técnicos em áreas como a filosofia da economia, filosofia social e filosofia política, entre outras.

Certamente, há vários outros verbetes da **Stanford Encyclopedia of Philosophy** que deveriam ser traduzidos e figurar neste volume sobre ferramentas formais para a filosofia. Infelizmente, não houve espaço para todos eles, porém acreditamos ter realizado uma boa seleção de textos, que apresentam resultados e teorias interessantes e instrutivas que efetivamente servem para discutir filosofia e ampliar a reflexão filosófica em várias áreas importantes, sendo, portanto, indispensáveis para o estudante com a intenção de aprofundar seu conhecimento e trabalhar nessas áreas. Também esperamos que o estudo de teorias científicas e matemáticas (mesmo que de forma amadora) possa promover a autonomia e a originalidade de pensamento, a exemplo de muitos filósofos contemporâneos. Uma última observação: confiamos na qualidade das traduções aqui publicadas, mas isso não impede a ocorrência de erros. Portanto, ficaremos gratos com críticas e sugestões que possam tornar o trabalho ainda melhor. Os leitores devem se sentir à vontade para enviar essas críticas e sugestões aos editores, nos seguintes e-mails: guilherme.cardoso@ufop.edu.br e sergiornmiranda@ufop.edu.br.

Agradecemos a todos que contribuíram para esse volume, em especial ao colega Prof. Dr. Rodrigo Cid, coordenador do projeto, e aos estudantes Arthur de Castro Machado (UFMG), Débora Oliveira (UNICAMP) e Hulian Ferreira de Araújo (UFMG), que participaram das traduções. Reconhecemos também o valioso apoio da Fundação John Templeton, que financiou este trabalho.

Prof. Dr. Guilherme A. Cardoso e Prof. Dr. Sérgio R. N. Miranda
Organizadores

Teoria dos Conjuntos*

Autoria: Joan Bagaria
Tradução: Sérgio R. N. Miranda
Revisão: Guilherme A. Cardoso

A teoria dos conjuntos é uma teoria matemática sobre coleções bem definidas, denominadas **conjuntos**, de objetos, denominados **membros** ou **elementos** do conjunto. A teoria dos conjuntos pura lida exclusivamente com conjuntos, portanto os únicos conjuntos sob consideração são aqueles cujos membros são também conjuntos. A teoria dos conjuntos finitos hereditários, que são conjuntos finitos cujos elementos são conjuntos finitos, os elementos dos quais são também finitos e assim sucessivamente, é formalmente equivalente à aritmética. Desse modo, a essência da teoria dos conjuntos é o estudo dos conjuntos infinitos e ela pode, portanto, ser definida como uma teoria matemática do infinito atual - em oposição ao infinito potencial.

---

*BAGARIA, J. Set Theory, In: ZALTA, E. N. (ed.) **Stanford Encyclopedia of Philosophy.** Winter Edition. Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/set-theory/. Acesso em: 20 jan. 2022.

The following is the translation of the entry on Set Theory by Joan Bagaria in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/set-theory/. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

A noção de conjunto é tão simples que ela é comumente introduzida de modo informal e encarada como auto-evidente. No entanto, na teoria dos conjuntos, como é comum na matemática, conjuntos são definidos axiomaticamente, portanto a existência dos conjuntos e das suas propriedades básicas são postuladas por axiomas formais apropriados. Os axiomas da teoria dos conjuntos implicam a existência de um universo teórico de conjuntos tão rico que todos os objetos matemáticos podem ser construídos como conjuntos. Além disso, a linguagem formal da teoria dos conjuntos permite a formalização de todas as noções e argumentos matemáticos. Desse modo, a teoria dos conjuntos se tornou a fundamentação padrão da matemática, à medida que todo objeto matemático pode ser encarado como um conjunto e todo teorema da matemática pode ser logicamente deduzido no Cálculo de Predicados a partir dos axiomas da teoria dos conjuntos.

Ambos os aspectos da teoria dos conjuntos, como ciência matemática do infinito e como a fundamentação da matemática, têm grande importância filosófica.

1. As Origens

A teoria dos conjuntos, como uma disciplina matemática independente, começa com a obra de Georg Cantor. Podemos dizer que a teoria dos conjuntos nasceu no final de 1873, quando Cantor fez a descoberta surpreendente de que o contínuo, isto é, a reta numérica dos números reais, não é contável, o que quer dizer que os seus pontos não podem ser contados por meio dos números naturais. Desse modo, mesmo que o conjunto dos números naturais e o conjunto dos números reais sejam ambos infinitos, há mais números reais do que há números naturais, o que abre a porta para a investigação sobre os diferentes tamanhos do infinito. Consulte o verbete Early Development of Set theory[1] na SEP para uma discussão sobre a origem das ideias teóricas sobre conjuntos e o uso dessas ideias por diferentes matemáticos e filósofos anteriores e contemporâneos de Cantor.

De acordo com Cantor, dois conjuntos *A* e *B* têm o mesmo tamanho, ou a mesma **cardinalidade**, se são bijetáveis, isto é, se os elementos de *A* podem ser colocados em correspondência um-para-um com os elementos de *B*. Desse

[1] N. T.: Disponível em: https://plato.stanford.edu/entries/settheory-early/. Acesso em 20 jan. 2022.

modo, o conjunto $\mathbb{N}$ de números naturais e o conjunto $\mathbb{R}$ dos números reais têm cardinalidades diferentes. Em 1878, Cantor formulou a famosa **Hipótese do Contínuo** (HC), que afirma que todo conjunto infinito de números reais é ou contável, isto é, tem a mesma cardinalidade de $\mathbb{N}$, ou tem a mesma cardinalidade de $\mathbb{R}$. Em outros termos, há somente dois tamanhos possíveis de conjuntos infinitos de números reais. HC é o problema mais famoso da teoria dos conjuntos. O próprio Cantor dedicou muito tempo a essa hipótese, como o fizeram muitos outros matemáticos de destaque na primeira metade do século XX, entre os quais Hilbert, que listou HC como o primeiro problema em sua celebrada lista dos 23 problemas matemáticos não resolvidos apresentada em 1900 no Segundo Congresso Internacional de Matemática, realizado em Paris. As tentativas de provar HC levaram às maiores descobertas na teoria dos conjuntos, como a teoria dos conjuntos construtíveis e a técnica de *forcing*, que mostrou que HC não pode ser provada e nem refutada a partir dos axiomas usuais da teoria dos conjuntos. Até hoje HC permanece em aberto.

Logo no começo, algumas inconsistências, ou paradoxos, surgiram do uso ingênuo da noção de conjunto. Tais inconsistências surgem especificamente da suposição natural enganadora de que qualquer propriedade define um conjunto, que seria justamente o conjunto dos objetos que têm a propriedade em questão. Um exemplo disso é o famoso **Paradoxo de Russell**, também conhecido de Zermelo:

> considere a propriedade de conjuntos de não ser membro de si mesmo. Se essa propriedade define um conjunto, chame-o de *A*, então *A* é um membro de si mesmo se e somente se não é um membro de si mesmo.

Assim, algumas coleções, como a coleção de todos os conjuntos, a coleção de todos os números ordinais, ou a coleção de todos os números cardinais, não são conjuntos. Tais coleções são chamadas de **classes próprias**.

Para evitar os paradoxos e colocá-la sobre bases firmes, a teoria dos conjuntos tinha de ser axiomatizada. A primeira axiomatização foi devida a Zermelo (1908) e resultou da necessidade de esclarecer os princípios da teoria dos conjuntos subjacentes à sua prova do Princípio da Boa Ordenação de Cantor. A axiomatização de Zermelo evita o paradoxo de Russell por meio do axioma da

Separação, que é formulado como quantificando sobre propriedades de conjuntos e, portanto, é uma asserção de segunda ordem. Trabalhos complementares de Skolem e Fraenkel levaram à formalização do axioma da Separação em termos de fórmulas de primeira ordem, em vez da noção informal de propriedade, como também à introdução do axioma da Substituição, que é também formulado como um esquema de axiomas para fórmulas de primeira ordem (*vide* a próxima seção). O axioma da Substituição é necessário para um desenvolvimento adequado da teoria dos ordinais e cardinais transfinitos usando a recursão transfinita (*vide* **Seção 3**). Esse axioma é também necessário para provar a existência de conjuntos simples, como o conjunto de conjuntos finitos hereditários, isto é, conjuntos finitos cujos elementos são finitos, os elementos dos quais são finitos e assim por diante; ou para provar fatos básicos da teoria dos conjuntos, como o fato de que todo conjunto é contido em um conjunto transitivo, isto é, um conjunto que contém todos os elementos de seus elementos (para uma discussão dos pontos fracos da teoria de Zermelo, *vide* MATHIAS, 2001). Uma adição a mais, devida a von Neumann, do axioma da Fundação, levou ao sistema axiomático padrão da teoria dos conjuntos, conhecido como axiomas de Zermelo-Fraenkel mais o axioma da Escolha [*Choice*], ou ZFC.

Axiomatizações da teoria dos conjuntos como as de Neumann-Bernays-Gödel (NBG) ou de Morse-Kelley (MK) permitem também um tratamento formal das classes próprias.

2. Os Axiomas da Teoria dos Conjuntos

ZFC é um sistema de axiomas formulado na lógica de primeira ordem com identidade e com só um símbolo de relação binária $\in$ para a pertinência. Assim, escrevemos $A \in B$ para exprimir que A é um membro do conjunto B. Para mais detalhes, *vide* o complemento **A Teoria Básica dos Conjuntos**. Para uma versão formalizada dos axiomas e comentários adicionais, *vide* o complemento **A Teoria dos Conjuntos de Zermelo- Fraenkel**. Os axiomas de ZFC são informalmente apresentados abaixo.

2.1. Os Axiomas de ZFC

- **Extensionalidade**: Se dois conjuntos *A* e *B* têm os mesmos elementos, então eles são iguais.
- **Conjunto Vazio**: Há um conjunto, denotado por $\emptyset$ e chamado de **conjunto vazio**, que não tem elementos.
- **Paridade**: Dados quaisquer conjuntos *A* e *B*, há um conjunto, denotado por $\{A, B\}$, que contém *A* e *B* como os seus únicos elementos. Em particular, existe o conjunto $\{A\}$ que tem *A* como o seu único elemento.
- **Conjunto Potência**: Para todo conjunto *A*, existe um conjunto, denotado por $\mathcal{P}(A)$ e chamado de **conjunto potência** de *A*, cujos elementos são todos os subconjuntos de *A*.
- **União**: Para todo conjunto *A*, existe um conjunto, denotado por $\bigcup A$ e chamado de **união** de *A*, cujos elementos são todos os elementos dos elementos de *A*.
- **Infinito**: Existe um conjunto infinito. Em particular, existe um conjunto *Z* que contém $\emptyset$ e tal que se $A \in Z$, então $\bigcup\{A, \{A\}\} \in Z$.
- **Separação**: Para todo conjunto *A* e toda propriedade dada, há um conjunto que contém exatamente os elementos de *A* com essa propriedade. Uma **propriedade** é dada por uma fórmula φ da linguagem de primeira ordem da teoria dos conjuntos.
  Assim, Separação não é um único axioma, mas um esquema de axiomas, ou seja, uma lista infinita de axiomas, um para cada fórmula φ.
- **Substituição**: Para toda função definida dada que tem como domínio um conjunto *A*, há um conjunto cujos elementos são todos os valores da função. Substituição é também um esquema de axiomas, à medida que funções definíveis são dadas por fórmulas.
- **Fundação**: Todo conjunto não vazio *A* contém um elemento $\in$-minimal, isto é, um elemento tal que nenhum elemento de *A* pertence a ele.

Esses são os axiomas da teoria dos conjuntos de Zermelo-Fraenkel, ou ZF. Os axiomas do Conjunto Vazio e Paridade se seguem dos outros axiomas de ZF e, portanto, podem ser omitidos. Da mesma forma, Substituição implica Separação.

Finalmente, há o Axioma da Escolha (AE):

- **Escolha**: Para todo conjunto A de conjuntos não vazios e disjuntos entre si, há um conjunto que contém exatamente um elemento de cada conjunto em *A*.

O AE foi durante muito tempo um axioma controverso. Por um lado, ele é muito útil e de uso amplo na matemática. Por outro, ele tem consequências bastante contra-intuitivas, tais como o paradoxo de Banach-Tarski, que afirma que uma unidade esférica pode ser dividida finitamente em muitas partes que podem então ser rearranjadas para formar duas unidades esféricas. As objeções ao axioma surgem do fato de que ele afirma a existência de conjuntos que não podem ser explicitamente definidos. Mas a prova da sua consistência oferecida por Gödel em 1938, relativa à consistência de ZF, afastou qualquer suspeita a respeito desse axioma.

O Axioma da Escolha é equivalente, com referência a ZF, ao **Princípio da Boa Ordenação**, que afirma que todo conjunto pode ser bem ordenado, isto é, ele pode ser linearmente ordenado de tal forma que todo subconjunto não vazio tenha um elemento minimal.

Embora não seja formalmente necessário, ao lado do símbolo $\emptyset$, normalmente se usa por conveniência outros símbolos auxiliares definidos. Por exemplo, $A \subseteq B$ exprime que *A* é um **subconjunto** de *B*, isto é, todo membro de *A* é um membro de *B*. Outros símbolos são usados para denotar conjuntos obtidos pela realização de operações básicas, tais como $A \cup B$, que denota a **união** de *A* e *B*, isto é, o conjunto cujos elementos são aqueles de *A* e de *B*; ou $A \cap B$, que denota a **interseção** de *A* e de *B*, isto é, o conjunto cujos elementos são comuns a *A* e a *B*. O **par ordenado** $\langle A, B \rangle$ é definido como o conjunto $\{\{A\}, \{A, B\}\}$. Assim, dois pares ordenados $\langle A, B \rangle$ e $\langle C, D \rangle$ são iguais se e somente se $A = C$ e $B = D$. E o produto cartesiano $A \times B$ é definido como o conjunto dos pares ordenados $\langle C, D \rangle$ tais que $C \in A$ e $D \in B$. Dada a fórmula $\varphi(x, y_1, ..., y_n)$, e os conjuntos $A, B_1, ..., B_n$, pode-se formar o conjunto de todos aqueles elementos de *A* que satisfazem a fórmula $\varphi(x, B_1, ..., B_n)$. Esse conjunto é denotado por $\{a \in A | \varphi(a, B_1, ..., B_n)\}$. Em ZF, pode-se facilmente provar que todos esses conjuntos existem. Para mais discussões, *vide* o complemento **A Teoria Básica dos Conjuntos**.

3. A Teoria dos ordinais e cardinais transfinitos

Em ZFC, podemos desenvolver a teoria cantoriana dos números ordinais e cardinais transfinitos (isto é, infinitos). Seguindo a definição dada por Von Neumann no início da década de 1920, os números ordinais, ou **ordinais**, para abreviar, são obtidos a partir do conjunto vazio e a realização de duas operações: tomar o sucessor imediato e passar ao limite. Assim, o primeiro número ordinal é $\emptyset$. Dado um ordinal α, o seu **sucessor imediato**, denotado por $\alpha + 1$, é o conjunto $\alpha \cup \{\alpha\}$. E dado um conjunto não vazio X de ordinais tais que para todo $\alpha \in X$ o sucessor $\alpha + 1$ está também em X, pode-se obter o **ordinal limite** $\bigcup X$. Pode-se mostrar que todo ordinal é (estritamente) bem ordenado por $\in$, ou seja, ele é linearmente ordenado por $\in$ e não há uma sequência infinita $\in$-descendente. Além disso, todo conjunto bem ordenado é isomórfico a um único ordinal, chamado de **tipo de ordem**.

Todo ordinal é o conjunto dos seus predecessores. No entanto, a classe ON de todos os ordinais não é um conjunto. Do contrário, ON seria um ordinal maior do que todos os ordinais, o que é impossível. O primeiro ordinal infinito, que é o conjunto de todos os ordinais finitos, é denotado pela letra grega ômega (ω). Em ZFC, os ordinais finitos são identificados com os números naturais. Desse modo, $\emptyset = 0, \{\emptyset\} = 1, \{\emptyset, \{\emptyset\}\} = 2$ e assim sucessivamente, e ω é justamente o conjunto $\mathbb{N}$ dos números naturais.

Pode-se estender as operações de adição e multiplicação dos números naturais a todos os ordinais. Por exemplo, o ordinal $\alpha + \beta$ é o tipo de ordem da boa ordenação obtida pela concatenação de um conjunto bem ordenado do tipo de ordem α e um conjunto bem ordenado do tipo de ordem β. A sequência de ordinais, bem ordenada por $\in$, começa como se segue:

$$0, 1, 2, ..., n, ..., \omega, \omega+1, \omega+2, ..., \omega+\omega, ..., \omega \cdot n, ..., \omega \cdot \omega, ..., \omega^n, ..., \omega^\omega, ...$$

Os ordinais satisfazem o princípio de **indução transfinita**: suponha que *C* é a classe dos ordinais tais que sempre que *C* contiver todos os ordinais β menores do que algum ordinal α, então α estará também em *C*. Assim, a classe *C* contém todos os ordinais. Usando a indução transfinita, pode-se provar em ZFC (e aqui o axioma da Substituição se faz necessário) o importante princípio de **recursão**

transfinita, segundo o qual dada qualquer função de classe definível $G : V \longrightarrow V$, pode-se definir a função de classe $F : ON \longrightarrow V$ tal que $F(\alpha)$ é o valor da função G aplicada à função F restrita a α. Usa-se a recursão transfinita, por exemplo, para definir apropriadamente as operações aritméticas de adição, produto e exponenciação nos ordinais.

Lembre-se de que um conjunto infinito é **contável** se ele é bijetável com ω, isto é, se ele pode ser colocado em uma correspondência um-para-um com ω. Todos os ordinais apresentados acima são ou finitos ou contáveis. Mas o conjunto de todos os ordinais finitos e contáveis é também um ordinal, chamado de ω_1, e não é contável. Similarmente, o conjunto de todos os ordinais que são bijetáveis com algum ordinal menor ou igual a ω_1 é também um ordinal, chamado de ω_2, e não é bijetável com ω_1, e assim por diante.

3.1. Cardinais

Um **cardinal** é um ordinal que não é bijetável com qualquer ordinal menor. Assim, todo ordinal finito é um cardinal e $\omega, \omega_1, \omega_2$, etc. são também cardinais. Os cardinais infinitos são representados pela letra aleph ($\aleph$) do alfabeto hebráico, e a sua sequência é indexada por ordinais. Ela começa assim:

$$\aleph_0, \aleph_1, \aleph_2, ..., \aleph_\omega, \aleph_{\omega+1}, ..., \aleph_{\omega+\omega}, ...\aleph_{\omega^2}, ..., \aleph_{\omega^\omega}, ..., \aleph_{\omega_1}, ...\aleph_{\omega_2}, ...$$

Assim, $\omega = \aleph_0, \omega_1 = \aleph_1, \omega_2 = \aleph_2$, etc. Para cada cardinal há um maior e o limite de uma sequência crescente de cardinais é também um cardinal. Portanto, a classe de todos os cardinais não é um conjunto, mas uma classe própria.

Um cardinal infinito κ é chamado de **regular** se não é a união de cardinais menores do que κ. Assim, $\aleph_0$ é regular, e também o são todos os cardinais infinitos sucessores, tais como $\aleph_1$. Cardinais infinitos não regulares são chamados de **singulares**. O primeiro cardinal singular é $\aleph_\omega$, visto que é a união de muitos cardinais menores contáveis, isto é, $\aleph_\omega = \bigcup_{n<\omega} \aleph_n$.

A **cofinalidade** de um cardinal κ, denotada por $cf(\kappa)$, é o menor cardinal λ tal que κ é a união de λ-muitos ordinais menores. Assim, $cf(\aleph_\omega) = \aleph_0$.

Pelo AE (na forma do Princípio da Boa Ordenação), todo conjunto *A* pode ser bem ordenado, por isso ele é bijetável com um único cardinal, chamado de

cardinalidade de A. Dados dois cardinais κ e λ, a adição $\kappa + \lambda$ é definida como a cardinalidade do conjunto que consiste na união de quaisquer dois conjuntos disjuntos, um da cardinalidade κ e o outro de cardinalidade λ. E a multiplicação $\kappa \cdot \lambda$ é definido como a cardinalidade do produto cartesiano $\kappa \times \lambda$. As operações de adição e multiplicação de cardinais infinitos são triviais, pois se κ e λ são cardinais infinitos, então $\kappa + \lambda = \kappa \cdot \lambda = maximum\{\kappa, \lambda\}$.

Contrariamente, a exponenciação cardinal é altamente não trivial, pois mesmo o valor do mais simples exponencial infinito não trivial $2^{\aleph_0}$ não é conhecida e não se pode determinar em ZFC (*vide* abaixo). O cardinal κ^{λ} é definido como a cardinalidade do produto cartesiano de λ cópias de κ; de forma equivalente, como a cardinalidade do conjunto de todas as funções de λ em κ. O teorema de König assevera que $\kappa^{cf(\kappa)} > \kappa$, o que implica que a cofinalidde do cardinal $2^{\aleph_0}$, seja qual for esse cardinal, tem de ser incontável. Mas isso é essencialmente tudo o que ZFC pode provar sobre o valor do exponencial $2^{\aleph_0}$.

No caso da exponenciação de cardinais singulares, ZFC tem muito mais a dizer. Em 1989, Shelah provou o resultado notável que se $\aleph_\omega$ é um limite forte, ou seja, $2^{\aleph_n} < \aleph_\omega$, para todo $n < \omega$, então $2^{\aleph_\omega} < \aleph_{\omega_4}$ (*vide* SHELAH, 1994). A técnica desenvolvida por Shelah para provar esse teorema e outros teoremas similares em ZFC é chamada de **teoria pcf** (de **possiveis cofinalidades**) e teve muitas aplicações em outras áreas da matemática.

4. O Universo *V* de Todos os Conjuntos

Os axiomas de ZF além do axioma da Extensionalidade - que não precisa de justificação porque apenas declara uma propriedade definidora de conjuntos - podem ser justificados *a posteriori* pelo uso na construção de uma **hierarquia cumulativa de conjuntos**. De fato, em ZF definimos, por meio de recursão transfinita, a função de classe que atribui a cada ordinal α o conjunto V_α, dado do seguinte modo:

- $V_0 = \emptyset$
- $V_{\alpha+1} = \mathcal{P}(V_\alpha)$
- $V_\alpha = \bigcup_{\beta<\alpha} V_\beta$, sempre que α é um ordinal limite.

O axioma do Conjunto Potência é usado para se obter $V_{\alpha+1}$ de V_α. Substituição e União permitem formar V_α para um ordinal limite α . De fato, considere a função que atribui a cada $\beta < \alpha$ o conjunto V_β. Por Substituição, a coleção de todos os V_β, para $\beta < \alpha$, é um conjunto, por isso o axioma da União aplicado a esse conjunto gera V_α. O axioma do Infinito é necessário para provar a existência de ω e, portanto, da sequência transfinita de ordinais. Finalmente, o axioma da Fundação é equivalente, assumindo os demais axiomas, à afirmação de que todo conjunto pertence a algum V_α, para algum ordinal α. Assim, ZF prova que o universo da teoria dos conjuntos, denotado por V, é a união de todos os V_α, α sendo um ordinal.

A classe própria *V*, junto com a relação $\in$, satisfaz todos os axiomas de ZFC e é assim um modelo de ZFC. Ela é um modelo pretendido de ZFC e pode-se pensar em ZFC como oferecendo uma descrição de V, uma descrição, no entanto, que é altamente incompleta, como veremos adiante.

Uma propriedade importante de V é o assim chamado **Princípio de Reflexão**: para cada fórmula $\varphi(x_1, ..., x_n)$, ZFC prova que há algum V_α que a reflete, ou seja, para todo $a_1, ..., a_n \in V_\alpha$,

$\varphi(a_1, ..., a_n)$ vale em V se e somente se $\varphi(a_1, ..., a_n)$ vale em V_α.

Assim, V não pode ser caracterizado por uma sentença, visto que qualquer sentença que é verdadeira em V tem também de ser verdadeira em algum segmento inicial V_α. Em particular, ZFC não é finitamente axiomatizável, pois de outro modo ZFC provaria que, para ilimitadamente muitos ordinais α, V_α é um modelo de ZFC, contradizendo o segundo teorema da incompletude de Gödel (*vide* **Seção 5.2**).

O Princípio de reflexão encapsula a essência da teoria ZF de conjuntos, pois, como mostrado por Levy (1960), os axiomas da Extensionalidade, Separação e Fundação, juntamente com o Princípio de Reflexão, formulado como um esquema de axiomas asseverando que toda fórmula é refletida por algum conjunto que contém todos os elementos e todos os subconjuntos dos seus elementos (note que V_α é assim), é equivaente a ZF.

5. A Teoria dos Conjuntos como Fundamento da Matemática

Todo objeto matemático pode ser encarado como um conjunto. Por exemplo, os números naturais são identificados com os ordinais finitos, portanto $\mathbb{N} = \omega$. O conjunto de inteiros $\mathbb{Z}$ pode ser definido como o conjunto de classes equivalentes de pares de números naturais sob a relação de equivalência $(n, m) \equiv (n', m')$ se e somente se $n + m' = m + n'$. Identificando todo número natural n com a classe equivalente do par $(n, 0)$, pode-se estender naturalmente as operações de adição e multiplicação dos números naturais a $\mathbb{Z}$ (para mais detalhes, *vide* ENDERTON, 1977, e para uma construção diferente, *vide* LEVY, 1979). Além disso, pode-se definir os números racionais $\mathbb{Q}$ como o conjunto de classes equivalentes dos pares (n, m) de inteiros, em que $m \neq 0$, sob a relação de equivalência $(n, m) \equiv (n', m')$ se e somente se $n \cdot m' = m \cdot n'$. Novamente, as operações + e $\cdot$ em $\mathbb{Z}$ podem ser naturalmente estendidas a $\mathbb{Q}$. Além disso, a ordenação $\leq_{\mathbb{Q}}$ nos racionais é dada por: $r \leq_{\mathbb{Q}} s$ se, e somente se, existe $t \in \mathbb{Q}$ tal que $s = r + t$. Os números reais podem ser definidos como cortes de Dedekind de $\mathbb{Q}$, ou seja, um número real é dado pelo par (A, B) de conjuntos não vazios disjuntos tais que $A \cup B = \mathbb{Q}$, e $a \leq_{\mathbb{Q}} b$ para todo $a \in A$ e $b \in B$. Pode-se então estender novamente as operações + e $\cdot$ em $\mathbb{Q}$, como também a ordenação $\leq_{\mathbb{Q}}$, ao conjunto dos números reais $\mathbb{R}$.

Vale enfatizar, que não se afirma aqui que, por exemplo, os números reais **são** cortes de Dedekind de racionais, visto que eles poderiam também ser definidos por sequências de Cauchy ou de modos diferentes. O que é importante, de um ponto de vista fundacional, é que a versão conjuntista de $\mathbb{R}$, junto com as operações algébricas usuais, satisfazem os axiomas categóricos que os números reais satisfazem, a saber, aqueles de um campo ordenado completo. A questão metafísica sobre o que são realmente os números naturais é aqui irrelevante.

Estruturas algébricas podem ser igualmente encaradas como conjuntos, à medida que qualquer relação n-ária entre os elementos de um conjunto *A* pode ser encarada como um conjunto de n-uplas $(a_1, ..., a_n)$ de elementos de *A*. E qualquer função n-ária *f* em *A*, com valores em algum conjunto *B*, pode ser encarada como o conjunto de $n+1$-uplas $((a_1, ..., a_n), b)$ tal que b é o valor de *f* em $(a_1, ..., a_m)$. Assim, por exemplo, um **grupo** é uma tripla $(A, +, 0)$, em que *A* é um conjunto

não vazio, + é uma função binária em *A* que é associativa, 0 é um elemento de *A* tal que $a + 0 = 0 + a = a$, para todo $a \in A$, e para todo $a \in A$ há um elemento de *A*, denotado por $-a$, tal que $a + (-a) = (-a) + a = 0$. Da mesma forma, um **espaço topológico** é um conjunto X junto com uma topologia τ, isto é, τ é um subconjunto do $\mathcal{P}(X)$ contendo X e $\emptyset$, e fechada sob uniões arbitrárias e interseções finitas. Qualquer objeto matemático pode sempre ser encarado como um conjunto ou como uma classe própria. As propriedades de um objeto podem ser então exprimidas na linguagem da teorias dos conjuntos. Qualquer afirmação matemática pode ser formalizada na linguagem da teoria dos conjuntos, e qualquer teorema matemático pode ser derivado, usando o cálculo da lógica de primeira ordem, a partir de axiomas de ZFC, ou de alguma extensão de ZFC. É nesse sentido que a teoria dos conjuntos fornece uma fundamentação para a matemática.

O papel fundacional da teoria dos conjuntos para a matemática, embora significativo, não é a única justificativa para o seu estudo. As ideias e técnicas desenvolvidas na teoria dos conjuntos, tais como combinatória infinita, *forcing* ou a teoria dos grandes cardinais, tornaram-na uma teoria matemática profunda e interessante, digna de estudo em si mesma, e com aplicações importantes em praticamente todas as áreas da matemática.

5.1. Metamatemática

O fato notável de que virtualmente toda a matemática pode ser formalizada em ZFC torna possível um estudo matemático da própria matemática. Desse modo, a qualquer questão sobre a existência de algum objeto matemático, ou a qualquer questão sobre a demonstrabilidadede uma conjectura ou hipótese, pode ser dada uma formulação matematicamente precisa. Isso torna a **metamatemática** possível, quer dizer, torna possível o estudo matemático da própria matemática. Portanto, a questão sobre a demonstrabilidade ou indemonstrabilidade de qualquer afirmação matemática torna-se uma questão matemática razoável. Frente a um problema ou conjectura matemática em aberto, faz sentido perguntar sobre a sua demonstrabilidade ou indemonstrabilidade no sistema formal de ZFC. Infelizmente, a resposta pode ser nem demonstrável nem indemonstrável, porque ZFC, se consistente, é incompleta.

5.2. O Fenômeno da Incompletude

O teorema da completude de Gödel para a lógica de primeira ordem implica que ZFC é **consistente** - isto é, nenhuma contradição pode ser aí derivada - se e somente se ZFC tem um modelo. Um **modelo** de ZFC é um par (*M,E*), em que *M* é um conjunto não vazio e *E* é uma relação binária em *M* tal que todos os axiomas de ZFC são verdadeiros quando interpretados em (*M,E*), ou seja, quando as variáveis que aparecem nos axiomas percorrem os elementos de M e $\in$ é interpretado como *E*. Desse modo, se φ é uma sentença da linguagem da teoria dos conjuntos e é possível encontrar um modelo de ZFC no qual φ é verdadeira, então a sua negação $\neg\varphi$ não pode ser demonstrada em ZFC. Por essa razão, se é possível encontrar um modelo de φ e também um modelo de $\neg\varphi$, então φ não é demonstrável e nem indemonstrável em ZFC, e nesse caso diremos que φ é indecidível em, ou **independente** de, ZFC.

Em 1931, Gödel anunciou os seus impactantes teoremas da incompletude, que afirmam que qualquer sistema formal razoável para a matemática é necessariamente incompleto. Em particular, se ZFC é consistente, então há proposições indecidíveis em ZFC. Além disso, o segundo teorema da incompletude de Gödel implica que a sentença formal (aritmética) $CON(ZFC)$, que afirma que ZFC é consistente, mesmo que verdadeira, não pode ser demonstrada em ZFC, assim como a sua negação não pode ser demonstrada. Portanto, $CON(ZFC)$ é indecidível em ZFC.

Se ZFC é consistente, então não pode provar a existência de um modelo de ZFC, pois, de outro modo, ZFC provaria a sua própria consistência. Assim, uma prova de consistência ou de indecidibilidade de uma dada sentença φ é sempre uma prova de **consistência relativa**. Quer dizer, assume-se em primeiro lugar que ZFC é consistente e que portanto tem um modelo, e em seguida se constrói outro modelo de ZFC em que a sentença φ é verdadeira. Veremos muitos exemplos nas próximas seções.

6. A Teoria do Contínuo

Desde Cantor, até cerca de 1940, a teoria dos conjuntos se desenvolveu principalmente em torno do estudo do contínuo, isto é, da linha real $\mathbb{R}$. O principal tópico era o estudo das assim chamadas "propriedades de regularidade", como também de outras propriedades estruturais, de conjuntos simplesmente definíveis de números reais, uma área da matemática que é conhecida como **Teoria Descritiva dos Conjuntos**.

6.1. A Teoria Descritiva dos Conjuntos

A Teoria Descritiva dos Conjuntos é o estudo de propriedades e estruturas de conjuntos de números reais definíveis, de modo mais geral, de subconjuntos definíveis de $\mathbb{R}^n$ e outros **Espaços Polonêses** (i.e, espaços topológicos que são homeomórficos a um espaço métrico completo separável), tais como o **espaço de Baire** $\mathcal{N}$ de todas as funções $f : \mathbb{N} \longrightarrow \mathbb{N}$, o espaço dos números complexos, o espaço de Hilbert e os espaços separáveis de Banach. Os conjuntos mais simples de números reais são os conjuntos abertos básicos (isto é, os intervalos abertos com extremidades racionais), e seus complementos. Os conjuntos que são obtidos em um número contável de passos começando com os conjuntos abertos básicos e então aplicando as operações de tomar o complemento e formar uma união contável de conjuntos previamente obtidos são chamados de **conjuntos borelianos**. Todos os conjuntos borelianos são **regulares**, isto é, eles gozam das **propriedades de regularidade** clássicas. Um exemplo de uma propriedade de regularidade é a **medida de Lebesgue**: um conjunto de reais é Lebesgue mensurável se difere de um conjunto boreliano por um conjunto vazio, mais especificamente, por um conjunto que pode ser coberto por conjutos de intervalos abertos básicos de comprimento total arbitrariamente pequeno. Assim, trivialmente, todo conjunto boreliano é Lebesgue mensurável, porém conjuntos mais complicados do que os conjuntos borelianos podem não o ser. Outras propriedades de regularidade clássicas são a **propriedade Baire** (um conjunto de reais tem a propriedade Baire se ele difere de um conjunto aberto por um conjunto escasso, isto é, por um conjunto que é uma união de conjuntos

contáveis que não são densos em quaquer intervalo), e a **propriedade de conjunto perfeito** (um conjunto de reais tem a propriedade de conjunto perfeito se ele é ou contável ou contém um conjunto perfeito, qual seja, um conjunto fechado não vazio sem pontos isolados). Em ZFC, pode-se provar que existem conjuntos não regulares de reais, mas o AE é necessário para essa prova (SOLOVAY, 1970).

Os **conjuntos analíticos**, também denotados por Σ^1_1, são as imagens contínuas dos conjuntos borelianos. E os conjuntos *co-analíticos*, ou Π^1_1, são os complementos dos conjuntos analíticos.

Começando com os conjuntos analíticos (ou com os co-analíticos) e aplicando as operações de projeção (do produto $\mathbb{R} \times \mathcal{N}$ a $\mathbb{R}$) e complementação, obtemos os **conjuntos projetivos**. Os conjuntos projetivos formam uma hierarquia de complexidade crescente. Por exemplo, se $A \subseteq \mathbb{R} \times \mathcal{N}$ é co-analítico, então a projeção $\{x \in \mathbb{R} \mid \exists y \in \mathcal{N}((x, y) \in A)\}$ é um conjunto projetivo no nível seguinte de complexidade acima dos conjuntos co-analíticos. Esses conjuntos são chamados Σ^1_2 e os seus complementos são chamados Π^1_2.

Os conjuntos projetivos surgem muito naturalmente na prática matemática, pois se constata que um conjunto *A* de reais é projetivo se e somente se ele é definível na estrutura:

$$\mathcal{R} = (\mathbb{R}, +, \cdot, \mathbb{Z}).$$

Ou seja, há uma fórmula de primeira ordem $\varphi(x, y_1, ..., y_n)$ na linguagem para a estrutura tal que para alguns $r_1, ..., r_n \in \mathbb{R}$,

$$A = \{x \in \mathbb{R} \mid \mathcal{R} \models \varphi(x, r_1, ..., r_n)\}.$$

ZFC prova que todo conjutno analítico, e portanto todo conjunto co-analítico, é Lebesgue mensurável e tem a propriedade Baire. Ela também prova que todo conjunto analítico tem a propriedade de conjunto perfeito. Mas a propriedade de conjunto perfeito para conjuntos co-analíticos implica que o primeiro cardinal incontável, $\aleph_1$, é um grande cardinal no universo construtível *L* (**Seção 7**), precisamente um **cardinal inacessível** (**Seção 10**), o que implica que não se pode provar em ZFC que todo conjunto co-analítico tem a propriedade de conjunto perfeito.

A teoria dos conjuntos projetivos de complexidade maior do que co-analítica é completamente subdeterminada por ZFC. Por exemplo, em *L* há um conjunto Σ^1_2 que não é Lebesgue mensurável e não tem a propriedade Baire, ao passo que se o axioma de Martin é válido (**Seção 11**), todo conjunto assim tem aquelas propriedades de regularidade. No entanto, há um axioma, chamado de "Projeção de Determinação", ou PD, que é consistente com ZFC, com referência à consistência de alguns grandes cardinais (de fato, ele se segue da existência de alguns grandes cardinais), e implica que todos os conjuntos projetivos são regulares. Além disso, PD decide essencialmente todas as questões sobre os conjuntos projetivos. Consulte o verbete **Large Cardnals and Determinacy**[2] da SEP para mais detalhes.

6.2. Determinação

Uma propriedade de regularidade de conjuntos que inclui todas as outras propriedades de regularidade é a propriedade de ser **determinado**. Para simplificar, trabalhemos com o espaço de Baire $\mathcal{N}$. Lembre-se de que os elementos de $\mathcal{N}$ são funções $f : \mathbb{N} \longrightarrow \mathbb{N}$, ou seja, sequências de números naturais de comprimento ω. O espaço $\mathcal{N}$ é topologicamente equivalente (i.e, homeomórfico) ao conjunto de pontos irracionais de $\mathbb{R}$. Portanto, visto que estamos interessados nas propriedades de regularidade de subconjuntos de $\mathbb{R}$, e visto que conjuntos contáveis, como o conjunto dos racionais, são insignificantes em termos dessas propriedades, podemos muito bem trabalhar com $\mathcal{N}$ em vez de $\mathbb{R}$.

Dado $A \subseteq \mathcal{N}$, o **jogo** associado a *A*, denotado por $\mathcal{G}_A$, tem dois jogadores, I e II, que jogam alternativamente $n_i \in \mathbb{N}$: I joga n_0, então II joga n_1, então I joga n_2, e assim por diante. Desse modo, no estágio *2k*, o jogador I joga n_{2k} e no estágio *2k + 1*, o jogador II joga n_{2k+1}. Podemos visualizar uma sequência do jogo assim:

Depois de infinitamente muitos movimentos, os dois jogadores produzem uma sequência infinita $n_0, n_1, n_2, \cdots$ de números naturais. O jogador I ganha o

[2]N.T.: Disponível em: https://plato.stanford.edu/entries/large-cardinals-determinacy/. Acesso em: 20 jan. 2022.

I	n_0		n_2		n_4	$\cdots$		n_{2k}		$\cdots$		
II		n_1		n_3	$\cdots$		$\cdots$		n_{2k+1}		$\cdots$	

jogo se a sequência pertence a *A*. De outro modo, ganha o jogador II.

O jogo $\mathcal{G}_A$ é **determinado** se há uma estratégia vencedora para um dos jogadores. Uma **estratégia vencedora** para um dos jogadores, digamos, para o jogador II, é uma função σ do conjunto de sequências finitas dos números naturais em $\mathbb{N}$, tal que se o jogador joga de acordo com a função, isto é, se joga $\sigma(n_0, ..., n_{2k})$ no *k-ésimo* lance, ele sempre ganhará o jogo, não importa o que faça o outro jogador.

Dizemos que um subconjunto *A* de $\mathcal{N}$ é **determinado** se e somente se o jogo $\mathcal{G}_A$ é determinado.

Pode-se provar em ZFC - e o uso de AE é necessário - que existem conjuntos não determinados. Assim, o **Axioma da Determinação** (AD), que afirma que todos os subconjuntos de $\mathcal{N}$ são determinados, é incompatível com AE. Mas Donald Martin provou, em ZFC, que todo conjunto boreliano é determinado. Além disso, ele mostrou que se existe um grande cardinal **mensurável** (**Seção 10**), então mesmo os conjuntos analíticos são determinados. O axioma da **Determinação Projetiva** (DP) assevera que todo conjunto projetivo é determinado. Verificou-se que DP implica que todos os conjuntos projetivos de reais são regulares, e Woodin mostrou que, em certo sentido, DP resolve essencialmente todas as questões sobre os conjuntos projetivos. Além disso, DP parece ser necessário para tanto. Outro axioma, $AD^{L(\mathbb{R})}$, assevera que AD vale em $L(\mathbb{R})$, que é a menor classe transitiva que contém todos os ordinais e todos os números reais, e satisfaz os axiomas de ZF (**Seção 7**). Assim, $AD^{L(\mathbb{R})}$ implica que todo conjunto que pertence a $L(\mathbb{R})$ é regular. Portanto, visto que $L(\mathbb{R})$ contém todos os conjuntos projetivos, $AD^{L(\mathbb{R})}$ implica DP.

6.3. A Hipótese do Contínuo

A Hipótese do Contínuo (HC), formulada por Cantor em 1878, assevera que todo conjunto infinito de números reais tem a cardinalidade de $\aleph_0$ ou tem a mesma cardinalidade de $\mathbb{R}$. Assim, HC é equivalente a $2^{\aleph_0} = \aleph_1$.

Cantor provou, em 1883, que conjuntos fechado de números reais têm a propriedade de conjunto perfeito, do que se segue que todo conjunto fechado incontável de números reais tem a mesma cardinalidade de $\mathbb{R}$. Assim, HC vale para conjuntos fechados. Mais de trinta anos depois, Pavel Aleksandrov estendeu o resultado para todos os conjuntos borelianos, e Mikhail Suslin, para todos os conjuntos analíticos. Portanto, todos os conjuntos analíticos satisfazem HC. No entanto, os esforços para provar que todos os conjuntos co-analíticos satisfazem HC não seriam bem sucedidos, à medida que tal coisa não é provável em ZFC.

Em 1938, Gödel provou a consistência de HC em ZFC. Assumindo que ZF é consistente, ele construiu um modelo de ZFC, conhecido como **universo construtível**, em que vale HC. Portanto, a prova mostra que se ZF é consistente, então ZF junto com AE e HC é consistente. Por essa razão, assumindo que ZF é consistente, o AE não pode ser refutado em ZF e a HC não pode ser refutada em ZFC.

Para o estado corrente do problema, incluindo os últimos resultados de Woodin, *vide* o verbete **Continuum Hypothesis**[3] da SEP.

7. O Universo Construtível de Gödel

O universo construtível de Gödel, denotado por *L*, é definido por recursão transfinita nos ordinais, similarmente a *V*, mas, em passos sucessivos, em vez de tomar o conjunto potência de V_α para obter $V_{\alpha+1}$, toma-se apenas aqueles subconjuntos de L_α que são definíveis em L_α, usando os elementos de L_α como parâmetros. Desse modo, sendo $\mathcal{P}^{Def}(X)$ a denotação do conjunto de todos os subconjuntos de *X* que são definíveis na estrutura $(X, \in)$ por uma fórmula da linguagem da teoria dos conjuntos, e usando os elementos de X como parâmetros da definição, supomos:

- $L_0 = \emptyset$
- $L_{\alpha+1} = \mathcal{P}^{Def}(L_\alpha)$
- $L_\lambda = \bigcup_{\alpha<\lambda} L_\alpha$, sempre que λ é um ordinal limite.

[3]N.T.: Disponível em: https://plato.stanford.edu/entries/continuum-hypothesis/. Acesso em 20 jan. 2022.

Então *L* é a união de todo L_α, sendo α um ordinal, isto é, $L = \bigcup_{\alpha \in ON} L_\alpha$. Gödel mostrou que *L* satisfaz todos os axiomas de ZFC e também a HC. De fato, ele satisfaz a Hipótese Generalizada do Contínuo (HGC), qual seja, $2^{\aleph_\alpha} = \aleph_{\alpha+1}$, para todo ordinal α.

A afirmação de que $V = L$, chamada de **Axioma da Construtibilidade**, assevera que todo conjunto pertence a L. Ela vale em L, por isso é consistente com ZFC, e implica tanto AE quanto HGC.

A classe própria L, junto com a relação $\in$ restrita a L, é um **modelo interno** de ZFC, ou seja, é uma classe **transitiva** (isto é, contém todos os elementos de seus elementos) que contém todos os ordinais e satisfaz todos os axiomas de ZFC. Ela é de fato o menor modelo interno de ZFC, à medida que está contida em todos os outros modelos.

De modo mais geral, dado algum conjunto A, pode-se construir o menor modelo transitivo de ZF que contenha A e todos os ordinais de maneira similar a L, mas então começando com o fechamento transitivo de $\{A\}$, isto é, o menor conjunto transitivo que contém A, em vez de $\emptyset$. O modelo resultante, $L(A)$, não precisa ser, no entanto, um modelo do AE. Um modelo desses muito importante é $L(\mathbb{R})$, o menor modelo transitivo de ZF que contém todos os ordinais e todos os números reais.

A teoria dos conjujntos construtíveis deve muito ao trabalho de Ronald Jansen. Ele desenvolveu a teoria da **estrutura fina** de *L* e isolou alguns princípios combinatórios, tais como o diamante ($\diamondsuit$) e o quadrado ($\square$), que podem ser usados para realizar construções complicadas de objetos matemáticos incontáveis. A teoria das estruturas finas tem também um papel importante na análise de modelos maiores similares a L, tais como $L(\mathbb{R})$ ou os modelos internos para grandes cardinais (**Seção 10.1**).

8. *Forcing*

Em 1963, vinte e cinco anos depois da prova de Gödel da consistência de HC e do AE, relativa à consistência de ZF, Paul Cohen (1966) provou a consistência da negação da HC, e também da negação do AE, relativa à consistência de ZF. Assim, se ZF é consistente, então HC é indecidível em ZFC, e o AE é indecidível em

ZF. Para chegar a isso, Cohen criou uma técnica nova e extremamente poderosa, chamada de *forcing*, para expandir modelos transitivos contáveis de ZF.

Visto que o axioma $V = L$ implica AE e HC, qualquer modelo da negação de AE ou de HC tem de violar $V = L$. Desse modo, ilustremos a ideia de *forcing* no caso da construção de um modelo para a negação de $V = L$. Começamos com um modelo transitivo *M* de ZFC, que podemos assumir, sem perda de generalidade, como sendo um modelo de $V = L$. Para violar $V = L$, temos de expandir *M* acrescentando um novo conjunto *r* tal que, no modelo expandido, *r* será não construtível. Visto que todos os conjuntos herediatariamente finitos são construtíveis, visamos acrescentar um conjunto infinito de números naturais. O primeiro problema que encontramos é que *M* pode já conter todos os subconjuntos de ω. Felizmente, pelo teorema de Löwenheim-Skolem para a lógica de primeira ordem, *M* tem um submodelo elementar contável *N*. Desse modo, visto que estamos interessados apenas nas afirmações que valem em *M* e não no próprio modelo *M*, podemos também trabalhar com *N* em vez de *M*, e assim podemos assumir que o próprio *M* é contável. Portanto, visto que $\mathcal{P}(\omega)$ é incontável, há muitos subconjuntos de ω que não pertencem a *M*. A razão é que *r* pode codificar um bocado de informação, de tal forma que quando acrescentado a *M*, *M* não é mais um modelo de ZFC ou é ainda um modelo de $V = L$. Para evitar isso, temos de selecionar *r* com muito cuidado. A ideia é selecionar *r* **genérico** em relação a *M*, querendo dizer com isso que *r* será construído a partir de suas aproximações finitas de tal modo que ele não tenha qualquer propriedade que seja definível em *M* e possa ser evitada. Por exemplo, encarando *r* como uma sequência infinita de números naturais em ordem crescente, a propriedade de *r* de só conter finitamente muitos números pares pode ser evitada, porque dada qualquer aproximação finita de *r*, isto é, qualquer sequência crescente finita de números naturais, pode-se sempre estendê-la pelo acréscimo de mais números pares, de tal modo que no final da construção *r* conterá infinitamente muitos números pares; enquanto a propriedade de conter o número 7 não pode ser evitada, porque quando uma aproximação finita de *r* contiver o número 7, então ele permanecerá aí não importa como a construção de *r* siga adiante. Visto que *M* é contável, existe um tal *r* genérico. Assim, o modelo expandido $M[r]$, que inclui *M* e contém o novo conjunto *r*, é

chamado de **extensão genérica** de *M*. Visto que assumimos que *M* é um modelo transtivo de $V = L$, o modelo $M[r]$ é exatamente $L_\alpha(r)$, em que α é o supremo dos ordinais de *M*. Assim, pode-se mostrar, usando a relação de *forcing* entre aproximações finitas de *r* e fórmulas na linguagem da teoria dos conjuntos expandida com os assim chamados **nomes** para conjuntos na extensão genérica, que *M[r]* é um modelo de ZFC e *r* não é construtível em *M[r]*, portanto o axioma da construtibilidade $V = L$ falha.

De modo geral, uma extensão *forcing* de um modelo *M* é obtida pelo acréscimo a *M* de um subconjunto genérico *G* de algum conjunto parcialmente ordenado $\mathbb{P}$ que pertence a *M*. No exemplo acima, $\mathbb{P}$ seria o conjunto de todas as sequência crescentes finitas dos números naturais, vistas como aproximações finitas da sequência infinita *r*, ordenadas por $\subseteq$; e *G* seria o conjunto de todos os segmentos iniciais finitos de *r*.

No caso da prova de consistência da negação de HC, começamos com o modelo *M* e acrescentamos $\aleph_2$ novos subconjuntos de ω, de tal modo que na extensão genérica a HC falha. Nesse caso, precisamos usar uma ordenação parcial apropriada $\mathbb{P}$ de tal modo que $\aleph_2$ de *M* não seja **colapsado**, isto é, seja o mesmo que $\aleph_2$ da extensão genérica, e assim a extensão genérica *M[G]* satisfará a sentença que afirma que há $\aleph_2$ números reais.

8.1. Outras Aplicações de *Forcing*

Além da HC, muitas outras conjecturas e problemas matemáticos sobre o contínuo, e outros objetos matemáticos infinitos, mostraram-se indecidíveis em ZFC usando-se a técnica de *forcing*.

Um importante exemplo é a **Hipótese de Suslin** (HS). Cantor mostrou que todo conjunto linearmente ordenado *S* sem extremidades que seja denso (isto é, entre dois elementos diferentes de *S* há outro elemento), completo (isto é, todo subconjunto de *S* que é limitado superiormente tem um supremo), e com um subconjunto denso contável é isomórfico à reta real. Suslin conjecturou que isso é verdadeiro ainda se a exigência de conter um subconjunto denso contável para ser ***ccc*** for relaxada, isto é, toda coleção de intervalos pares disjuntos é contável. No início da década de 1970, Thomas Jech produziu um contraexemplo

consistente usando *forcing*, e Ronald Jensen mostrou que um contraexemplo existe em *L*. Na mesma época, Robert Solovay e Stanley Tennenbaum (1971) desenvolveram e usaram pela primeira vez a técnica iterada de *forcing* para produzir um modelo em que vale HS, assim mostrando a sua independência de ZFC. A fim de tornar seguro que HS vale na extensão genérica, precisa-se destruir todos os contraexemplos, mas ao destruir um contraexemplo particular pode-se inadvertidamente criar novos contraexemplos, e assim precisa-se de aplicar *forcing* mais uma vez e assim sucessivamente; de fato, precisa-se seguir adiante ao menos ω_2 vezes. Essa é a razão por que uma iteração de *forcing* é requerida.

Entre os problemas matemáticos famosos que se mostraram indecidíveis em ZFC graças à técnica de *forcing*, especialmente pelo uso da iteração de *forcing*, algumas vezes combinada com os grandes cardinais, podemos mencionar o Problema da Medida e a Conjectura de Borel na teoria da medida, a Conjectura de Kaplansky sobre as álgebras de Banach e o Problema de Whitehead na teoria dos grupos.

9. A Busca por Novos Axiomas

Como resultado de 50 anos de desenvolvimento da técnica de *forcing* e da sua aplicação a muitos problemas em aberto na matemática, há agora literalmente milhares de questões, praticamente em todas as áreas da matemática, que se mostraram independentes de ZFC. Entre elas estão incluídas quase todas as questões sobre a estrutura dos conjuntos incontáveis. Pode-se dizer que o fenômeno da indecidibilidade é impregnante, ao ponto da investigação do incontável ter se tornado quase impossível apenas em ZFC (no entanto, para exceções notáveis, *vide* SHELAH, 1994).

Isso levanta a questão sobre o valor de verdade das afirmações que são indecidíveis por ZFC. Devemos nos contentar com a ideia de que elas são indecidíveis? Afinal, faz sentido perguntar sobre o valor de verdade dessas afirmações? Há muitas reações possíveis a tal questão. Uma é a posição cética: as afirmações que são indecidíveis em ZFC não têm uma resposta definida; e elas podem mesmo ser inerentemente vagas. Outra, a resposta comum entre os

matemáticos, é a posição de Gödel: a indecidibilidade só mostra que o sistema ZFC é muito fraco para responder a essas questões e, portanto, deve-se buscar novos axiomas que, uma vez acrescentados a ZFC, poderiam respondê-las. A busca por novos axiomas é conhecida como **Programa de Gödel**. Para uma discussão filosófica minuciosa desse programa, *vide* HAUSER, 2006, e para considerações filosóficas sobre a justificação de novos axiomas para a teoria dos conjuntos, *vide* o verbete **Large Cardinal and Determinacy**[4] da SEP.

Um tema central da teoria dos conjuntos é assim a busca e classificação de novos axiomas. Esses axiomas caem atualmente em dois grandes tipos: os axiomas dos grandes cardinais e os axiomas de *forcing*.

10. Grandes Cardinais

Em ZFC não se pode provar que existe um cardinal limite regular κ, pois se κ é um tal cardinal, então L_κ é um modelo de ZFC, e assim ZFC provaria a sua própria consistência, contradizendo o segundo teorema da incompletude de Gödel. Desse modo, a existência de um cardinal limite regular tem de ser postulada como um novo axioma. Esse cardinal é chamado de **fracamente inacessível**. Se, além disso, κ é um limite forte, isto é, $2^\lambda < \kappa$, para todo cardinal $\lambda < \kappa$, então κ é chamado de **fortemente inacessível**. Um cardinal κ é fortemente inacessível se e somente se ele é regular e V_κ é um modelo de ZFC. Se HGC vale, então todo cardinal fracamente inacessível é fortemente inacessível.

Grandes cardinais são cardinais incontáveis satisfazendo algumas propriedades que os tornam muito grandes; a existência desses cardinais não pode ser provada em ZFC. O primeiro cardinal fracamente inacessível é exatamente o menor de todos os grandes cardinais. Além dos cardinais inacessíveis há uma variedade rica e complexa de grandes cardinais, que formam uma hierarquia em termos de força de consistência, e em muitos casos também em termos de implicações óbvias. Para mais detalhes, *vide* o verbete

[4] N.T.: Disponível em: https://plato.stanford.edu/entries/large-cardinals-determinacy/. Acesso em 20 jan. 2022

Independence and Large Cardinals[5]da SEP.

Para formular a próxima noção de grande cardinal mais forte, vamos dizer que um subconjunto *C* de um cardinal infinito κ é **fechado** se todo limite de elementos de *C* está também em *C*; e é **ilimitado** se para todo $\alpha < \kappa$ existe $\beta \in C$ maior do que α. Por exemplo, o conjunto de ordinais limite menores que κ é fechado e ilimitado. Assim, um subconjunto *S* de κ é chamado de **estacionário** se intersecta todo conjunto fechado ilimitado de κ. Se κ é regular e incontável, então o conjunto de todos os ordinais menores do que κ de cofinalidade ω é um exemplo de conjunto estacionário. Desse modo, o primeiro cardinal Mahlo é muito maior que o primeiro cardinal fortemente inacessível, à medida que há muitos κ cardinais fortemente inacessíveis menores que κ.

Noções de grandes cardinais mais fortes surgem de considerações sobre as propriedades fortes de reflexão. Relembre que o Princípio de Reflexão (**Seção 4**), que é provável em ZFC, assevera que toda sentença **verdadeira** (isto é, toda sentença que vale em *V*) é verdadeira em alguma V_α. Uma fortificação desse princípio para sentenças de segunda ordem gera grandes cardinais. Por exemplo, κ é fortemente inacessível se e somente se toda sentença Σ^1_1 (isto é, uma sentença existêncial de segunda ordem na linguagem da teoria dos conjuntos com um símbolo de predicado adicional) verdadeira em alguma estrutura da forma $(V_\kappa, \in, A)$, em que $A \subseteq V_\kappa$, é verdadeira em alguma $(V_\alpha, \in, A \cap V_\alpha)$, com $\alpha < \kappa$. O mesmo tipo de raciocínio, mas agora para sentenças Π^1_1 (isto é, sentenças universais de segunda ordem), gera uma propriedade de κ de grande cardinal mais forte, chamada de **compacidade fraca**. Todo cardinal κ fracamente compacto é Mahlo, e o conjunto de cardinais Mahlo menores do que κ é estacionário. Ao permitir a reflexão para sentenças mais complexas de segunda ordem, ou mesmo de ordens superiores, obtemos noções de grande cardinal mais forte do que a compacidade fraca.

O mais famoso grande cardinal, chamado de **mensurável**, foi descoberto por Stanislaw Ulam em 1930 como um resultado da sua solução do Problema da Medida. Uma **medida** de dois valores, ou **ultrafiltro**, em um cardinal κ é um subconjunto *U* do $\mathcal{P}(\kappa)$ que tem as seguintes propriedades: (i) a

[5]N.T.: Disponível em: https://plato.stanford.edu/entries/independence-large-cardinals/. Acesso em 20 jan. 2022.

interseção de quaisquer dois elementos de *U* está em *U*; (ii) se $X \in U$ e Y é um subconjunto de κ tal que $X \subseteq Y$, então $Y \in U$; e (iii) para todo $X \subseteq \kappa$, ou $X \in U$ ou $\kappa - X \in U$, mas não ambos. Uma medida U é chamada de κ-completa se toda interseção de menos que κ elementos de U está também em U. Uma medida é chamada de **não principal** se não há $\alpha < \kappa$ que pertence a todos os elementos de *U*. Um cardinal κ é chamado de **mensurável** se existe uma medida de κ que é κ-completa e não principal.

Cardinais mensuráveis podem ser caracterizados por imersões elementares do universo *V* em alguma classe transtiva *M*. Que uma tal imersão $j : V \longrightarrow M$ seja **elementar** significa que *j* preserva a verdade, ou seja, para toda fórmula $\varphi(x_1, ..., x_n)$ da linguagem da teoria dos conjuntos, e todo $a_1, ..., a_n$, a sentença $\varphi(a_1, ..., a_n)$ vale em *V* se e somente se $\varphi(j(a_1)..., j(a_n))$ vale em *M*. Constatou-se que um cardinal κ é mensurável se e somente se existe uma imersão elementar $j : V \longrightarrow M$, com *M* transitivo, de tal modo que κ é o primeiro ordinal movido por *j*, isto é, o primeiro ordinal tal que $j(\kappa) \neq \kappa$. Dizemos que κ é o **ponto crítico** de *j*, e escrevemos $crit(j) = \kappa$. A imersão *j* é definível de uma medida não principal κ-completa, usando a assim chamada construção *ultrapower*. Contrariamente, se $j : V \longrightarrow M$ é uma imersão elementar, com *M* transitivo e $\kappa = crit(j)$, então o conjunto $U = \{X \subseteq \kappa \mid \kappa \in j(X)\}$ é um ultrafiltro não principal κ-completo de κ. Uma medida *U* obtida desse modo a partir de *j* é chamada **normal**.

Todo cardinal mensurável κ é fracamente compacto, e há muitos cardinais fracamente compactos menores que κ. De fato, abaixo de κ há muitos cardinais que são **totalmente indescritíveis**, isto é, eles refletem todas as sentenças, de qualquer complexidade, e em qualquer linguagem de ordem superior.

Se κ é mensurável e $j : V \longrightarrow M$ é dado pela construção *ultrapower*, então $V_\kappa \subseteq M$, e toda sequência de comprimento menor ou igual a κ de elementos de *M* pertence a *M*. Assim, *M* é bastante similar a *V*, mas não pode ser o próprio *V*. De fato, um famoso teorema de Kenneth Kunen mostra que não pode haver qualquer imersão elementar $j : V \longrightarrow V$, outra que uma imersão trivial, isto é, a identidade. Todas as provas conhecidas desse resultado usam o Axioma da Escolha e é uma questão excepcionalmente importante se o axioma é mesmo necessário. O Teorema de Kunen abre a porta para formular noções de grande

cardinal mais fortes que a mensurabilidade por meio da exigência de que *M* seja mais próximo de *V*.

Por exemplo, κ é chamado de **forte** se para todo ordinal α existe uma imersão elementar $j : V \longrightarrow M$, para algum *M* transitivo, tal que $\kappa = crit(j)$ e $V_\alpha \subseteq M$.

Outra noção de grande cardinal importante, e muito mais forte, é a supercompactação. Um cardinal κ é **supercompacto** se para todo α existe uma imersão elementar: $j = V \longrightarrow M$, com M transitivo e ponto crítico κ, de tal modo que $j(\kappa) > \alpha$ e toda sequência de elementos de *M* de comprimento α pertence a *M*.

Os cardinais de Woodin ficam entre os fortes e os supercompactos. Todo cardinal supercompacto é Woodin, e se δ é Woodin, então V_δ é um modelo de ZFC em que há uma classe própria de cardinais fortes. Assim, enquanto um cardinal Woodin δ não precisa ser ele mesmo muito forte - o primeiro não é nem mesmo compacto -, ele implica a existência de muitos grandes cardinais em V_δ.

Além dos cardinais supercompactos, temos os cardinais **extensíveis**, os **enormes**, os **superenormes**, etc.

O teorema de Kunen sobre a inexistência de uma imersão elementar não trivial $j : V \longrightarrow V$ realmente mostra que não pode haver uma imersão elementar $j : V_{\lambda+2} \longrightarrow V_{\lambda+2}$ diferente da identidade, para qualquer λ.

As noções de grandes cardinais mais fortes que não se sabe serem inconsistentes, com referência a ZFC, são as seguintes:

- Existe uma imersão elementar $j : V_{\lambda+1} \longrightarrow V_{\lambda+1}$ diferente da identidade.
- Existe uma imersão elementar $L(V_{\lambda+1}) \longrightarrow L(V_{\lambda+1})$ diferente da identidade.

Os grandes cardinais formam uma hierarquia de crescente força de consistência. De fato, eles são o ponto de partida da hierarquia de interpretabilidade de teorias matemáticas. Para mais detalhes, *vide* o verbete **Independence and Large Cardinals**[6] da SEP. Dada qualquer sentença φ, ocorre exatamente uma de três possibilidades sobre a teoria ZFC mais φ:

[6]N.T.: Disponível em: https://plato.stanford.edu/entries/independence-large-cardinals/. Acesso em 20 jan. 2022.

- ZFC mais φ é inconsistente.
- ZFC mais φ é equiconsistente com ZFC.
- ZFC mais φ é equiconsistente com ZFC mais a existência de algum grande cardinal.

Portanto, grandes cardinais podem ser usados para provar que uma dada sentença φ não implica outra sentença ψ, com referência a ZFC, mostrando que ZFC mais ψ implica a consistência de algum grande cardinal, enquanto ZFC mais φ é consistente com a suposição da existência de um grande cardinal menor, ou só com a suposição da consistência de ZFC. Em outros termos, ψ tem uma consistência maior do que φ, com referência a ZFC. Assim, pelo segundo teorema da incompletude de Gödel, ZFC mais φ não pode provar ψ, assumindo que ZFC mais φ é consistente.

Como já foi observado, não se pode provar em ZFC que grandes cardinais existem. Mas tudo indica que a existência desses cardinais não só não pode ser refutada, mas na verdade a suposição da existência desses cardinais é um axioma muito razoável da teoria dos conjuntos. A princípio, há muita evidência para as suas consistências, especialmente para aqueles grandes cardinais para os quais é possível construir um modelo interno.

10.1. Modelos Internos dos Grandes Cardinais

Um **modelo interno** de ZFC é uma classe própria transitiva que contém todos os ordinais e satisfaz todos os axiomas de ZFC. Assim, *L* é o menor modelo interno, enquanto *V* é o maior. Alguns grandes cardinais, como o inacessível, Mahlo, ou fracamente compacto, podem existir em *L*. Quer dizer, se κ tem um dessas propriedades de grandes cardinais, então também tem a propriedade em *L*. Mas alguns grandes cardinais não podem existir em *L*. De fato, Scott (1961) mostrou que, se existe um cardinal mensurável κ em L, então $V \neq L$. É importante notar que κ pertence a *L*, visto que *L* contém todos os ordinais, mas ele não é mensurável em *L*, porque uma medida não principal κ-completa de κ não pode aí existir.

Se κ é um cardinal mensurável, então é possível construir um modelo como de *L* em que κ é mensurável assumindo uma medida não principal

κ-completa e normal *U* de κ, e procedendo como na definição de *L*, mas agora mostrando *U* como um predicado adicional. O modelo resultante, chamado de $L[U]$, é um modelo interno de ZFC no qual κ é mensurável, e de fato κ é o único cardinal mensurável. O modelo é canônico, no sentido de que qualquer outra medida normal atestando a mensurabilidade de κ produziria o mesmo modelo, e tem muitas das propriedade de *L*. Por exemplo, ele tem uma boa ordenação projetiva dos reais, e satisfaz HGC.

Construindo modelos similares como de *L* para grandes cardinais, como o forte, ou de Woodin, é bem mais difícil. Tais modelos são da forma *L[E]*, em que *E* é uma sequência de **extensores**, cada extensor sendo um sistema de medidas, que codifica as imersões elementares relevantes.

O maior modelo interno como *L* para grandes cardinais que se obteve até agora pode conter os limites de Woodin de cardinais de Woodin (NEEMAN, 2002). No entanto, construir um modelo como *L* para um cardinal supercompacto é ainda um desafio. A barreira supercompacta parece ser crucial, pois Woodin mostrou que para um tipo de modelo interno como *L* para um cardinal supercompacto, que ele chama de *Ultimate-L*, todos os cardinais mais fortes que podem existir em *V*, como extensível, enorme, I1, etc., também existiriam no modelo. A construção do *Ultimate-L* é ainda incompleta, e não é claro ainda que ela ocorrerá, pois ela se baseia sobre algumas hipóteses técnicas que precisam de ser confirmadas.

10.2. Consequências de Grandes Cardinais

A existência de grandes cardinais tem consequências dramáticas, mesmo para os pequenos conjuntos definíveis de modo simples, como os conjuntos projetivos de números reais. Por exemplo, Solovay (1970) provou, assumindo que existe um cardinal mensurável, que todos os conjuntos Σ^1_2 de reais são Legesgue mensuráveis e têm a propriedade Baire, o que não pode ser provado em ZFC apenas. Shelah e Woodin (1990) mostraram que a existência de uma classe própria de cardinais de Woodin implica que a teoria de $L(\mathbb{R})$, mesmo com números reais como parâmetros, não pode ser alterada por *forcing*, o que implica que todos os conjuntos de números reais que pertencem a $L(\mathbb{R})$ são regulares. Além disso, sob a hipótese mais fraca de grandes cardinais,

particularmente a existência de infinitamente muitos cardinais de Woodin, Martin e Steel (1989) provaram que todo conjunto projetivo de números reais é determinada, isto é, vale o exioma de DP, e por isso todos os conjuntos projetivos são regulares. Além disso, Woodin mostrou que a existência de infinitamente muitos cardinais de Woodin, mais o cardinal mensurável acima de todos eles, implica que todo conjunto de reais em $L(\mathbb{R})$ é determinado, isto é, vale o axioma $AD^{L(\mathbb{R})}$, por isso todos os conjuntos de números reais que pertencem a $L(\mathbb{R})$, e, portanto, todos os conjuntos projetivos, são regulares. Woodin mostrou, também, que os cardinais de Woodin oferecem as pressuposições ideais dos grandes cardinais provando que as duas seguintes afirmações:

1 Há infinitamente muitos cardinais de Woodin.
2 $AD^{L(\mathbb{R})}$.

são equiconsistentes, isto é, ZFC mais 1 é consistente se e somente se ZFC mais 2 é consistente. Para mais detalhes e resultados relacionados, *vide* o verbete **Large Cardinals and Determinacy**[7] da SEP.

Outra área em que os grandes cardinais desempenham um papel importante é a exponenciação de cardinais singulares. A **Hipótese do Cardinal Singular** (HCS) completamente determina o comportamento da exponenciação para cardinais singulares, com referência à exponenciação para cardinais regulares. HCS se segue de HGC e, portanto, vale em *L*. Uma consequência de HCS é que, se $2^{\aleph_n} < \aleph_\omega$, para todo n finito, então $2^{\aleph_\omega} = \aleph_{\omega+1}$. Portanto, se a HCS vale para cardinais menores do que $\aleph_\omega$, então também vale para $\aleph_\omega$. HCS vale acima do primeiro cardinal supercompacto (Solovay), mas Magiro (1977) mostrou que, surpreendentemente, assumindo a existência de grandes cardinais, é possível construir um modelo de ZFC que em que HGC primeiramente falha em $\aleph_\omega$ e, por isso, HCS também falha. Grandes cardinais mais fortes que mensuráveis são realmente necessários para tanto. Contrariamente, no entanto, ZFC sozinha é suficiente para provar que se HCS vale para todos os cardinais menores que $\aleph_{\omega 1}$, então também vale para $\aleph_{\omega 1}$. Além disso, se a HCS vale para todos os cardinais singulares de cofinalidade contável, então vale para todos os

[7] N.T.: Disponível em: https://plato.stanford.edu/entries/large-cardinals-determinacy/. Acesso em: 20 jan. 2022

cardinais singulares (Silver).

11. Axiomas de *Forcing*

Os axiomas de *forcing* são axiomas da teorias dos conjuntos que asseveram que acertas afirmações existenciais são absolutas entre o universo *V* de todos os conjuntos e as suas extensões *forcing* (ideais), ou seja, algumas afirmações existenciais que valem em algumas extensões *forcing* de *V* são já verdadeiras em *V*. O primeiro axioma de *forcing* foi formulado por Donald Martin, estimulado pela prova de Solovay-Tennenbaum da consistência da Hipótese de Suslim, e é bem conhecido como **Axioma de Martin** (AM). Antes de apresentá-lo, notemos que uma **ordem parcial** é um conjunto não vazio *P* junto com uma relação binária $\leq$ em *P* que é reflexiva e transitiva. Dois elementos, *p* e *q*, de *P* são chamados de **compatíveis** se existe um $r \in P$ tal que $r \leq p$ e $r \leq q$. Uma **anti-cadeia** de *P* é um subconjunto de *P* cujos elementos são pareadamente incompatíveis. Uma ordem parcial *P* é chamada de **ccc** se toda anti-cadeia de *P* é contável. Um subconjunto não vazio *G* de *P* é chamado de **filtro** se (i) todo pareamento de elementos de *G* são compatíveis, e (ii) se $p \in G$ e $p \leq q$, então também $q \in G$. Finalmente, um subconjunto D de P é chamado de **denso** se para todo $p \in P$ existe $q \in D$ tal que $q \leq p$.

AM assevera o seguinte:

> Para toda ordem parcial **ccc** P e todo conjunto $\{D_\alpha \mid \alpha < \omega_1\}$ de subconjuntos densos de *P*, existe um filtro $G \subseteq P$ que é **genérico** para o conjunto, ou seja, $G \cap D_\alpha \neq \emptyset$, para todo $\alpha < \omega_1$.

Martin e Solovay (1970) provaram que AM é consistente com ZFC usando a iteração de *forcing* com a propriedade **ccc**. À primeira vista, AM pode não parecer um axioma, ou seja, uma asserção óbvia, ou ao menos uma asserção razoável, sobre conjuntos, mas sim uma afirmação técnica sobre as ordens parciais **ccc**. No entanto, ele parece mais natural quando exprimido em termos topológicos, pois é simplesmente uma generalização do bem conhecido Teorema da Categoria de Baire, que afirma que em cada espaço topológico compacto Hausdorff a interação de contavelmente muitos conjuntos abertos densos é não vazia. De

fato, AM é equivalente a:

> Em todo espaço topológico **ccc** Hausdorf compacto, a interseção de $\aleph_1$-muitos conjuntos abertos densos é não vazia.

AM tem muitas formulações diferentes equivalentes e tem sido usado com muito sucesso para decidir um grande números de problemas em aberto em outras áreas da matemática. Por exemplo, ele implica a Hipótese de Suslin e que todo conjunto de reais Σ^1_2 é Leguesgue mensurável e tem a propriedade Baire. Ele também implica a negação da HC e que $2^{\aleph_0}$ é um cardinal regular, mas não decide qual cardinal. Para muitas outras consequências de AM e outras formulações equivalentes, *vide* FREMLIN, 1984. A despeito disso, o status de AM como axioma da teoria dos conjuntos é ainda incerto. Talvez a formulação mais natural de AM, de um ponto de vista fundacional, é em termos de reflexão. Escrevendo Hec para o conjunto de conjuntos hereditariamente contáveis (isto é, conjuntos contáveis cujos elementos são contáveis, os elementos dos quais são também contáveis e assim por diante), AM é equivalente a:

> Para toda ordem parcial **ccc** *P*, se uma afirmação existencial sobre Hec vale em uma extensão genérica (ideal) de *V* obtida por *forcing* com *P*, então a afirmação é verdadeira, ou seja, ela vale em *V*. Em outros termos, se um conjunto tendo uma propriedade que depende somente dos conjuntos em Hec existe um alguma extensão genérica (ideal) de *V* obtida por *forcing* com uma ordem parcial ***ccc***, então um conjunto com tal propriedade já existe em *V*.

A noção de uma extensão genérica ideal de *V* pode ser tornada precisa em termos dos assim chamados modelos Boole-valorados, que fornece uma versão alternativa de *forcing*.

Axiomas de *forcing* mais fortes foram introduzidos na década de 1980, como o **Axioma de Forcing Apropriado** (AFA) de J. Baumgartner, e o mais forte **Máximo de Martin** (MM) de Foreman, Magidor e Shelah (1988), que é o axioma de *forcing* mais forte possível. Tanto AFA quanto MM são consistentes reletivamente à existência de um cardinal supercompacto. AFA assevera o mesmo que AM, mas para ordens parciais que têm uma propriedade mais fraca do que **ccc**, chamada de

adequatibilidade, introduzida por Shelah. E MM assevera o mesmo para a classe ampla de ordens parciais que, quando *forcing* é aplicadas a elas, os subconjuntos estacionários de ω_1 não são destruídos.

Axiomas de *forcing* fortes, como AFA e MM, implicam que todos os conjuntos projetivos de reais são determinados (DP), e têm muitas outras consequências fortes em combinatória infinita. Notavelmente, eles implicam que a cardinalidade do contínuo é $\aleph_2$.

Bibliografia

BAGARIA, J. "Set Theory", in: **The Princeton Companion to Mathematics**, editor Timothy GOWERS; editores associados June BARROW-GREEN e Imre LEADER. Princeton: Princeton University Press, 2008.

COHEN, P.J. **Set Theory and the Continuum Hypothesis**, Nova York: W. A. Benjamin, Inc. 1966.

ENDERTON, H.B. **Elements of Set Theory**, Nova York: Academic Press, 1977.

FERREIRÓS, J. **Labyrinth of Thought: A History of Set Theory and its Role in Modern Mathematics**, 2ª Edição Revisada, Basel: Birkhäuser, 2007.

FOREMAN, M., M. MAGIDOR, e S. SHELAH "Martin's maximum, saturated ideals and non-regular ultrafilters", Parte I, **Annals of Mathematics**, 127: 1–47,1988.

FREMLIN, D.H. "Consequences of Martin's Axiom", **Cambridge tracts in Mathematics 84**. Cambridge: Cambridge University Press, 1984.

GÖDEL, K. 1931, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," **Monatshefte für Mathematik Physik**, 38: 173–198, 1931.

GÖDEL, K. "The consistency of the axiom of choice and of the generalized continuum hypothesis", **Proceedings of the National Academy of Sciences**, U.S.A. 24: 556–557, 1938.

GÖDEL, K. **Collected Works I.** Publications 1929–1936, S. FEFERMAN et al. (eds.), Oxford: Oxford University Press, 1986.

HAUSER, K. "Gödel's program revisited, Part I: The turn to phenomenology", **Bulletin of Symbolic Logic**, 12(4): 529–590, 2006.

JECH, T. **Set theory,** 3ª Edição, Nova York: Springer, 2003.

JENSEN, R.B. “The fine structure of the constructible hierarchy”, **Annals of Mathematical Logic**, 4(3): 229–308, 1972.

KANAMORI, A. **The Higher Infinite**, 2ª Edição, Springer Monographs in Mathematics, Nova York: Springer, 2003.

KECHRIS, A.S. Classical Descriptive Set Theory, Graduate Texts in Mathematics, Nova York: Springer Verlag, 1995.

KUNEN, K. **Set Theory, An Introduction to Independence Proofs**, Amsterdam: North-Holland, 1980.

LEVY, A. “Axiom schemata of strong infinity in axiomatic set theory”, **Pacific Journal of Mathematics**, 10: 223–238, 1960.

LEVY, A. **Basic Set Theory**, Nova York: Springer, 1979.

MAGIDOR, M. “On the singular cardinals problem, II”, **Annals of Mathematics**, 106: 514–547, 1977.

MARTIN, D.A. e R. SOLOVAY “Internal Cohen Extensions”, **Annals of Mathematical Logic**, 2: 143–178, 1970.

MARTIN, D.A. e J.R. STEEL “A proof of projective determinacy”, **Journal of the American Mathematical Society,** 2(1): 71–125, 1989.

MATHIS, A.R.D. “Slim models of Zermelo Set Theory”, **Journal of Symbolic Logic**, 66: 487–496, 2001.

NEEMAN, I. “Inner models in the region of a Woodin limit of Woodin cardinals”, **Annals of Pure and Applied Logic**, 116: 67–155, 2002.

SCOTT, D. “Measurable cardinals and constructible sets”, **Bulletin de l’Académie Polonaise des Sciences. Série des Sciences Mathématiques, Astronomiques et Physiques**, 9: 521–524, 1961.

SHELAH, S. “Cardinal Arithmetic”, **Oxford Logic Guides**, 29, Nova York: The Clarendon Press, Oxford University Press, 1994.

SHELAH, S. **Proper and improper forcing**, 2ª Edição, Nova York: Springer-Verlag, 1998.

SHELAH, S. e W.H. WOODIN , “Large cardinals imply that every reasonably definable set of reals is Lebesgue measurable”, **Israel Journal of Mathematics**, 70(3): 381–394, 1990.

SOLOVAY, R. “A model of set theory in which every set of reals is Lebesgue

measurable", **Annals of Mathematics**, 92: 1–56, 1970.

SOLOVAY, R. e S. TENNENBAUM "Iterated Cohen extensions and Souslin's problem", **Annals of Mathematics** (2), 94: 201–245, 1971.

TODORCEVIC, S. "Partition Problems in Topology", **Contemporary Mathematics**, Volume 84. American Mathematical Society, 1989.

ULAM, S. "Zur Masstheorie in der allgemeinen Mengenlehre", **Fundamenta Mathematicae**, 16: 140 – 150, 1930.

WOODIN, W.H. **The Axiom of Determinacy, Forcing Axioms, and the Nonstationary Ideal**, De Gruyter Series in Logic and Its Applications 1, Berlin-Nova York: Walter de Gruyter, 1999.

WOODIN, W.H. 2001, "The Continuum Hypothesis, Part I", Notices of the AMS, 48(6): 567–576, e "The Continuum Hypothesis, Part II", **Notices of the AMS** 48(7): 681–690, 2001.

ZEMAN, M. "Inner Models and Large Cardinals", **De Gruyter Series in Logic and Its Applications** 5, Berlin-Nova York: Walter de Gruyter, 2001.

ZERMELO, E. "Untersuchungen über die Grundlagen der Mengenlehre, I", Mathematische Annalen 65: 261–281, 1908. Publicado em Zermelo 2010: 189–228, com introdução de Ulrich Felgner (2010).

Complemento 1 - A Teoria Básica dos Conjuntos*

Autoria: Joan Bagaria
Tradução: Sérgio R. N. Miranda
Revisão: Guilherme A. Cardoso

Conjuntos são coleções bem definidas que são completamente caracterizadas por seus elementos. Assim, dois conjuntos são iguais se e somente se eles têm exatamente os mesmos elementos. A relação básica na teoria dos conjuntos é a de pertinência ou pertencimento. Escrevemos $a \in A$ para indicar que o objeto a é um **elemento** ou um **membro** do conjunto *A*. Também dizemos que a **pertence a** *A*. Assim, um conjunto *A* é igual ao conjunto

*BAGARIA, J. "Set Theory", In: ZALTA, E. N. (ed.) **The Stanford Encyclopedia of Philosophy**, Winter 2021 Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em https://plato.stanford.edu/archives/win2021/entries/set-theory/basic-set-theory.html. Acesso em 20 jan. 2022

The following is the translation of the supplement to Set Theory by Joan Bagaria in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/set-theory/basic-set-theory.html. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

B se e somente se para todo a, $a \in A$ se e somente se $a \in B$. Em particular, há só um conjunto que não tem elementos. Essa conjunto é chamado, obviamente, de **conjunto vazio**, e é representado pelo símbolo $\emptyset$.

Dizemos que *A* é um subconjunto de *B*, escrito $A \subseteq B$, se e somente se todo elemento de *A* é também um elemento de *B*. Assim, $A = B$ se e somente se $A \subseteq B$ e $B \subseteq A$. Note que $\emptyset \subseteq A$, para qualquer A.

Dados os conjuntos *A* e *B*, pode-se realizar algumas operações básicas com eles gerando os seguintes conjuntos:

- O conjunto $A \cup B$, chamado de **união** de *A* e *B*, cujos elementos são os elementos de *A* e os elementos de *B*.
- O conjunto $A \cap B$, chamado de **interseção** de *A* e *B*, cujos elementos são os elementos comuns a *A* e a *B*.
- O conjunto $A - B$, chamado de **diferença** de *A* e *B*, cujos elementos são todos os elementos de *A* que não são membros de *B*.

É rotina checar que essas operações satisfazem as seguintes propriedades:

- Associatividade
  $A \cup (B \cup C) = (A \cup B) \cup C$
  $A \cap (B \cap C) = (A \cap B) \cap C$
- Comutatividade
  $A \cup B = B \cup A$
  $A \cap B = B \cap A$
- Distributividade
  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- Idempotência
  $A \cup A = A$
  $A \cap A = A$
  $A \cup \emptyset = A$
  $A \cap \emptyset = \emptyset$
  $A - A = \emptyset$
- Se $A \subseteq B$, então
  $A \cup B = A \cup (B - A) = B$

$$A \cap B = A$$

Dado um objeto a, podemos formar o conjunto que tem a como o seu único elemento. Esse conjunto é denotado por $\{a\}$. De modo mais geral, dados $a, b, c, ...$, podemos formar o conjunto que tem $a, b, c, ...$ como seus elementos, que denotamos por $\{a, b, c, ...\}$. Podemos realmente anotar todos os elementos do conjunto quando eles não são em grande número. No caso de conjuntos infinitos, isso claramente não é possível.

Se $a = b$, então $\{a, b\} = \{a\}$. Além disso, para qualquer a e b, o par $\{a, b\}$ é o mesmo que o par $\{b, a\}$. Portanto, se queremos levar em conta a ordem na qual os dois elementos do par são dados, precisamos encontrar outro modo de representar o par. Assim, definimos um **par ordenado** $\langle a, b\rangle$ como o conjunto $\{\{a\}, \{a, b\}\}$. Pode-se facilmente checar que dois pares ordenados $\langle a, b\rangle$ e $\langle c, d\rangle$ são iguais se e somente se $a = c$ e $b = d$. A ordem é agora importante, pois se $a \neq b$, então $\langle a, b\rangle \neq \langle b, a\rangle$.

O **produto cartesiano** $A \times B$ de dois conjuntos, *A* e *B*, é definido como o conjunto de todos os pares ordenados $\langle a, b\rangle$ tal que $a \in A$ e $b \in B$.

O produto cartesiano $A_1 \times ... \times A_n$, dos conjuntos $A_1, ..., A_n$, é o conjunto de todas as n-uplas $(a_1, ..., a_n)$ tal que $a_i \in A_i$ para todo $1 \leq i \leq n$. Em particular, para $n \geq 2$, o produto cartesiano de um conjunto multiplicado *n* vezes, denotado por A^n, é a soma de todas as *n*-uplas de elementos de *A*.

Relações

Uma **relação binária** em um conjunto *A* é um conjunto de pares ordenados dos elementos de *A*, quer dizer, um subconjunto de $A \times A$. Em geral, uma relação *n*-ária em um conjunto *A* é um subconjunto de A^n.

Uma relação binária *R* no conjunto *A* é chamada de **reflexiva** se $\langle a, a\rangle \in R$ para todo $a \in A$. Ela é chamada de **simétrica** se $\langle b, a\rangle$ sempre que $\langle a, b\rangle \in R$. E é chamada de **transitiva** se $\langle a, c\rangle \in R$ sempre que $\langle a, b\rangle \in R$ e $\langle b, c\rangle \in R$. Uma relação que é reflexiva, simétrica e transitiva é chamada de **relação de equivalência**. A relação de identidade em um conjunto *A* é o exemplo paradigmático de uma relação de equivalência. Outro exemplo é a relação no conjunto de todos os conjuntos finitos de números naturais que consiste de todos

os pares $\langle a, b\rangle$ tais que a e b têm o mesmo número de elementos.

Se *R* é uma relação de equivalência em um conjunto *A*, e $\langle a, b\rangle \in R$, então dizemos que a e b são *R*-equivalentes. Para todo $a \in A$, a **classe de equivalência** de a, usualmente denotada por $[a]_R$, é o conjunto de todos os elementos de *A* que são *R*-equivalentes a a. O conjunto de todas as classes *R*-equivalentes é chamado de **conjunto quociente** e é denotado por A/R. Pode-se facilmente checar que A/R é uma **partição** de *A*, ou seja, nenhum elemento de *A/R* é vazio, e quaisquer dois elementos de *A/R* são disjuntos, e todo $a \in A$ pertence a (exatamente) um elemento de *A/R*, exatamente a classe $[a]_R$.

Se *R* é uma relação binária, então usualmente se escreve aRb em vez de $\langle a, b\rangle \in R$.

Uma relação binária *R* em um conjunto *A* é chamada de **antissimétrica** se $a = b$ sempre que *aRb* e *bRa*. Uma relação *R* em um conjunto *A* que é reflexiva, antissimétrica e transtiva é chamada de **ordem parcial** (reflexiva). Se removemos de *R* todos os pares $\langle a, a\rangle$, para todo $a \in A$, então temos uma ordem parcial **estrita**. A relação $\subseteq$ em algum conjunto de conjuntos é um exemplo de ordem parcial. Uma ordem parcial em um dado conjunto *A* é usualmente representada pelo símbolo $\leq$, e a ordenção parcial estrita correspondente por <. Uma ordem parcial $\leq$ em um conjunto *A* com a propriedade adicional que ou $a \leq b$ ou $b \leq a$, para todos os elementos a e b de A, é chamada de **ordem total** ou **ordem linear**. As ordenações usuais do conjunto $\mathbb{N}$ de números naturais, do conjunto $\mathbb{Z}$ de inteiros, do conjunto $\mathbb{Q}$ de números racionais, ou do conjunto $\mathbb{R}$ e números reais, são ordens lineares.

Note que se $\leq$ é uma ordem linear no conjunto *A*, e $B \subseteq A$, então $\leq \cap B^2$ é também uma ordem linear de *B*. Se $\leq$ é uma ordem linear do conjunto A, então dizemos que $a \in A$ é o $\leq$-menor elemento de A se não há um $b \in A$ distinto de a tal que $b \leq a$. O número 0 é o menor elemento de $\mathbb{N}$, mas $\mathbb{Z}$ não tem o menor elemento.

Uma ordem linear $\leq$ no conjunto *A* é uma **boa ordem** se todo subconjunto não vazio de *A* tem um $\leq$-menor elemento. De forma equivalente, se não há uma sequência estritamente descendente

$$\ldots < a_2 < a_1 < a_0$$

de elementos de *A*. Desse modo, a ordenação usual de $\mathbb{N}$ é uma boa ordem. Mas a ordenação usual de $\mathbb{Z}$ não é uma boa ordem, porque ela não tem um menor elemento.

2 Funções

Uma função (1-ária) em um conjunto *A* é uma relação binária *F* em *A* tal que para todo $a \in A$ há exatamente um par $\langle a, b\rangle \in F$. O elemento b é chamado de **valor** de *F* em a, e é denotado por *F(a)*. E o conjunto *A* é chamado de **domínio** de *F*. A notação $F : A \longrightarrow B$ indica que *F* é uma função com domíno *A* e valores no conjunto *B*. Para $n \geq 2$, uma função *n*-ária em *A* é uma função $F : A^n \longrightarrow B$ para algum *B*.

Uma função $F : A \longrightarrow B$ é **injetora** se para todos os elementos a e b de *A*, se $a \neq b$, então $F(a) \neq F(b)$. E ela é **sobrejetora** se para todo $b \in B$ há algum $a \in A$ tal que $F(a) = b$. Finalmente, *F* é **bijetora** se ela é injetora e sobrejetora. Assim, uma bijeção $F : A \longrightarrow B$ estabelece uma correspondência um-para-um entre os elementos de *A* e aqueles de *B*, e *A* é **bijetável** com *B* se há tal bijeção. A **função de identidade** no conjunto *A*, denotada por $Id : A \longrightarrow A$, e que consiste nos pares $\langle a, a\rangle$, com $a \in A$, é trivialmente uma bijeção.

Dadas as funções $F : A \longrightarrow B$ e $G : B \longrightarrow C$, a **composição** de *F* e *G*, escrita $G \circ F$, é uma função $G \circ F : A \longrightarrow C$ cujos elementos são todos os pares $\langle a, G(F(a))\rangle$, em que $a \in A$. Se *F* e *G* são bijeções, então $G \circ F$ é também uma bijeção.

3 Conjuntos e Fórmulas

A **linguagem formal da teoria dos conjuntos** é a linguagem de primeira ordem cujo único símbolo não lógico é o símbolo da relação binária $\in$.

Dada qualquer fórmula $\phi(x, y_1, ..., y_n)$ da linguagem da teoria dos conjuntos, e os conjuntos $A, B_1, ...B_n$, pode-se formar o conjunto de todos os elementos de *A* que satisfazem a fórmula $\phi(x, B_1, ..., B_n)$. Esse conjunto é denotado por $\{a \in A | \phi(a, B_1, ..., B_2)\}$. Eis alguns exemplos:

- $\emptyset = \{a \in A | a \neq a\}$

- $A = \{a \in A | a = a\}$
- $A - B = \{a \in A | a \notin B\}$.
- $A \cap B = \{a \in A | a \in B\}$.

E se *A* e *C* são subconjuntos de *A*, então

- $B \cup C = \{a \in A | a \in B \vee a \in C\}$.

Dado um subconjunto $C \subseteq A \times B$, a **projeção** de *C* (sobre a primeira coordenada) é o conjunto

- $\{a \in A | \exists b \in B(\langle a, b \rangle \in C)\}$.

Não é o caso, no entanto, que dada alguma fórmula $\phi(x, y_1, ..., y_2)$, e os conjuntos $B_1, ..., B_n$, pode-se formar o conjunto de todos aqueles conjuntos que satisfazem a fórmula $\phi(x, B_1, ..., B_2)$. A razão é a seguinte. Seja $\phi(x)$ a fórmula $x \notin x$. Se *A* fosse o conjunto de todos os conjuntos que satisfazem a fórmula, então $A \in A$ se e somente se $A \notin A$. Uma contradição! Essa contradição é conhecida como **Paradoxo de Russell**, em homenagem a Bertrand Russell, que o descobriu em 1901 (*vide* o verbete **Russell's Paradox** [8] da SEP).

4 Ordinais

O primeiro número ordinal é $\emptyset$. Dado um ordinal α, o próximo ordinal maior, chamado de **sucessor** (imediato) de α, é o conjunto $a \cup \{a\}$. Portanto, o sucessor de α é o conjunto α junto com um elemento a mais, justamente o próprio α.

Na teoria dos conjuntos, os **números naturais** são definidos como ordinais finitos. Assim:

- $0 = \emptyset$
- $1 = \emptyset \cup \{\emptyset\} = \{\emptyset\}$
- $2 = \{\emptyset\} \cup \{\{\emptyset\}\} = \{\emptyset, \{\emptyset\}\}$

[8] N.T.: Disponível em: https://plato.stanford.edu/entries/russell-paradox/. Acesso em: 20 jan. 2022

- $3 = \{\emptyset, \{\emptyset\}\} \cup \{\{\emptyset, \{\emptyset\}\}\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$

  $\vdots$

Note que $1 = \{0\}, 2 = \{0, 1\}, 3 = \{0, 1, 2\}$, e em geral temos $n = \{0, 1, 2, ..., n - 1\}$. Assim, todo número natural *n* é exatamente o conjuntos de seus predecessores.

Um conjunto *A* é **finito** se não há uma correspondência um-para-um entre algum número natural *n* e os elementos de *A*, ou seja, uma bijeção $F : n \longrightarrow A$, caso no qual dizemos que *A* tem *n* elementos. Um conjunto é **infinito** se ele não é finito.

O conjunto de todos os ordinais finitos é denotado pela letra grega ômega (ω). Assim, ω é o conjunto $\mathbb{N}$ de números naturais. ω é também um ordinal, o primeiro ordinal infinito. Note que ω não é o sucessor de algum ordinal, e assim é chamado de **ordinal limite**. Agora que temos ω, podemos continuar gerando mais ordinais produzindo o seu sucessor $\omega \cup \{\omega\}$, então o seu sucessor $(\omega \cup \{\omega\}) \cup \{\omega \cup \{\omega\}\}$, e assim por diante. Todos os números ordinais maiores do que 0 são produzidos deste modo: ou produzindo o sucessor do último ordinal produzido ou, se não há esse último ordinal, produzindo o conjunto de todos os ordinais até então produzidos, como no caso de ω que gera um novo ordinal limite. Note, no entanto, que não se pode produzir o conjunto de **todos** os ordinais, pois, então, esse conjunto seria um novo ordinal limite, o que é impossível, visto que já teríamos todos eles.

Como no caso dos ordinais finitos, todo ordinal infinito é o conjunto dos seus predecessores. Uma consequência disso é que a relação $\in$ é uma boa ordem estrita em qualquer conjunto de ordinais. Portanto, para quaisquer ordinais α e β, definimos $\alpha < \beta$ se e somente se $\alpha \in \beta$. Assim, a boa ordem reflexiva associada é definida como $\alpha \leq \beta$ se e somente se $\alpha < \beta$ ou $\alpha = \beta$. Observe agora que $\alpha \subseteq \beta$ se e somente se $\alpha \leq \beta$.

5 Conjuntos Contáveis e Incontáveis

Se *A* é um conjunto finito, há uma bijeção $F : n \longrightarrow A$ entre um número natural *n* e *A*. Qualquer bijeção dessas oferece uma **contagem** dos

elementos de *A*, a saber, *F(0)* é o primeiro elemento de *A*, *F(1)* é o segundo, e assim por diante. Portanto, todos os conjuntos finitos são contáveis. Um conjunto infinito *A* é chamado de **contável** se há uma bijeção $F : \omega \longrightarrow A$ entre o conjunto dos números naturais e *A*. O conjunto $\mathbb{N}$ dos números naturais é (trivialmente) contável. Se *A* é um subconjunto infinito de ω, então *A* é também contável: pois seja $F : \omega \longrightarrow A$ tal que *F(n)* é o menor elemento de *A* que não está no conjunto $\{F(m) \in A | m < n\}$. Então *F* é uma bijeção.

Todo subconjunto infinito de uma conjunto contável é também contável: pois suponha que $F : \omega \longrightarrow A$ é uma bijeção e $B \subset A$ é infinito. Então o conjunto $\{n \in \omega | F(n) \in B\}$ é um subconjunto infinito de ω, por isso contável, e assim há uma bijeção $G : \omega \longrightarrow \{n \in \omega | F(n) \in B\}$. Então a função composta $F \circ G : \omega \longrightarrow B$ é uma bijeção.

A união de um conjunto contável e um conjunto finito é também contável. Pois dados os conjuntos *A* e *B*, que, sem perda de generalidade, podemos assumir como disjuntos, e dadas as bijeções $F : \omega \longrightarrow A$ e $G : n \longrightarrow B$, para algum $n < \omega$, seja $H : \omega \longrightarrow A \cup B$ a bijeção dada por: $H(m) = G(m)$, para todo *m* < *n*, e $H(m) = F(m - n)$, para todo $n \leq m$.

Além disso, a união de dois conjuntos contáveis é também contável: visto que já mostramos que a união de um conjunto contável e um conjujnto finito é contável, é suficiente ver que a união de dois conjuntos contáveis disjuntos é também contável. Desse modo, suponha que *A* e *B* são conjuntos contáveis e $F : \omega \longrightarrow A$ e $G : \omega \longrightarrow B$ são bijeções, então a função $H : \omega \longrightarrow A \cup B$, consistindo de todos os pares $\langle 2n, F(n) \rangle$, mais todos os pares $\langle 2n + 1, G(n) \rangle$ é uma bijeção.

Assim, o conjunto $\mathbb{Z}$, sendo a união de dois conjuntos contáveis, qual seja,

$$\mathbb{N} \cup \{-1, -2, -3, -4, ...\}$$

é também contável.

O produto cartesiano de dois conjuntos infinitos contáveis é também contável. Pois suponha que $F : \omega \longrightarrow A$ e $G : \omega \longrightarrow B$ são bijeções. Assim, usando o fato de que a função $J : \omega \times \omega \longrightarrow \omega$ dada por $J(\langle m, n \rangle) = 2^m(2n + 1) - 1$ é uma bijeção, temos a função $H : \omega \longrightarrow A \times B$ dada por $H(2^m(2n + 1) - 1) = \langle F(m), G(n) \rangle$ é

também uma bijeção.

Visto que qualquer número racional é dado por um par de inteiros, ou seja, o quociente $\frac{m}{n}$, em que $m, n \in \mathbb{Z}$ e $n \neq 0$, o conjunto $\mathbb{Q}$ dos números racionais é também contável.

No entanto, Georg Cantor descobriu que o conjunto $\mathbb{R}$ dos números reais não é contável. Pois suponha, visando uma contradição, que $F : \omega \longrightarrow \mathbb{R}$ é uma bijeção. Seja $a_0 = F(0)$. Escolha o menor *k* tal que $a_0 < F(k)$ e coloque $b_0 = F(k)$. Dados a_n e b_n, escolha o menor *l* tal que $a_n < F(l) < b_n$, e coloque $a_{n+1} = F(l)$. E escolha o menor *m* tal que $a_{n+1} < F(m) < b_n$, e $b_{n+1} = F(m)$. Desse modo, temos $a_0 < a_1 < a_2 < ... < b_2 < b_1 < b_0$. Agora, seja a o limite de a_n. Então a é um número real diferente de *F(n)*, para todo *n*, o que é impossível porque *F* é uma bijeção.

A existência de conjuntos incontáveis se segue de um fato muito mais geral, também descoberto por Cantor. O fato é o seguinte: dado qualquer conjunto *A*, o conjunto de todos os subconjuntos, chamado de **conjunto potência** de *A*, e denotado por $\mathcal{P}(A)$, não é bijetável com *A*: pois suponha que $F : A \longrightarrow \mathcal{P}(A)$ é uma bijeção. Então o subconjunto $\{a \in A | a \notin F(a)\}$ de *A* é o valor de *F(a)* para algum $a \in A$. Mas então $a \in F(a)$ se e somente se $a \notin F(a)$. Portanto, se *A* é um conjunto infinito qualquer, então $\mathcal{P}(A)$ é incontável.

Há também ordinais incontáveis. O conjunto de todos os ordinais finitos e contáveis é também um ordinal, chamado de ω_1, que é o primeiro ordinal incontável. Similarmente, o conjunto de todos os ordinais que são bijetáveis com algum ordinal menor ou igual a ω_1 é também um ordinal, chamado de ω_2, que não é bijetável com ω_1, e assim por diante.

5.1 Cardinais

A **cardinalidade**, ou tamanho, de um conjunto finito *A* é o único número natural *n* tal que há uma bijeção $F : n \longrightarrow A$.

No caso dos conjuntos finitos, a cardinalidade é dada, não por um número natural, mas por um ordinal infinito. No entanto, contrariamente aos conjuntos finitos, um conjunto infinito *A* é bijetável com muitos números ordinais diferentes. Por exemplo, o conjunto $\mathbb{N}$ é bijetável com ω, mas também com seu

sucessor $\omega \cup \{\omega\}$: atribuindo 0 a ω e $n+1$ a *n*, para todo $n \in \omega$, obtemos uma bijeção entre $\omega \cup \{\omega\}$ e ω. Mas visto que os ordinais são bem ordenados, podemos definir a cardinalidde de um conjunto infinito como o menor ordinal que é bijetável com ele.

Em particular, a cardinalidade de um número ordinal α é o menor ordinal κ que é bijetável com ele. Note que κ não é bijetável com qualquer ordinal menor, pois, de outro modo, ele seria α. Os números ordinais que não são bijetáveis com algum ordinal menor são chamados de **números cardinais**. Assim, todos os números naturais são cardinais, e da mesma forma $\omega, \omega_1, \omega_2$, e assim por diante. Em geral, dado algum cardinal κ, o conjunto de todos os ordinais que são bijetáveis com algum ordinal $\leq \kappa$ é também um cardinal; ele é o menor cardinal maior do que κ.

Os cardinais infinitos são representados pela letra grega aleph ($\aleph$) do alfabeto hebraico. Assim, o menor cardinal infinito é $\omega = \aleph_0$, o próximo é $\omega_1 = \aleph_1$, que é o primeiro cardinal incontável, então vem $\omega_2 = \aleph_2$, etc.

A cardinalidade de qualquer conjunto, denotada por |*A*|, é o número cardinal único que é bijetável com *A*. Vimos que $|\mathbb{R}|$ é incontável, por isso maior do que $\aleph_0$, mas não se sabe qual número cardinal ele é. A conjectura que $|\mathbb{R}| = \aleph_1$, formulada por Cantor em 1878, é a famosa **Hipótese do Contínuo**.

Leituras Recomendadas

DEVLIN, K. **The Joy of Sets: Fundamentals of Contemporary Set Theory**. Undergraduate Texts in Mathematics, Nova York: Springer, 1993.

ENDERTON, H.B. **Elements of Set Theory,** Nova York: Academic Press, 1977.

JECH, T. e K. HRBAČEK **Introduction to set theory,** Nova York: Marcel Dekker, 3ª Edição, 1999 [1978].

Complemento 2 - A Teoria dos Conjuntos de Zermelo-Fraenkel*

Autoria: Joan Bagaria
Tradução: Sérgio R. N. Miranda
Revisão: Guilherme A. Cardoso

Axiomas de ZF

Extensionalidade

$\forall x \forall y[\forall z(z \in x \leftrightarrow z \in y) \rightarrow x = y]$

Esse axioma assevera que quando os conjuntos *x* e *y* têm os mesmos

*BAGARIA, J. "Set Theory", In: ZALTA, E. N. (ed.) **The Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/set-theory/ZF.html. Acesso em: 20 jan. 2022.

The following is the translation of the supplement entry on Zermelo Fraenkel Set Theory by Joan Bagaria in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/set-theory/ZF.html. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry.Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

membros, eles são o mesmo conjunto.

O próximo axioma assevera a existência do conjunto vazio:

Conjunto Vazio
$\exists x \neg \exists y (y \in x)$

Visto que é demonstrável, a partir desse axioma e do anterior, que há um único conjunto vazio, podemos itnroduzir a notação $\emptyset$ para denotá-lo.

O próximo axioma assevera que se dados dois conjuntos quaisquer *x* e *y*, existe um conjunto par de *x* e *y*, ou seja, um conjunto que tem apenas *x* e *y* como membros:

Paridade:
$\forall x \forall y \exists z \forall w (w \in z \leftrightarrow w = x \vee w = y)$

Visto que é provável que há apenas um conjunto par para cada *x* e *y* dados, introduzimos a notação $\{x, y\}$ para denotá-lo.

O próximo axioma assevera que para qualquer conjunto *x*, há um conjunto *y* que contém como membros todos aqueles conjuntos cujos membros são também elementos de *x*, ou seja, *y* contém todos os subconjujntos de *x*:

Conjunto Potência:
$\forall x \exists y \forall z [z \in y \leftrightarrow \forall w (w \in z \rightarrow w \in x)]$

Visto que todo conjunto provavelmente tem um único "conjunto potência", introduzimos a notação $\mathcal{P}(x)$ para denotá-lo. Note também que podemos definir a noção de **é um subconjunto de y** ($x \leq y$) como: $\forall z (z \in x \rightarrow z \in y)$. Assim, podemos simplificar a apresentação do conjunto potência deste modo:

$$\forall x \exists y \forall z (z \in y \leftrightarrow z \leq x)$$

O próximo axioma assevera que para um dado conjunto *x*, há um conjunto *y* que tem como membro todos os membros de todos os membros de *x*:

Uniões:
$\forall x \exists y \forall z [z \in y \leftrightarrow \exists w (w \in x \wedge z \in w)]$.

Visto que é provável que há uma única "união" de qualquer conjunto *x*, introduzimos a notação $\bigcup x$ para denotá-lo.

O próximo axioma assevera a existência de um conjunto infinito, ou seja, um conjunto que tem um número infinito de membros:

Infinitude:
$\exists x [\emptyset \in x \wedge \forall y (y \in x \rightarrow \bigcup \{y, \{y\}\} \in x)]$.

Podemos entendê-lo nestes termos. Vamos definir a **união** de *x* e *y* ($x \cup y$) como a união do conjunto par de *x* e *y*, ou seja, como $\bigcup \{x, y\}$. Então o axioma da Infinitude assevera que há um conjunto *x* que contém $\emptyset$ como membro e que é tal que sempre que um conjunto *y* é membro de *x*, então $y \cup \{y\}$ é membro de *x*. Consequentemente, esse axioma garante a existência de um conjunto da seguinte forma:

$$\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, ...\}$$

Note que o segundo elemento, $\{\emptyset\}$, está nesse conjunto porque (1) o fato de que $\emptyset$ está no conjunto implica que $\emptyset \cup \{\emptyset\}$ está no conjunto e (2) $\emptyset \cup \{\emptyset\}$ é examente $\{\emptyset\}$. Similarmente, o terceiro elemento, $\{\emptyset, \{\emptyset\}\}$, está nesse conjunto porque (1) o fato de que $\{\emptyset\}$ está no conjunto implica que $\{\emptyset\} \cup \{\{\emptyset\}\}$ está no conjunto e (2) $\{\emptyset\} \cup \{\{\emptyset\}\}$ é examente $\{\emptyset, \{\emptyset\}\}$. E assim por diante.

O próximo axioma é o **Esquema de Separação**, que assevera que a existência de um conjunto que contém os elementos de um dado conjunto *w* que satisfaz uma certa condição ψ. Quer dizer, suponha que $\psi(x, \hat{u})$ tenha *x* livre e pode ou não ter $u_1, ..., u_k$ livre. Seja $\psi_{x,\hat{u}}[r, \hat{u}]$ o resultado de substituir *r* por *x* em $\psi(x, \hat{u})$. Então o Esquema de Separação assevera:

Esquema de Separação

$\forall u_1 ... \forall u_k [\forall w \exists v \forall r (r \in v \leftrightarrow r \in w \wedge \psi_{x,\hat{u}}[r, \hat{u}])]$

Em outros termos, se dada uma fórmula ψ e um conjunto *w*, existe um conjunto *v* que tem como membros exatamente os membros de *w* que satisfazem a fórmula ψ.

O próximo axioma de ZF é o **Esquema de Substituição**. Suponha que $\phi(x, y, \hat{u})$ é uma fórmula com *x* e *y* livres, e seja $\hat{u}$ a representação das variáveis $u_1, ..., u_k$, que podem ou não ser livres em ϕ. Além disso, seja $\phi_{x,y,\hat{u}}[s, r, \hat{u}]$ o resultado de substituir *s* e *r* por *x* e *y*, respectivamente, em $\phi(x, y, \hat{u})$. Assim, toda instância do seguinte axioma é um axioma:

Esquema de Substituição:
$\forall u_1 ... \forall u_k [\forall x \exists! y \phi(x, y, \hat{u}) \rightarrow \forall w \exists v \forall r (r \in v \leftrightarrow \exists s (s \in w \wedge \phi_{x,y,\hat{u}}[s, r, \hat{u}]))]$

Em outros termos, se sabemos que ϕ é uma fórmula funcional (que relaciona cada conjunto *x* a um único conjunto *y*), então se nos é dado um conjunto *w*, podemos formar um novo conjunto *v* do seguinte modo: junte todos os conjuntos aos quais os membros de *w* são exclusivamente relacionados por ϕ.

Note que o Esquema de Substituição pode levar você para fora do conjunto w quando formar o conjunto *v*. Os elementos de *v* não precisam ser elementos de w. Contrariamente, o Esquema de Separação de Zermele produz somente conjunto de um dado conjunto w.

O último axioma assevera que todos conjunto é "bem-fundado":

Regularidade:
$\forall x [x \neq \emptyset \rightarrow \exists y (y \in x \wedge \forall z (z \in x \rightarrow \neg (z \in y)))]$

Um membro de *y* de um conjunto *x* com essa propriedade é chamado de elemento "mínimo". Essa axioma exclui a existência de cadeias circulares de conjuntos (tais como $x \in y \wedge y \in z \wedge z \in x$), como também cadeias infinitamente descendentes de conjuntos (tais como $...x_3 \in x_2 \in x_1 \in x_0$).

Os Teoremas da Incompletude de Gödel*

Autoria: Panu Raatikainen
Tradução: Guilherme A. Cardoso
Revisão: Sérgio R. N. Miranda

Os dois teoremas da incompletude de Gödel estão entre os mais importantes resultados da lógica contemporânea e têm implicações profundas para várias questões. Esses teoremas dizem respeito aos limites da demonstrabilidade em teorias axiomáticas formais. O primeiro teorema da incompletude afirma que em qualquer sistema formal consistente *F* no qual uma certa quantidade de aritmética pode ser feita, há afirmações da linguagem de *F*

*RAATIKAINEN, P. "Gödel's Incompleteness Theorems", In: ZALTA, E. N. **The Stanford Encyclopedia of Philosophy**, Spring Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/goedel-incompleteness/. Acesso em: 20 jan. 2022.

The following is the translation of the entry on Gödel's Incompleteness Theorems by Panu Raatikainen in the **Stanford Encyclopedia of Philosophy**. The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/goedel-incompleteness/. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

que não podem ser demonstradas e nem refutadas em *F*. De acordo com o segundo teorema da incompletude, tal sistema formal não pode demonstrar que ele mesmo é consistente (assumindo que seja realmente consistente). Esses resultados tiveram grande impacto na filosofia da matemática e na lógica. Há tentativas de aplicar os resultados também em outras áreas da filosofia, como a filosofia da mente, mas essas tentativas são mais controversas. (Para uma discussão que contextualiza os teoremas da incompletude no trabalho matemático e filosófico de Gödel, *vide* o verbete **Kurt Gödel**[9] da SEP.)

1. Introdução

1.1. Panorama

Os teoremas da incompletude de Gödel estão entre os mais importantes resultados da lógica contemporânea.Essas descobertas revolucionaram o entendimento da matemática e da lógica, e tiveram implicações dramáticas para a filosofia da matemática. Há também tentativas de aplicá-las em outras áreas da filosofia, mas a legitimidade de muitas dessas aplicações é mais controversa.

Para entender os teoremas de Gödel, primeiramente temos de explicar os conceitos-chave que lhes são essenciais, como "sistema formal", "consistência" e "completude". *Grosso modo*, uma **sistema formal** é um sistema de axiomas equipado com regras de inferência, que permitem a geração de novos teoremas. Há a exigência de que o conjunto de axiomas seja finito ou ao menos decidível, isto é, tem de haver um algorítmo (um método efetivo) que permita decidir mecanicamente se uma dada afirmação é ou não um axioma. Se essa condição é satisfeita, a teoria é chamada de "recursivamente axiomatizável", ou, simplesmente, "axiomatizável". As regras de inferência (de um sistema formal) são também operações efetivas, de tal modo que possamos sempre decidir mecanicamente se temos em mãos uma aplicação legítima de uma regra de inferência. Portanto, é também possível decidir para qualquer sequência finita de fórmulas se essa sequência é uma derivação genuína, ou uma prova, no sistema -

[9]N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/goedel/. Acesso em: 20 jan. 2022

dados os axiomas e as regras de inferência desse sistema.

Um sistema formal é **completo** se para toda afirmação da linguagem do sistema ou a afirmação ou a sua negação pode ser derivada (isto é, demonstrada) no sistema. Um sistema formal é **consistente** se não há uma afirmação tal que tanto a afirmação ela mesma quanto a sua negação sejam deriváveis no sistema. Somente sistemas consistentes são de algum interesse nesse contexto, pois é um fato elementar da lógica que em um sistema formal inconsistente qualquer afirmação é derivável e, consequentemente, esse sistema é trivialmente completo.

Gödel estabeleceu dois teoremas da incompletude diferentes, embora relacionados, usualmente chamados de "primeiro teorema da incompletude" e "segundo teorema da incompletude". A expressão "Teorema de Gödel" é algumas vezes usada para se referir à conjunção desses dois teoremas, mas pode se referir a cada um deles separadamente - usualmente ao primeiro. Com um desenvolvimento devido a J. Barkley Rosser em 1936, o primeiro teorema pode ser colocado, grosseiramente, nos seguintes termos:

> **Primeiro Teorema da Incompletude**
> Todo sistema formal consistente *F* no qual uma certa quantidade de aritmética elementar possa ser executada é incompleto; ou seja, há afirmações na linguagem de *F* que não podem ser demonstradas e nem refutadas em *F*.

O teorema de Gödel não só diz que tais afirmações existem: o método da prova de Gödel explicitamente produz uma sentença especial que não é demonstrada e nem refutada em *F*; a afirmação "indecidível" pode ser produzida mecanicamente a partir de uma especificação de *F*. A sentença em questão é uma afirmação relativamente simples da teoria dos números, uma sentença pura da aritmética universal.

Um mal-entendido frequente é interpretar o primeiro teorema de Gödel como tendo mostrado que há verdades que não podem ser demonstradas. Contudo, isso é um erro, porque o teorema da incompletude não trata da demonstrabilidade em um sentido absoluto, mas só diz respeito à derivabilidade em algum sistema formal particular. Para qualquer afirmação *A* que não pode ser demonstrada em um sistema formal particular *F*, há, trivialmente, outros sistemas

formais em que *A* é provável (considere *A* como axioma). Por outro lado, há o sistema axiomático padrão extremamente poderoso da teoria de conjuntos de Zermelo-Fraenkel (denotado por **ZF**, ou, com o Axioma da Escolha [*Choice*], **ZFC**; *vide* a seção sobre os axiomas de **ZFC** no verbete sobre a **Teoria dos Conjuntos** publicado neste livro), que é mais do que suficiente para a derivação de toda a matemática ordinária. Pelo primeiro teorema de Gödel, há verdades aritméticas que não são prováveis em **ZFC**. Portanto, seria necessário um sistema que incorporasse métodos que fossem além de **ZFC** para prová-las. Há assim um sentido em que tais verdades não são prováveis com o uso dos métodos matemáticos e axiomas "ordinários" atuais e nem podem ser demonstradas de um modo que os matemáticos encarariam hoje como não problemático e conclusivo.

O segundo teorema de Gödel diz respeito aos limites das provas de consistência. Uma apresentação grosseira é a seguinte:

> **Segundo Teorema da Incompletude**
> Para qualquer sistema consistente *F* no qual uma certa quantidade de aritmética elementar possa ser executada, a consistência de F não pode ser demonstrada em *F* mesmo.

No caso do segundo teorema, *F* tem de conter um pouco mais de aritmética do que no caso do primeiro teorema, que vale em condições muito fracas. É importante notar que esse resultado, como o primeiro teorema da incompletude, é um teorema sobre demonstrabilidade formal, ou derivabilidade (que é sempre relativa a algum sistema formal; nesse caso, o próprio *F*). Ele não diz qualquer coisa sobre se, para uma teoria particular *T* que satisfaz as condições do teorema, a afirmação "T é consistente" pode ser demonstrada, no sentido de ser mostrada verdadeira por um argumento conclusivo ou por uma prova geralmente aceita pelos matemáticos. Para muitas teorias, isso é perfeitamente possível.

1.2. Algumas Teorias Formalizadas

A existência de teorias incompletas é dificilmente surpreendente. Pegue uma teoria, mesmo uma teoria completa (*vide* exemplos abaixo), e retire algum axioma; a não ser que o axioma seja redundante, o sistema resultante será

incompleto. Os teoremas da incompletude, no entanto, lidam com um tipo de fenômeno de incompletude muito mais radical. Diferentemente do tipo de teorias trivialmente incompletas acima, que podem ser facilmente completadas, não há como completar as teorias relevantes; todas as suas extensões, enquanto forem ainda sistemas formais e por isso axiomatizáveis, são também incompletas. Elas permanecem, por assim dizer, eternamente incompletas e jamais podem ser completadas. Elas são "essencialmente incompletas".

Nas primeiras (e um tanto relaxadas) apresentações dos teoremas da incompletude oferecidas acima, ocorre a exigência vaga de que "uma certa quantidade de aritmética elementar possa ser executada". Agora é o momento de tornar isso mais preciso.

1.2.1. Teorias Aritméticas

O sistema de aritmética mais fraco que é usualmente considerado em conexão com a incompletude e indecidibilidade é a assim chamada "Aritmética de Robinson" (devida a Raphael M. Robinson; *vide* TARSKI; MOSTOWSKI; ROBINSON, 1953), denotada de maneira padrão por **Q**. Como axiomas, ela tem as sete seguintes suposições:

- $\neg(0 = x')$
- $x' = y' \rightarrow x = y$
- $\neg(x = 0) \rightarrow \exists y(x = y')$
- $x + 0 = 0$
- $x + y' = (x + y)'$
- $x \times 0 = 0$
- $x \times y' = (x \times y) + x$

A interpretação pretendida de "x'" é a função de sucessor, e, obviamente, de $+$ e $\times$, as funções de adição e multiplicação, respectivamente. "0" é a única constante e denota o número zero.

Acrescentando a esses axiomas elementares o axioma esquema da indução:

(IND) $$\phi(0) \wedge \forall x[\phi(x) \rightarrow \phi(x')] \rightarrow \forall x\phi(x),$$

obtemos como resultado a Aritmética de Peano [*Peano Arithmetic* (**PA**)] de primeira ordem. Note que, diferentemente de **Q**, **PA** contém infinitamente muitos axiomas, porque todas as instâncias (infinitamente muitas) do esquema de indução, uma correspondente a cada fórmula $\phi(x)$ (com ao menos uma variável livre) da linguagem, são consideradas axiomas. Mas é uma tarefa mecânica e rotineira checar se uma dada sentença é uma instância desse esquema. **PA** é geralmente considerado o sistema padrão de aritmética de primeira ordem.

Outro sistema de aritmética natural e muito estudado, cuja força fica entre **Q** e **PA**, é a Aritmética Primitiva Recursiva (**APR**). Ela contém não só os axiomas de **Q** acima governando as funções de sucessor, adição e multiplicação, mas também axiomas definidores para todas as funções recursivas (*vide* o verbete **Recursive Functions**[10] da SEP), e a aplicação do esquema de indução é restrita a fórmulas não quantificadas (isto é, não se permite que $\phi(x)$ contenha quaisquer quantificadores (ilimitados)).

No entanto, essencialmente o mesmo sistema é obtido se consideramos apenas os axiomas de **Q** e o esquema de indução restrito a, grosseiramente falando, fórmulas puramente existenciais (em termos técnicos, Σ_1^0-fórmulas; *vide* abaixo) (isso foi primeiramente mostrado por PARSONS, 1970). Além disso, pode-se mostrar que a Σ_1^0-indução é equivalente ao esquema de indução restrito a (grosseiramente falando) fórmulas puramente universais (Π_1^0-fórmulas) (*vide* PARIS; KIRBY, 1978). **APR** pode também ser formulada como um cálculo equacional de "lógica livre". **APR** é frequentemente considerada como a teoria de fundo não problemática em que vários outros sistemas, a legitimidade dos quais pode ser mais controversa, são estudados.

Um sistema muito mais forte do que **PA**, importante para os fundamentos da matemática, que será mencionado agora e na sequência, é a aritmética de segunda ordem **PA^2** (também frequentemente denotada por **Z_2**). Ele é mais do que suficiente para desenvolver toda a análise e álgebra ordinárias. A sua linguagem é uma linguagem de primeira ordem duplamente sortida, ou seja, ela contém dois tipos de variáveis, variáveis de números $x_1, x_2, ...$ (ou $x, y, z...$) e variáveis de propriedades $X_1, X_2, ...$ (ou X, Y, Z), em que propriedades são

[10] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/recursive-functions/. Acesso em: 20 jan. 2022

extensionalmente concebidas (*vide* o verbete **Second-order and Higher-order Logic**[11] da SEP). Como axiomas, ele inclui, além dos axiomas básicos de **PA**, todas as instâncias do esquema de compreensão de segunda ordem:

$$\exists X \forall x [Xx \leftrightarrow \phi(x)]$$

em que $\phi(x)$ pode ser qualquer fórmula da linguagem de **PA²** na qual X não ocorra livre. (Deve-se mencionar que **PA²** pode também ser formulado com o acréscimo à linguagem da noção primitiva de pertinência a conjuntos ($\in$), encarando as variáveis $X, Y, Z\ldots$ como percorrendo explicitamente conjuntos e reformulando a compreensão de segunda ordem nos seguintes termos: $\exists X \forall x$ $[x \in X \leftrightarrow \phi(x)]$.)

PA² é uma teoria muito forte. Pelo método de interpretações (*vide* abaixo), pode-se mostrar que ela é, do ponto de vista da teoria da prova, tão forte quanto a teoria dos conjuntos de Zermelo-Fraenkel **ZFC** sem o axioma do conjunto potência [*Power-set*], que podemos chamar de **ZFC** $-$ **Pow** (enquanto a teoria padrão, **PA** de primeira ordem, é similarmente equivalente, do ponto de vista da teoria da prova, a **ZFC** sem o axioma da infinitude, **ZFC** $-$ **Inf**). (*vide* a seção sobre os axiomas de **ZFC** no verbete sobre a **Teoria dos Conjuntos** publicado neste livro).

Obviamente, assume-se que os nossos sistemas formais são todos equipados com um sistema de **regras de inferência** (e possivelmente de alguns axiomas lógicos), usualmente algum sistema padrão de lógica clássica (embora os teoremas da incompletude não pressuponham essencialmente a lógica clássica, mas também se apliquem a sistemas como a lógica intuicionista). Os sistemas padrões acima vêm com a lógica clássica. A notação padrão $F \vdash A$ é usada para exprimir (no nível metalógico) que A é derivável de F, ou seja, que há uma prova de A em F, ou, em outros termos, que A é um **teorema** de F. De acordo com isso, $F \nvdash A$ significa que A *não* é derivável em F.

Para resumir: quando se diz, no contexto dos teoremas da incompletude, que "uma certa quantidade de arimética elementar possa ser executada" no sistema, isso usualmente significa que o sistema contém **ARP** ou ao menos **Q**.

[11] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/logic-higher-order/. Acesso em: 20 jan. 2022.

Para o primeiro teorema da incompletude, **Q** é suficiente; para as provas padrão do segundo teorema, algo como **ARP**, no mínimo, é necessário. Há uma versão do segundo teorema da incompletude para **Q** (*vide* BEZBORUH; SHEPHERDSON, 1976), mas tem havido algum debate sobre se a afirmação relevante em **Q** pode realmente ser entendida como exprimindo a consistência, visto ser **Q** muito fraco (*vide* KREISEL, 1958; BEZBORUAH; SHEPHERDSON, 1976; PUDIÁK, 1996; FRANKS, 2009).

1.2.2. Teorias não Formuladas na Linguagem da Aritmética

Obviamente, há muitas teorias importantes e interessantes na matemática que não são formuladas na linguagem da aritmética. No entanto, a aplicabilidade dos teoremas da incompletude podem ser dramaticamente estendidas para fora da linguagem da aritmética de primeira ordem e suas extensões, quando se nota que tudo o que é preciso é que teorias fracas como **Q** e **ARP** possam ser **interpretadas** no sistema em questão. Mais importante ainda, isso envolve vários sistemas de teoria dos conjuntos. Por exemplo, os teoremas da incompletude valem para **ZFC** $-$ **Inf** (isto é, **ZFC** sem o axioma da infinitude) e todas as suas extensões, não importa quão fortes (à medida que sejam axiomatizáveis).

Grosso modo, uma teoria T_1 é interpretável em outra teoria T_2 se os conceitos primitivos e o escopo das variáveis de T_1 são definíveis em T_2 de tal modo que é possível traduzir todo teorema de T_1 em um teorema de T_2. Não se deve entender essas interpretações como oferecendo algo como a sinonímia intuitiva. Duas teorias podem dizer respeito a assuntos radicalmente diferentes e, ainda assim, como sistemas formais, uma pode ser interpretada na outra. (Como ilustração: uma teoria simples de ancestrais pode ser, tomada como um sistema formal, interpretada na aritmética; obviamente, isso não quer dizer que coisas como avós sejam números). O que é importante é que a interpretabilidade preserva certas propriedades **formais** de teorias, sendo a consistência a mais importante: se T_1 é interpretável em T_2 e T_2 é consistente, T_1 é também consistente. É certo que qualquer sistema em que **Q** possa ser intepretada é essencialmente incompleto. Para qualquer teoria em que **Q** possa ser

interpretada, a incompletude poderia ser demonstrada também diretamente; por exemplo, em várias teorias da teoria dos conjuntos, pode-se codificar fórmulas e derivações (em vez de números) por conjuntos, "Conjuntos de Gödel", e proceder de modo usual (*vide* FITTING, 2007). No entanto, para a maioria dos propósitos, é muito mais simples estabelecer a interpretabilidade de **Q** na teoria em questão.

Em suma, quando se diz que "uma certa quantidade de aritmética elementar pode ser executada no sistema", o que se quer dizer é ou que o sistema é uma extensão axiomatizável de **Q** ou que **Q** pode ser interpretada nesse sistema. (No caso envolvendo (provas padrão do) segundo teorema da incompletude, substitua **ARP** por **Q**).

1.2.3. Algumas exceções: Teorias Completas

Por outro lado, nem todas as teorias da aritmética são incompletas. A teoria com apenas a adição de números naturais sem a multiplicação (frequentemente chamada de "Aritmética de Presburger"), por exemplo, é completa (e decidível) (PRESBURGER, 1929), assim como é a teoria da multiplicação de inteiros positivos (SKOLEM, 1930). Essas teorias, no entanto, são muito fracas. E de qualquer modo, ao menos uma teoria que trata tanto da adição quanto da multiplicação é necessária. Mais importante ainda, a teoria da aritmética dos **números reais** de primeira ordem mais natural (com tanto adição quanto multiplicação), a assim chamada "teoria de campos fechados reais" (**CFR**), é tanto completa quanto decidível, como foi mostrado em TARSKI, 1948; ele também demonstrou que a teoria de primeira ordem da geometria euclidiana é completa e decidível. Assim, deve-se ter em mente que há algumas teorias não triviais e interessantes às quais os teoremas de Gödel não se aplicam.

1.3. A Relevância de Tese de Church-Turing

Gödel originalmente estabeleceu apenas a incompletude de uma teoria **P** particular, embora muito compreensiva, uma variante do sistema de Russell de teoria dos tipos **PM** (de **Principia Mathematica**, *vide* as seções sobre Paradoxos

e Teorias dos Tipos de Russell nos verbetes **Type Theory**[12] e **Principia Mathematica**[13] da SEP), e toda extensão de **P** com a mesma linguagem, cujo conjunto de axiomas seja primitivo recursivo. Ele também sugeriu, embora não tenha demonstrado, que a prova poderia ser adaptada a fim de também ser aplicada a sistemas de axiomas padrão da teoria dos conjuntos, como **ZFC**. Embora tenha sido mostrado que Gödel de fato já tinha um resultado muito geral, na época não era claro quão geral era esse resultado (*vide* **Seção 5**.)

O que então faltava era uma análise da noção intuitiva de decidibilidade, necessária na caracterização da noção de um sistema formal arbitrário. Lembre-se que o conjunto de axiomas e a relação de prova de um sistema formalizado são necessários para a decidibilidade. Os matemáticos e lógicos têm implicitamente usado a noção intuitiva de método de decisão desde a antiguidade, e quando se perguntava por uma solução positiva, era suficiente que se apresentasse um método concreto que fosse intutivamente percebido por todos como um método mecânico. Para os resultados limitativos gerais, como os teoremas gerais da incompletude, ou os resultados de indecidibilidade (*vide* Seção **4.2**), no entanto, uma explicação matemática precisa da noção seria necessária. Em vez de conjuntos decidíveis ou propriedades, geralmente se considera funções ou operações efetivas ou computáveis, mas de fato esses são apenas dois lados da mesma moeda - a conversa sobre um pode ser facilmente transcrita na conversa sobre o outro.

Gödel (1934), Alonzo Church (1936a,1936b) e Alan Turing (1936-1937) chegaram independentemente a diferentes propostas para uma definição matemática exata de funções computáveis, e, consequentemente, de conjuntos decidíveis (de números). Essas propostas, no entanto, se mostraram equivalentes. A cuidadosa análise conceitual de Turing que usa máquinas ficcionais e abstratas (hoje em dia convencionalmente chamadas de "Máquinas de Turing"; *vide* o capítulo sobre as **Máquinas de Turing** publicado neste livro) foi particularmente importante, como o próprio Gödel enfatizou (*vide* GÖDEL, 1963).

[12]N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/type-theory/index.html#ParaRussTypeTheo. Acesso em: 20 jan. 2022.

[13]N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/principia-mathematica/. Acesso em: 20 jan. 2022.

A equação da noção intuitiva com algumas dessas explicações matemática é geralmente chamada de "A Tese de Church-Turing". Por razões históricas, o rótulo "função recursiva" tem sido dominante na literatura lógica. Consequentemente, conjuntos decidíveis são frequentemente chamado de "conjuntos recursivos". (*vide* os verbetes **Computability and Complexity**[14], **Recursive Functions**[15] e **Church-Turing Thesis**[16] da SEP).

Para uma compreensão adequada dos resultados de incompletude e indecidibilidade, é vital entender a diferença entre duas noções centrais a respeito de conjuntos. Em primeiro lugar, pode haver um método mecânico que decide se algum número dado pertence ou não ao conjunto em questão (e nesse caso o conjunto é chamado de "decidível" ou "recursivo"), e, em segundo lugar, pode haver um método mecânico que gera ou lista os elementos do conjunto, número por número. Neste último caso, o conjunto é chamado de "recursivamente enumerável" (r.e.), quer dizer, ele pode ser efetivamente gerado ou é "semi-decidível". É um resultado fundamental da teoria da computabilidade (ou "Teoria das Funções Recursivas") que há conjuntos semi-decidíveis, conjuntos que podem ser efetivamente gerados (isto é, são recursivamente enumeráveis), mas **não** são decidíveis (isto é, não recursivos). De fato, essa é, em um nível bem abstrado, a essência do primeiro teorema da incompletude. No entanto, se tanto um conjunto quanto o seu complemento são recursivamente enumeráveis, o conjunto é recursivo, ou seja, decidível.

2. O Primeiro Teorema da Incompletude

Nesta seção, as linhas principais da prova do primeiro teorema da incompletude são esboçadas. Para mais detalhes da prova, o leitor interessado pode consultar os complementos deste capítulo: **Numeração de Gödel** e **O Lema da Diagonalização**.

---

[14] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/computability/. Acesso em: 20 jan. 2022.

[15] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/recursive-functions/. Acesso em: 20 jan. 2022.

[16] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/church-turing/. Acesso em: 20 jan. 2022.

2.1. Preliminares

O termo formal ("numeral") que canonicamente denota o número natural **n** é abreviado como $\underline{n}$. Na linguagem padrão da aritmética usada neste verbete, o número **n** é denotado pelo termo $0'^{\cdots'}$, em que o símbolo de sucessor "′" é repetido *n* vezes. Ou seja, numerais que nomeiam 1,2,3,... são $0'$, $0''$, $0'''$,... e abreviados por $\underline{1}, \underline{2}, \underline{3}, \ldots$

Em sua prova original, Gödel usou a sua noção específica de ω-consistência, e para alguns propósitos é ainda conveniente seguir a abordagem inicial de Gödel. Uma teoria formalizada *F* é ω-consistente se **não** é o caso que para alguma fórmula $A(x)$, tanto $F \vdash \neg A(\underline{n})$ para todo **n quanto** $F \vdash \exists x A(x)$. Naturalmente, isso implica a consistência normal, e segue-se da suposição de que os números naturais satisfazem os axiomas de F.

Na verdade, um caso especial simples de ω-consistência é aqui suficiente: a suposição é apenas necessária com respeito ao que os lógicos chamam de Σ_1^0-fórmulas; essas são, grosseiramente falando, as fórmulas existenciais puras; de modo mais preciso, fórmulas da forma $\exists x_1 \exists x_2 \ldots \exists x_n A$, em que A não contém qualquer quantificador livre (A pode conter quantificadores universais ligados $\forall x < t$ e existenciais ligados $\exists x < t$). Essa ω-consistência restrita é chamada de 1-consistência.

A ω-consistência e a 1-consistência são noções puramente sintáticas. Se o uso das noções de verdade e falsidade é permitido, a suposição de 1-consistência pode ser exprimida intuitivamente simplesmente como a exigência de que o sistema formal em questão não prove quaisquer Σ_1^0-sentenças falsas (isto é, o sistema é correto ao menos no caso de tais sentenças). De agora em diante, assume-se que os sistemas formalizados sob consideração contêm **Q** e são ao menos 1-consistentes, a não ser que seja de outro modo especificado.

2.2. Representabilidade

A prova de Gödel também requer a noção de representabilidade de conjuntos e relações no sistema formal F. Mais precisamente, duas noções relacionadas são necessárias.

Um conjunto *S* de números naturais é **fortemente representável** em F se há uma fórmula $A(x)$ da linguagem de F com uma variável livre x tal que para todo número natural **n**:

$$\begin{aligned} \mathbf{n} \in S \quad &\Rightarrow \quad F \vdash A(\underline{n}); \\ \mathbf{n} \notin S \quad &\Rightarrow \quad F \vdash \neg A(\underline{n}), \end{aligned}$$

Um conjunto *S* de números naturais é **fracamente representável** em *F* se há uma fórmula $A(x)$ na linguagem de F tal que para todo número natural **n**:

$$\mathbf{n} \in S \Leftrightarrow F \vdash A(\underline{n}).$$

É óbvio como todas essas noções são generalizadas para relações de muitos lugares. Há também noções relacionadas de representabilidade para funções. Como os resultados de incompletude em particular nos ensinam, há conjuntos que são apenas fracamente, mas não fortemente, representáveis (o exemplo central sendo o conjunto de afirmações prováveis no sistema).

[**Aviso:** Aqui a terminologia na literatura varia um pouco: "fortemente representa" é algumas vezes denominado por expressões como "representa", "exprime numericamente", "bi-numera", "define" ou "define fortemente"; "fracamente representa" é por sua vez exprimido por expressões como "representa", "define", "define fracamente" ou "numera". Deve-se ter cuidado aqui, focar nas definições relevantes e não se deixar enganar pelas palavras].

No caso de ambos os tipos de representabilidade (fraca e forte), há sempre uma Σ_1^0-fórmula existencial simples, que (fraca ou fortemente) representa o conjunto em questão, e usualmente tal fórmula é usada para representar *S*.

Embora essas noções sejam relativas ao sistema formal, mostrou-se que a representabilidade forte e fraca são extremamente estáveis. De modo bem independente do sistema formal particular escolhido, exatamente os conjuntos (relações) decidíveis, ou recursivos, são fortemente representáveis, e exatamente os conjuntos (relações) semi-decidíveis, ou recursivamente enumeráveis, são fracamente representáveis. Isso vale para todos os sistemas formalizados que contêm a aritmética de Robinson **Q**, da própria aritimética de Robinson até os sistemas de axiomas mais fortes da teoria dos conjuntos como **ZFC** e além

(enquanto forem (recursivamente) axiomatizáveis). Em vez de usar a noção de "representabilidade", Gödel fez uma abordagem diferente falando de conjuntos como sendo "decidíveis no sistema formal *F*" ("*entscheidungsdefinit*"). Se as provas de *F* são sistematicamente geradas, será eventualmente determinado, para qualquer número dado **n**, se ele pertence a *S* ou não - dado que *S* é fortemente representável em *F*.

Em suma, temos:

Teorema da Representabilidade:
Em qualqeur sistema formal consistente que contém **Q**:

1 Um conjunto (ou relação) é fortemente representável se, e somente se, é recursivo;
2 Um conjunto (ou relação) é fracamente representável se, e somente se, é recursivamente enumerável.

Ambas as noções de representabilidade - forte e fraca - têm de ser claramente distinguidas da mera **definibilidade** (no sentido padrão da palavra). Um conjunto *S* é **definível** na linguagem da aritmética se há uma fórmula $A(x)$ na linguagem tal que $A(\underline{n})$ é verdadeiro na estrutura padrão dos números naturais (a interpretação pretendida) se, e somente se, **n** $\in$ S. Há muitos conjuntos que podem ser definidos na linguagem da aritmética, mas não (mesmo fracamente) representados em algum F, tal como o conjunto das fórmulas consistentes, o conjunto de sentenças que não podem ser demonstradas no sistema F, ou o conjunto de equações Diofantinas sem soluções (*vide* abaixo).

2.3. Aritmetização da Linguagem Formal

O próximo passo essencial na prova de Gödel é tomar a linguagem de um sistema formal, que é sempre precisamente definida (isso é parte de ser um sistema formal), e fixar uma correspondência de um certo tipo entre as expressões dessa linguagem e o sistema de números naturais - uma codificação, "aritmetização", ou "numeração de Gödel", da linguagem. Há muitos modos possíveis de realizar essa tarefa, e os detalhes não importam realmente (para

mais detalhes de uma abordagem bem padrão, *vide* o complemento deste capítulo sobre a **Numeração de Gödel**). O ponto essencial é que o mapeamento escohido seja efetivo: é sempre possível passar, de um modo puramente mecânico, de uma expressão ao seu número de codificação, e do número à expressão correspondente. Hoje em dia, devido à familiaridade com computadores e com o fato de que tantas coisas podem ser coficiadas com os números 0 e 1, a possibilidade de tal aritmização dificilmente é surpreendente.

Grosso modo, o procedimento é o seguinte. Em primeiro lugar, os símbolos primitivos da linguagem são emparelhados com números naturais distintos, "números simbólicos". Um pouco de teoria dos números é então suficiente para codificar **sequências** de números por números singulares. Consequentemente, a cada uma das fórmulas bem formadas, como sequências de símbolos primitivos, é atribuído um único número. Finalmente, derivações, ou provas, do sistema, sendo sequências de formulas, são aritmetizadas, e elas recebem números específicos. Tal código, o "número de Gödel" de uma fórmula, é denotado como $\ulcorner A \urcorner$, e similarmente para derivações.

Desse modo, propriedades sintáticas, relações e operações são refletidas na aritmética: por exemplo, $neg(x)$ é a função aritmética que envia o número de Gödel de uma fórmula ao número de Gödel de sua negação; em outros termos, $neg(\ulcorner A \urcorner) = (\ulcorner \neg A \urcorner)$; similarmente, $impl(x, y)$ é a função que mapeia os números de Gödel de um par de fórmulas ao número de Gödel da implicação das fórmulas: $impl(\ulcorner A \urcorner, \ulcorner B \urcorner) = \ulcorner A \to B \urcorner$; e assim por diante. Há uma fórmula aritmética, chamada de $Fmla(x)$, que é verdadeira para **n** sse **n** é o número de Gödel de uma fórmula bem formada do sistema. Há também uma fórmula aritmética $M(x, y, z)$ que é verdadeira exatamente se temos uma aplicação válida da regra de inferência do *modus ponens* para algumas fórmulas A e B com $x = \ulcorner A \urcorner$, $y = \ulcorner A \to B \urcorner$ e $z = \ulcorner B \urcorner$; etc. Desse modo, todas as propriedades e operações sintáticas podem ser simuladas ao nível numérico, e, além disso, elas são fortemente representáveis em todas as teorias que contêm **Q**.

Como é decidível (pela definição de sistemas formais) se uma dada sequência de fórmulas constitui uma prova de uma dada sentença, de acordo com as regras do sistema formal *F* escolhido, a relação binária "x é (o número de

Gödel de) uma prova da fórmula (com o número de Gödel) *y*" pode ser fortemente representada em todos os sistemas que contêm **Q**, e assim em *F* em particular. Vamos denotar a fórmula que fortemente representa essa relação em *F* mesmo como $Prf_F(x, y)$. A propriedade de ser provável em *F* pode então ser definida como $\exists x Prf_F(x, y)$. Vamos abreviar esse **predicado de demonstrabilidade** formalizado como $Prov_F(x)$. Segue-se que este último é fracamente representável (embora, como se verificou, não fortemente):

$$F \vdash A \Rightarrow F \vdash Prov_F(\ulcorner A \urcorner).$$

Sempre é possível escolher o predicado de demonstrabilidade $Prov_F(x)$ como sendo uma Σ_1^0-fórmula.

2.4. Diagonalização ou "Auto-referência"

O próximo e talvez um tanto surpreendente ingrediente da prova de Gödel é o seguinte lema importante (ainda assumimos que *F* é um sistema formal que contém **Q**):

> **Lema da Diagonalização**
> Seja $A(x)$ uma fórmula arbitrária da linguagem de *F* com uma única variável livre. Então, uma sentença D pode ser mecanicamente construída, tal que:
>
> $$F \vdash D \leftrightarrow A(\ulcorner D \urcorner).$$
>
> (para um esboço da prova, *vide* o complemento deste capítulo sobre o Lema da Diagonalização)

Na literatura, esse lema é algumas vezes chamado de "lema da auto-referência" ou "lema do ponto fixo". Ele tem muitas aplicações importantes além dos teoremas da incompletude.

Diz-se com frequência que, dada uma propriedade denotada por $A(x)$, a sentença D é uma sentença auto-referencial que "diz acerca de si mesma" que ela tem a propriedade A. Tais figuras de linguagem podem ser heuristicamente

úteis, mas elas são também facilmente enganadoras e sugerem em excesso. Por exemplo, note que o lema só oferece uma equivalência material (provável) entre D e $A(\ulcorner D \urcorner)$ (que afirma que ambos os lados têm de ter o mesmo valor de verdade) e não afirma qualquer tipo de semelhança de significado. Em particular, D e $A(\ulcorner D \urcorner)$ não são de forma alguma idênticas - e nem o são $\ulcorner D \urcorner$ e $\ulcorner A(\ulcorner D \urcorner) \urcorner$.

2.5. O Primeiro Teorema da Incompletude - Prova Completada

Para completar a prova, o Lema da Diagonalização é aplicado ao predicado de demonstrabilidade negada $\neg Prov_F(x)$: isso nos dá a sentença G_F tal que:

(G) $$F \vdash G_F \leftrightarrow \neg Prov_F(\ulcorner G_F \urcorner).$$

Assim, pode-se mostrar, mesmo em *F*, que G_F é verdadeira se e somente se não é provável em *F*.

Não é difícil mostrar que G_F não é nem provável e nem refutável em *F*, se *F* é somente 1-consistente.

Para a primeira parte, assuma que G_F seja provável. Então, pela representabilidade fraca de **demonstrabilidade-em-F** por $Prov_F(x)$, *F* provaria também $Prov_F(\ulcorner G_F \urcorner)$. No entanto, porque *F* de fato também prova a equivalência (G), ou seja, $F \vdash G_F \leftrightarrow \neg Prov_F(\ulcorner G_F \urcorner)$, *F* então demonstraria $\neg G_F$ também. Mas isso significaria que *F* é inconsistente. Em suma, se *F* é consistente, então G_F não é provável em *F*. Para essa primeira parte, a suposição da simples consistência é suficiente.

Para a segunda parte, tem-se de assumir que *F* é 1-consistente (se $Prov_F(\ulcorner G_F \urcorner)$ foi escolhido de tal modo que é uma Σ_1^0-sentença; de outro modo, a suposição mais geral da ω-consistência é exigida).

Assuma que $F \vdash \neg G_F$. Então *F* não pode demonstrar G_F, pois, de outro modo, *F* seria simplesmente inconsistente. Por isso, nenhum número natural **n** é o número de Gödel de uma prova de G_F, e porque a relação de prova é fortemente representável, para todo **n**, $F \vdash \neg Prf_F(\underline{n}, \ulcorner G_F \urcorner)$. Se também $F \vdash \exists x Prf_F(x, \ulcorner G_F \urcorner)$, *F* não é 1-consistente, contrariamente à suposição.

Portanto, *F* não prova $\exists x Prf_F(x, \ulcorner G_F \urcorner)$; em outros termos, pela definição de $Prov_F(x)$, *F* não prova $Prov_F(\ulcorner G_F \urcorner)$. Pela equivalência (G), *F* também não prova $\neg G_F$.

O Primeiro Teorema da Incompletude de Gödel
Assuma que *F* é um sistema formalizado que contém a aritmética de Robinson **Q**. Então uma sentença G_F da linguagem de *F* pode ser mecanicamente construída de *F*, tal que:

i. Se *F* é consistente, então $F \nvdash G_F$.
ii. Se *F* é 1-consistente, então $F \nvdash \neg G_F$.

Uma tal afirmação independente ou "indecidível" (ou seja, nem provável e nem refutável em *F*) G_F em *F* é geralmetne chamada de "sentença de Gödel de *F*".

De fato, em circunstâncias favoráveis, pode-se mostrar que G_F é verdadeira, desde que *F* seja realmente consistente. Esse é o caso se, por exemplo, o predicado de demonstrabilidade $Prov_F(x)$ for escolhido como uma Σ_1^0-fórmula: a sentença de Gödel pode então ser provada como equivalente à fórmula universal $\forall x \neg Prof_F(x, \ulcorner G_F \urcorner)$. Tais fórmulas podem ser provadas falsas sempre que forem de fato falsas: se falsa, haveria um número **n** tal que $F \vdash Prof_F(\underline{n}, \ulcorner G_F \urcorner)$ (isso já vale para **Q**). Isso, no entanto, contradiria o teorema da incompletude. Portanto, G_F não pode ser falsa, e tem de ser verdadeira. Por essa razão, a sentença de Gödel é frequentemente chamada de "verdadeira mas não provável".

Não se deve ficar confuso sobre este ponto: "O Teorema de Gödel" é o resultado geral de incompletude que diz respeito a uma grande classe de sistemas formais, enquanto a "sentença de Gödel" é a sentença construída, formalmente indecidível, que varia de um sistema formal para outro. Essa é a razão por que é importante incluir o subscrito *F* em G_F. Além disso, não se deve confundir dois sentidos diferentes de "indecidível" nesse contexto. Por um lado, uma **sentença particular**, como a sentença de Gödel, pode ser indecidível no sentido de ser independente, ou seja, nem provável e nem refutável no sistema escolhido. Por outro lado, uma **teoria** pode ser indecidível (*vide* abaixo) no

sentido de que não há um método de decisão para determinar, a respeito de uma dada sentença arbitrária da linguagem, se ela é derivável na teoria (assim, este último sentido de "indecidível" diz respeito, por assim dizer, a uma classe infinita de afirmações).

Em explicações informais do primeiro teorema da incompletude, frequentemente se diz que a sentença de Gödel G_F "diz de si mesma que ela não é provável". Essas afirmações imprecisas, no entanto, não deveriam ser tomadas literalmente. Há inúmeras razões para concluir que, ao menos em geral, as sentenças de Gödel não dizem realmente algo substancial acerca de si mesmas (MILNE, 2007, é uma análise cuidados dessas questões); por exemplo, como anteriormente notamos sobre o Lema da Diagonalização, usualmente trabalha-se aqui com equivalências materiais.

O Desenvolvimento de Rosser - da ω-consistência à Consistência

Em 1936, J. Barkley Rosser fez um avanço importante que tornou possível nos livrarmos da suposição desajeitada da ω-consistência na prova do **primeiro** teorema de Gödel. Para esse fim, Rosser introduziu um "predicado de demonstrabilidade" novo, e em certa medida artificial, $Prov^*(x)$, que era construído, informalmente, do seguinte modo:

> **Existe** um *y* tal que *y* é o número de Gödel de uma prova da fórmula com o número de Gödel *x*, e **não existe um z** menor do que *y* tal que *z* é o número de Gödel de uma prova da negação da fórmula com o número de Gödel *x*.
>
> Mais formalmente:

$$Prov^*(x) =_{def} \exists y[Prf_F(y,x) \wedge \forall z < y(\neg Prf_F(z, neg(x)))],$$

em que $Prf_F(y,x)$ é a relação de prova padrão discutida anteriormente.

Se o sistema formal *F* sob consideração é realmente consistente, o predicado de demonstrabilidade de Rosser é co-extensional com o predicado

ordinário de demonstrabilidade. Aplicando o Lema da Diagonalização à negação do predicado de demonstrabilidade de Rosser $Prov^*(x)$, obtemos:

> **A Modificação de Rosser do Primeiro Teorema (Rosser 1936**)
> Seja *F* um sistema consistente formalizado que contém **Q**. Então há uma sentença R_F da linguagem de *F* tal que nem R_F nem $\neg R_F$ é provável em *F*.

2.6. Incompletude e Modelos não-standard

É esclarecedor refletir sobre o primeiro teorema da incompletude também de uma perspectiva da teoria de modelos - embora o teorema em si mesmo de modo nenhum exija tal coisa. É possível concluir que qualquer teoria *F* que satisfaz as condições que o teorema coloca deve possuir, além da interpretação pretendida ou "modelo padrão" (no caso da teorias aritiméticas, a estrutura dos números naturais), interpretações não pretendidas ou "modelos não standard" - que nenhuma dessas teorias pode excluir estes últimos e fixar univocamente a interpretação pretendida. Se há afirmações independentes tais como G_F, F tem de ter modelos que satisfazem G_F e modelos que pelo contrário satisfazem $\neg G_F$. Como $\neg G_F$ é equivalente a $\exists x Prf_F(x, \ulcorner G_F \urcorner)$, os últimos modelos têm de ter entidades que satisfazem à fórmula $Prf_F(x, \ulcorner G_F \urcorner)$. E ainda sabemos (visto que $Prf_F(x, y)$ fortemente representa a relação de prova) que para qualquer numeral $\underline{n}$, *F* pode provar $\neg Prf_F(\underline{n}, \ulcorner G_F \urcorner)$. Portanto, nenhum número natural **n** pode servir de testemunha para a fórmula. Segue-se que qualquer modelo não standard como esse tem de conter, além dos números naturais (denotações dos numerais $\underline{n}$), "infinitos" números não naturais depois dos números naturais.

O estudo dos modelos não standard não começou com os resultados de Gödel - Skolem, em particular, já estava anteriormente ciente desses modelos em um contexto diferente (ele descobriu que as teorias de primeira ordem da teoria dos conjuntos têm modelos reduzidos não naturais, que são os modelos

contáveis, em SKOLEM, 1922; *vide* o verbete **Skolem's Paradox**[17] da SEP) - mas o primeiro teorema da incompletude esclarece a existência de modelos não standard no contexto da aritmética, enquanto os modelos não standard esclarecem o primeiro teorema da incompletude. Modelos não standard têm se tornado desde então um área rica na lógica matemática (*vide* BOOLOS; JEFFREY, 1989: cap. 17; KAYE, 1991).

3. O Segundo Teorema da Incompletude

3.1. Preliminares

Informalmente, o raciocínio que conduz ao segundo teorema da incompletude é relativamente simples. Dado o predicado da demonstrabilidade aritmetizado, é também fácil apresentar uma afirmação de consistência aritmetizada: escolha alguma fórmula manifestadamente inconsistente (nas teorias aritméticas, uma escolha padrão é $(\underline{0} = \underline{1})$; vamos denotá-la por $\bot$; (a contraparte artimetizada de) a consistência do sistema pode ser então definida como $\neg Prov_F(\ulcorner \bot \urcorner)$. Vamos abreviar essa fórmula por $Cons(F)$. A prova da primeira parte do primeiro teorema da incompletude (isto é, o caso (i) acima) pode ser presumivelmente formalizada em *F* (na prática, isso certamente seria intrincado). Obtemos então:

$$F \vdash Cons(F) \rightarrow G_F,$$

em que G_F é a sentença de Gödel para F oferecida pelo primeiro teorema. Se $Cons(F)$ fosse provável em *F*, então G_F também seria provável, por lógica simples. Isso contradiria o primeiro teorema de Gödel. Consequentemente, $Cons(F)$ também não pode ser provável em *F*.

> **O Segundo Teorema da Incompletude de Gödel**
> Assuma que *F* é um sistema consistente formalizado que contém a aritmética elementar. Então $F \nvdash Cons(F)$.

[17] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/paradox-skolem/. Acesso em: 20 jan. 2022.

Há uma questão filosófica importante que deveria ser aqui mencionada: como se encontra, o segundo teorema da incompletude de Gödel só estabelece a não demonstrabilidade de **uma** sentença, *Cons(F)*. Mas essa sentença realmente exprime que *F* é consistente? (Compare essa questão com a observação acima de que G_F não exprime, estritamente falando, a sua própria não demonstrabilidade). Além disso, não poderia haver **outras** sentenças que são prováveis e também exprimem a consistência de *F*?

Oferecer uma prova rigorosa do segundo teorema em uma forma mais geral cobrindo todas essas sentenças, no entanto, mostrou-se muito complicado. A razão básica para isso é a seguinte: diferentemente do primeiro teorema, não é qualquer predicado de demonstrabilidade que seja meramente extensionalmente adequado que funciona para a formalização da afirmação de consistência. O modo de apresentação faz toda a diferença. Por exemplo, o predicado de demonstrabilidade de Rosser mencionado acima não serviria; pode-se provar a "consistência" de *F* em *F*, se a consistência é exprimida nos termos do predicado de demonstrabilidade de Rosser. Tem-se então de acrescentar algumas condições para o predicado de demonstrabilidade para a prova do segundo teorema da incompletude seguir em frente. Seguindo Feferman (1960), costuma-se dizer que enquanto o primeiro teorema e seus relativos são resultados **extensionais**, o segundo teorema é **intensional**: tem de ser possível pensar que *Cons(F)* em algum sentido **exprime** a consistência de *F* - que ele realmente **significa** que *F* é consistente.

3.2. Condições de Derivabilidade

A prova do segundo teorema da incompletude requer que o predicado de demonstrabilidade em *F* satisfaça algumas condições que são usadas nos detalhes da prova. Há vários conjuntos diferentes de condições que servem a esse propósito.

A primeira prova detalhada do segundo teorema da incompletude apareceu em Hilbert e Bernays 1939 (principalmente redigido por Bernays), embora para só uma teoria específica, **PA**. Ela usa um complicado conjunto de condições para o predicado de demonstrabilidade. Essas condições eram mais

lemas técnicos requeridos para uma prova particular do que algum tipo de análise dos predicados de demonstrabilidade "naturais". Uma lista mais elegante, e agora padrão, de "condições de derivabilidade" foi apresentada por Löb (1955) - embora o seu uso pretendido fosse um pouco diferente (*vide* abaixo).

As Condições de Derivabilidade de Löb

(D1) $F \vdash A \Rightarrow F \vdash Prov_F(\ulcorner A \urcorner)$.
(D2) $F \vdash Prov_F(\ulcorner A \urcorner) \rightarrow (\ulcorner Prov_F(\ulcorner A \urcorner) \urcorner)$.
(D3) $F \vdash Prov_F(\ulcorner A \urcorner) \wedge Prov_F(\ulcorner A \rightarrow B \urcorner) \rightarrow Prov_F(\ulcorner B \urcorner)$.

(D1) é simplesmente a reafirmação da exigência da prova do primeiro teorema de que a demonstrabilidade seja fracamente representável. *Grosso modo*, (D2) exige que a totalidade da demonstração de (D1) para o candidato ao predicado de provabilidade $Prov_F$ possa ser ela mesma formalizada em *F*. Finalmente, (D3) exige que o predicado de demonstrabilidade seja fechado sob o *Modus Ponens*.

Se a aritmetização do predicado de demonstrabilidade realmente satisfizer essas condições, o segundo teorema pode ser provado. Seja G_F mais uma vez a sentença de Gödel para *F* dada no primeiro teorema. Não é difícil mostrar, usando as condições de derivação, que:

$$F \vdash G_F \leftrightarrow Cons(F).$$

Isso imediatamente produz a não demonstrabilidade de $Cons(F)$, dado o primeiro teorema da incompletude.

Além disso, Jeroslow (1973) demonstrou, com um artifício engenhoso, que é de fato possível estabelecer o segundo teorema sem (D3). No entanto, em alguns outros casos (como quando se prova o teorema de Löb; *vide* abaixo), e na Lógica da demonstrabilidade, todas as três condições são ainda requeridas.

3.3. A Abordagem Alternativa de Feferman do Segundo Teorema

Sob a suposição de que o predicado de demonstrabilidade para uma teoria satisfaz as condições de derivabildiade (ou, pelo artifício de Jeroslow, ao menos D1 e D2), é relativamente simples provar o caso relevante do segundo teorema da incompletude. No entanto, na prática temos de estabelecer caso a caso se um predicado de demonstrabilidade aritmetizado proposto realmente satisfaz as condições, e tipicamente esse procedimento é longo e tedioso.

Essa desvantagem, entre outras coisas (vide FEFERMAN, 1997), levou Solomon Feferman no final da década de 1950 a buscar uma linha alternativa de ataque ao segundo teorema (*vide* FEFERMAN, 1960). Feferman aborda a questão em dois passos. Primeiramente, ele isola a fórmula $Prov_{FOL}(x)$ que aritmetiza alguma noção padrão de derivabilidade na **lógica de primeira ordem** a fim de permitir-nos fixar uma fórmula escolhida para a demonstrabilidade na lógica. Como o conjunto de axiomas não lógicos do sistema em questão são apresentados é deixado em aberto nesse estágio. Em segundo lugar, Feferman busca uma restrição adequada para apresentar os axiomas. Entre as fórmulas da linguagem da aritmética, ele isola o que ele chama de PR- e RE-fórmulas. As primeiras correspondem às definições canônicas primitivas-recursivas (PR) na aritmética, e as segundas às generalizações existenciais das primeiras. Todo conjunto recursivamente enumerável (RE) pode ser definido por uma fórmula do segundo tipo; essas são justamente as Σ_1^0-fórmulas. Essas duas classes são facilmente discriminadas puramente por suas formas sintáticas. (De fato, pelo Teorema MRDP (*vide* abaixo), pode-se - ao invés das RE-fórmulas - focar em casos mais simples de equações diofantinas existencialmente quantificadas).

Acima, notamos o fato importante de que, em todas as teorias aritméticas *F* contendo **Q**, um conjunto é fortemente representável em *F* se, e somente se, ele é recursivo, e um conjunto é recursivamente enumerável se, e somente se, ele é fracamente representável. Além disso, pode-se sempre tomar a fórmula representando fraca ou fortemente o conjunto como sendo uma RE-fórmula (isto é, uma Σ_1^0-fórmula; e, pelo teorema MRDP, mesmo uma equação diofantiana existencialmente quantificada). É então natural exigir que o conjunto de axiomas não lógicos do sistema em questão seja representado por tal

fórmula. Se a definição artimetizada do conjunto dos números de Gödel de axiomas refletir como os axiomas, se infinitos, são indutivamente definidos, a fórmula resultante será Σ_1^0. (Para teorias que são axiomatizáveis com um número finito de axiomas, há uma única representação dos axiomas na forma de uma lista, e, consequentemente, uma única afirmação de consistência relativa a $Prov_{FOL}(x)$.) Em vez de determinar se as condições de derivabilidade são satisfeitas, é uma tarefa relativamente rotineira determinar que uma dada fórmula que formaliza os axiomas é realmente da forma exigida (Σ_1^0).

Agora a versão do segundo teorema da incompletude apresentada em Feferman 1960 é:

> **Uma Variante do Segundo Teorema da Incompletude (FEFERMAN, 1960)**
> Seja *F* uma extensão consistente de **PA**, seja $Ax_F(x)$ uma Σ_1^0-fórmula que fracamente representa os axiomas *F*, e *Cons(F)* uma afirmação de consistência construída a partir de $Ax_F(x)$ e $Prov_{FOL}(x)$. Então $Cons(F)$ não é provável em *F*.

Para outras abordagens diferentes ao segundo teorema da incompletude, *vide* Feferman, 1982, 1989a; Visser, 2011. Para algumas complicações filosóficas a respeito do segundo teorema, *vide* DETLEFSEN, 1979, 1986, 1990, 2001; AUERBACH, 1985, 1992; ROEPER, 2003; FRANKS, 2009 (*vide* também a seção sobre incompletude no verbete **Hilbert's Program**[18] da SEP).

4. Resultados Relacionados aos Teoremas da Incompletude

4.1. O Teorema da Indefinibilidade de Tarski

Primeiramente, Gödel chegou nos resultados de incompletude (*vide* **Seção 5** abaixo) observando que a verdade (da linguagem de um sistema) deve

[18] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/hilbert-program/index.html#4. Acesso em: 20 jan. 2022.

ser indefinível no sistema, um resultado convencionalmente atribuído a Tarski (há certas virtudes reais na maneira de Tarski de apresentar a questão; *vide* GÖMEZ TORRENTE, 2004). Vejamos agora o resultado no contexto da abordagem de Tarski à verdade.

Tarski distinguiu claramente a linguagem objeto, ou seja, a linguagem das sentenças cuja verdade está em jogo e a metalinguagem na qual a primeira é discutida. Ele também exigia que (*vide* o verbete **Tarski's Truth Definitions**[19] na SEP) qualquer definição satisfatória da verdade $T(x)$ para a linguagem objeto deveria satisfazer a sua "Convenção T", ou seja, que a definição deveria ter como consequências todas as equivalências ("Equivalências T") da forma

$$(T) \qquad T(\ulcorner A \urcorner) \leftrightarrow B,$$

em que $\ulcorner A \urcorner$ é um nome da sentença da linguagem objeto, e B é a sua tradução na metalinguagem.[20] Se a metalinguagem coincide com a linguagem objeto ou se é uma extensão dela, então B é simplesmente a sentença A, **ela mesma**, e as equivalências-T são da forma:

$$T(\ulcorner A \urcorner) \leftrightarrow A$$

O que o teorema da indefinibilidade nos mostra é que a linguagem objeto e a metalinguagem não podem coincidir, que elas devem ser linguagens distintas.

> **O Teorema da Indefinibilidade de Tarski**
> Seja F um sistema consistente formalizado que contém um fragmento suficiente da aritmética. Então não existe uma fórmula $T(x)$ na linguagem de F, tal que, para toda sentença A da linguagem de F:

[19] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/tarski-truth/. Acesso em: 20 jan. 2022.

[20] N.T.: Optamos por manter a letra "T" como notação do predicado-verdade, ao invés da alternativa evidente para o português, "V". O uso da notação "T" neste contexto já foi bastante difundido também na língua portuguesa. Em alguns trechos do original em inglês, o autor utiliza ainda "$True(x)$" e "$Tr(x)$". Nestes casos, mantivemos a notação "$T(x)$" para evitar confusão com a opção de tradução.

$$F \vdash T(\ulcorner A \urcorner) \leftrightarrow A.$$

A ideia da prova: Se houvesse uma tal fórmula da linguagem de F, uma simples aplicação do Lema da Diagonalização à sua negação resultaria na sentença paradoxal M (de "Mentiroso"; *vide* o verebete **Liar Paradox**[21] na SEP), tal que:

$$F \vdash \neg T(\ulcorner M \urcorner) \leftrightarrow M,$$

o que, juntamente com as equivalências-T, assumidas como deriváveis, levaria rapidamente a uma explícita contradição, contradizendo assim a suposição de que F é consistente

Similarmente, pode ser demonstrado que o conjunto das sentenças verdadeiras de F não é definível na interpretação pretendida de F - no sentido já padronizado de "definibilidade" (*vide* acima).

4.2. Os Resultados de Indecibilidade

As ferramentas utilizadas nas provas dos teoremas de Gödel também fornecem vários e importantes resultados de indecidibilidade. Uma teoria é dita **decidível** se o conjunto dos seus teoremas (sentenças deriváveis na teoria) é decidível, isto é (pela tese Church-Turing), recursivo. De outro modo, a teoria é indecidível. Informalmente, que ela seja decidível significa que existe um procedimento mecânico que nos habilita a decidir se uma dada sentença arbitrária (da linguagem da teoria) é ou não um teorema.

Se uma teoria é completa, ela é decidível (esboço da prova: dada uma sentença A, produza sistematicamente os teoremas da teoria. Pela completude, fatalmente, A ou $\neg A$ será produzida em um tempo finito). A conversa, entretanto, não vale sempre: existem teorias incompletas que são decidíveis. Todavia, a

[21] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/liar-paradox/. Acesso em: 20 jan. 2022.

incompletude ao menos torna possível a indecidibilidade. Ademais, todas as teorias que contém a aritmética de Robinson **Q** (contém diretamente **Q** ou **Q** pode ser interpretada nelas) são tanto incompletas quanto indecidíveis. Portanto, para uma classe muito ampla de teorias, a incompletude e a indecibilidade caminham lado a lado.

Um modo elegante e simples de demonstrar a indecidibilidade das extensões de **Q** é o seguinte: seja F uma teoria consistente qualquer que contém **Q**. Suponha que o conjunto dos seus teoremas é decidível, ou seja (pela tese Church-Turing), recursivo. Seguir-se-ía então que o conjunto (dos números de Gödel) dos teoremas de F seria **fortemente representável** em F, ela mesma. Lembre-se, isto significa que existe uma fórmula $B(x)$ da linguagem de F, tal que, não apenas $F \vdash B(\ulcorner A \urcorner)$ quando $F \vdash A$ (o que é garantido mesmo pela representabilidade fraca), mas ainda que $F \vdash \neg B(\ulcorner A \urcorner)$ quando $F \nvdash A$. No entanto, a técnica utilizada na prova do primeiro teorema da incompletude também mostra que sempre existem sentenças para as quais esta última consequência não vale: é possível construir uma sentença de Gödel G^B relativa a $B(x)$ para F, tal que:

$$(D) \qquad F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B \urcorner).$$

Como antes, segue-se que $F \nvdash G^B$. Foi antes assumido que $B(x)$ representa fortemente o conjunto dos teoremas, isso então acarreta que $F \vdash \neg B(\ulcorner G^B \urcorner)$, e portanto, por (D), $F \vdash G^B$, uma contradição. Logo, F deve ser indecidível.

Uma teoria F é dita **essencialmente indecidível** se toda extensão consistente dela na linguagem de F é indecidível. Em verdade, o esboço de prova acima estabelece que **Q** é essencialmente indecidível. (Existem algumas teorias muito fracas que são indecidíveis mas que não são essencialmente indecidíveis.)

Lembre-se que **Q** tem apenas finitamente muitos axiomas e tome A_Q por uma única sentença consistindo em uma conjunção dos axiomas de **Q**. Então, para qualquer sentença B da linguagem da aritmética,

> **Q** $\vdash B$ se, e somente se, $A_Q \rightarrow B$ é um teorema da lógica de primeira ordem.

Mas então um procedimento de decisão para a lógica de primeira ordem forneceria um método de decisão para **Q**. Como já foi mostrado, tal método é impossível. Portanto, pode-se concluir o seguinte:

Teorema de Church
A lógica de primeira ordem é indecidível.

(Este resultado de indecidibilidade foi primeiramente estabelecido por CHURCH, 1936a,b; o método de derivá-lo pela indecidibilidade de **Q** deve-se a TARSKI, MOSTOWSKI e ROBINSON, 1953).

Subsequentemente, várias teorias e problemas de diferentes áreas da matemática mostraram-se indecidíveis (*vide*, por exemplo, DAVIS, 1977; MURAWSKI, 1999, Cap. 3).

4.3. Princípios de Reflexão e o Teorema de Löb

Heuristicamente, pode-se ver a sentença de Gödel G_F como exprimindo a sua própria indemonstrabilidade - como dizendo "eu não sou provável" -; no entanto, como já foi enfatizado, tais afirmações não deveriam ser interpretadas literalmente. Leon Henkin levantou a questão de se a sentença exprimindo a sua própria indemonstrabilidade ("eu não sou provável") é verdadeira ou falsa, e provável ou não (HENKIN, 1952). Georg Kreisel apontou em seguida que a questão depende vitalmente de como se exprime demonstrabilidade; diferentes escolhas dão respostas opostas (KREISEL, 1953).

O artigo de Martin Löb (1955), ampliado pelos comentários de um parecerista, trouxe avanços substanciais em várias frentes. Em primeiro lugar, ele introduz as, hoje padronizadas, condições de derivabilidade, anteriormente discutidas no contexto do segundo teorema da incompletude. Em segundo lugar, o artigo contém a solução de Löb para o problema de Henkin acerca das sentenças que "exprimem a sua própria indemonstrabilidade". Em terceiro lugar, o artigo contém uma generalização hoje conhecida como "Teorema de Löb", mas que Löb creditava ao parecerista anônimo (este era niguém menos que o próprio Henkin; a história completa é narrada em Smoryński (1991)).

No intuito de compreender adequadamente o Teorema de Löb, é útil

considerar os assim denominados “princípios de reflexão”. Acima, o foco tinha sido em expressar, dentro de um sistema formal, que o sistema formal é consistente, isto é, em $Cons(F)$. Naturalmente, a teoria não deveria ser meramente consistente, mas também **correta**, isto é, provar apenas sentenças verdadeiras. Como a correção de um sistema, isto é, a alegação de que tudo aquilo que é derivável no sistema é vedadeiro, deveria ser expressa? Se queremos expressar tal coisa na própria linguagem do sistema, não podemos fazê-lo por meio de uma única afirmação, pois, pela indefinibilidade da verdade, não existe um predicado-verdade adequado disponível na linguagem. Entretanto, várias alegações de correção restritas e irrestritas podem ser expressas na forma de um esquema, os assim denonimados Princípios de Reflexão:

(Ref) $$Prov_F(\ulcorner A \urcorner) \rightarrow A.$$

Substituindo A por $\perp$ no esquema e considerando que $\perp$ é refutável em F, é fácil de ver que o Princípio de Reflexão acarreta a afirmação de consistência $Con(F)$, isto é, $\neg Prov_F(\ulcorner \perp \urcorner)$; logo, ele não pode ser demonstrado de forma generalizada no sistema.

O esquema também pode ser restrito. Equivalente à suposição de 1-consistência ou Σ^0_1-correção, por exemplo, o Princípio de Reflexão é restrito às Σ^0_1-sentenças (isto é, exige-se que a sentença A no esquema seja uma Σ^0_1-sentença). Ou ele pode ser restrito às Π^0_1-sentenças universais; e assim por diante.

Exatamente quais instâncias do Princípio de Reflexão são de fato prováveis no sistema? O Teorema de Löb dá uma resposta precisa para esta questão (assumindo que $Prov_F(x)$ satisfaça as condições de derivabilidade):

Teorema de Löb
Seja A uma sentença qualquer da linguagem de F. Então: $F \vdash Prov_F(\ulcorner A \urcorner) \rightarrow A$ se, e somente se, $F \vdash A$.

Portanto, as instâncias da correção (princípio de reflexão) prováveis em um sistema são exatamente aquelas que concernem às sentenças que são, elas mesmas, prováveis no sistema. Consequentemente, isso também resolve o

problema original de Henkin: assumindo que o predicado de demonstrabilidade aritmetizado seja "normal" (isto é, satisfaça as condições de derivabilidade), todas as sentenças "exprimindo a sua própria demonstrabilidade" são prováveis.

Em verdade, o Teorema de Löb pode ser provado rapidamente como uma consequência do segundo teorema da incompletude. Kreisel também notou que, na direção oposta, o segundo teorema da incompletude pode ser facilmente derivado como uma consequência do teorema de Löb.

4.4. O Décimo Problema de Hilbert e o Teorema MRDP

O décimo ítem na famosa lista de Hilbert dos problemas em aberto na matemática em 1900 pede por um método de decisão para as assim denominadas equações Diofantinas. Embora o termo "Diofantina" não seja familiar, o que está em jogo aqui é de fato elementar. Considere qualquer equação com uma ou mais variáveis, coeficientes inteiros e que envolva apenas adição e multiplicação, tais como $x^2 + y^2 = 2$ e $3x^2 + 5y^2 + 2xy = 0$. Se soluções com números reais são requisitadas, em geral fala-se simplesmente em uma "equação". Entretanto, na teoria dos números é tipicamente requisitado que uma solução consista apenas de inteiros. Isto faz uma grande diferença. A primeira das equações acima tem infinitamente muitas soluções entre os números reais, mas apenas quatro entre os inteiros. A equação $x^2 + y^2 = 3$ também tem infinitas soluções entre os reais, mas nenhuma solução entre os inteiros. Quando o foco é nas soluções com inteiros, fala-se em "equações Diofantinas" (em homenagem ao antigo teórico dos números Diofanto de Alexandria).

Para uma solução positiva do décimo problema de Hilbert, bastaria apresentar um método concreto particular que teria sido intuitivamente um método de decisão "mecânico". No entanto, a análise pioneira de Turing da noção de método de decisão colocou em foco a possibilidade de uma solução negativa. A partir do início dos anos 1950, Julia Robinson e Martin Davis trabalharam nesse problema, mais tarde acompanhados por Hilary Putnam. Como resultado de sua colaboração, o primeiro resultado importante nesse sentido foi alcançado. Chame uma equação de "equação diofantina exponencial" se ela envolver também exponenciação, bem como adição e multiplicação (ou seja, pode-se ter

constantes e variáveis como expoentes); naturalmente, o foco ainda está nas soluções inteiras. **D**avis, **P**utnam e **R**obinson (1961), mostraram que o problema da solubilidade de equações diofantinas exponenciais é indecidível. Em 1970, Yuri **M**atiyasevich acrescentou a última peça que faltava e demonstrou que o problema da solubilidade das equações diofantinas é indecidível. Portanto, o resultado geral é frequentemente chamado de Teorema MRDP (para uma exposição, *vide*, por exemplo, DAVIS, 1973; MATIYASEVICH, 1993).

A essência dessa façanha técnica é que a todo conjunto semidecidível (recursivamente enumerável) pode se dada uma representação Diofantina, ou seja, eles podem ser representados por uma fórmula simples da forma $\exists x_1, ..., x_n(s = t)$, em que $(s = t)$ é uma equação Diofantina. Mais exatamente, dado um conjunto recursivamente enumerável qualquer S, existe uma equação Diofantina $(s(y, x_1, ..., x_n) = t(y, x_1, ..., x_n))$, tal que, **n** $\in S$ se, e somente se, $\exists x_1 ... \exists x_n(s(\underline{n}, x_1, ..., x_n) = t(\underline{n}, x_1, ..., x_n))$.

Na medida em que existem conjuntos semidecidíveis (recursivamente enumeráveis) que não são decidíveis (recursivos), a conclusão geral segue-se imediatamente:

> **Teorema MRDP**
> Não existe um método geral para decidir se uma dada equação Diofantina tem solução ou não.

Isso também fornece uma variante elegante dos teoremas da incompletude em termos de equações Diofantinas:

> **Corolário**
> Para qualquer sistema formal axiomatizável 1-consistente F, existem equações Diofantinas insolúveis, mas cuja insolubilidade não pode ser provada em F.

(A questão de evitar a exigência da 1-consistência aqui é intrincada; *vide* DYSON, JONES; SHEPHERDSON, 1982).

4.5. Casos Concretos de Sentenças Indemonstráveis

As sentenças indecidíveis fornecidas pelas provas de Gödel são (tal como escritas) fórmulas extremamente complicadas e sem qualquer significado intuitivo, construídas apenas para os propósitos das provas de incompletude. Levanta-se então a questão de se existem quaisquer afirmações matemáticas simples e naturais igualmente indecidíveis nas teorias básicas escolhidas, por exemplo, em **PA**. Existem hoje várias afirmações matemáticas específicas, de conteúdo claro e que são reconhecidamente indecidíveis em algumas teorias padrão (no entanto, tem sido disputado o quão naturais são essas afirmações; *vide* FEFERMAN, 1989b). Alguns exemplos naturais e bem conhecidos são listados abaixo, começando com algumas afirmações matemáticas bastante naturais que são independentes de **PA** e prosseguindo para teorias cada vez mais poderosas. Ocasionalmente, tais resultados são designados como variantes dos teoremas de Gödel; ou suas provas de independência como provas alternativas dos teoremas de Gödel, mas isso é um erro: por mais interessante que sejam, esses resultados não têm a generalidade dos teoremas de Gödel propriamente ditos, eles apenas fornecem afirmações independentes de uma teoria particular.

Tem sido frequentemente afirmado que, antes do celebrado teorema de Paris-Harrington (*vide* abaixo), não eram conhecidas afirmações matemáticas naturais independentes. Estritamente falando, contudo, isso é incorreto. Muito antes, por volta de 1935, Gerhard Gentzen (*vide* o verbete **The Development of Proof Theory**[22] da SEP) havia fornecido uma tal afirmação. É muito natural generalizar a ideia de indução do domínio dos números naturais para o domínio dos números ordinais. Na teoria de conjuntos, tais generalizações são denominadas princípios de indução transfinita. Muito embora alguns construtivistas possam ser céticos acerca da legitimidade da teoria de conjuntos plena, existem casos mais concretos e limitados de indução transfinita (nos quais lida-se apenas com alguma classe bem definida de ordinais **contáveis**) perfeitamente aceitáveis do ponto de vista construtivista ou intuicionista. Um caso importante é o princípio da indução transfinita até o ordinal ε_0. Gentzen mostrou

[22] N.T.:Disponível em: https://plato.stanford.edu/archives/win2021/entries/proof-theory-development. Acesso em: 20 jan. 2022.

que a consistência de **PA** pode ser demonstrada se este princípio de indução transfinita for assumido. Logo, em virtude do segundo teorema da incompletude, o princípio ele mesmo não pode ser demonstrável em **PA** (GENTZEN, 1936).

O teorema de Ramsey é um resultado em combinatória infinitária, estabelecido por Frank Ramsey (1930), e lida com possibilidades de "coloração" para certos grafos. Jeff Paris e Leo Harrington formularam uma variante finitária do teorema de Ramsey e mostraram que ela não era demonstrável em **PA** (PARIS; HARRINGTON, 1977). Isso nos dá uma afirmação bastante natural de combinatorial finita que é independente de **PA**. Um exemplo talvez ainda mais limpo é o teorema de Goldstein, devido a Reuben Goldstein (1944), cuja natureza é pura teoria dos números. Primeiro define-se uma certa classe de sequências de números naturais, agora conhecidas como "sequências de Goodstein". O teorema afirma que toda sequência de Goodstein cedo ou tarde termina em 0. O teorema de Goodstein é certamente uma afirmação matemática natural, pois foi formulada e demonstrada (obviamente por métodos de prova que vão além de **PA**) por Goodstein muito antes (isto é, em 1944) de ser mostrado, em 1982, que o teorema não é demonstrável em **PA** (KIRBY; PARIS, 1982).

Em se tratando de teorias mais fortes, para além de **PA**, pode-se mencionar, por exemplo, o Teorema de Kruskal. Esse é um teorema que diz respeito a certas ordenações de árvores finitas (KRUSKAL, 1960). Harvey Friedman demonstrou que este teorema é indemonstrável mesmo em subsistemas de aritméticas de segunda ordem muito mais fortes que **PA** (*vide* SIMPSON, 1985). Em particular, o teorema não é demonstrável em qualquer teoria que seja predicativamente justificável (sob uma explicação de "predicativo" amplamente aceita; *vide* seção sobre predicativismo no verbete **Philosophy of Mathematics**[23] da SEP).

Existem alguns exemplos concretos de afirmações matemáticas indecididas mesmo em teorias mais fortes advindas da assim denominada teoria descritiva dos conjuntos. Esta área da matemática está relacionada com a topologia e foi iniciada pelos semi-intuicionistas franceses (Lebesque, Baire, Borel; *vide* a seção sobre teoria descritiva dos conjuntos, etc., no verbete sobre o

[23]N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/philosophy-mathematics/index.html#Pre. Acesso em: 20 jan. 2022.

Intuitionism in the Philosophy of Mathematics[24] da SEP). Nessa área, estudam-se conjuntos que possuem definições relativamente simples (em constradistinção às ideias de conjuntos arbitrários e de vários conjuntos potência superiores), denominados conjuntos projetivos ou conjuntos analíticos. Classicamente, esses conjuntos foram definidos como conjuntos que podem ser construídos de uma interseção contável de conjuntos abertos tomando-se, finitamente muitas vezes, imagens contínuas e complementos; eles coincidem com os conjuntos que são definíveis na linguagem de $\mathbf{P}^2$. Em particular, os assim denominados conjuntos de Borel podem ser definidos tanto por uma fórmula da forma $\exists X A(x)$, quanto por uma fórmula da forma $\forall X B(x)$, em que A e B não possuem quaisquer variáveis de conjuntos (na terminologia dos lógicos, conjuntos de Borel são conjuntos Δ^1_1). Uma função de Borel é definida de modo análogo (*vide*, por exemplo, MARTIN, 1977).

Harvey Friedman estabeleceu o seguinte teorema: *grosso modo*, se S é um conjunto de Borel, então existe uma função de Borel f tal que ou o grafo de f está incluído em S ou é disjunto de S. Muito embora esse teorema soe tão simples, Friedman mostrou que ele não é demonstrável, mesmo na aritmética de segunda ordem plena $\mathbf{P}^2$, mas que a sua prova requer necessariamente toda a força de **ZFC** (*vide* SIMPSON, 1999: 23).

Mais além, uma questão tradicional da teoria descritiva de conjuntos (uma questão que pode ser formulada na linguagem da aritmética de segunda ordem) era se todos os conjuntos projetivos (*vide* acima) são Lebesque mensuráveis. Isso permaneceu como um problema aberto por muitas décadas, e por uma boa razão: acontece que a afirmação é independente mesmo da teoria de conjuntos plena **ZFC** (*vide* SOLOVAY 1970). Apenas pela postulação da existência de cardinais extremamente grandes (os assim denominados cardinais de Woodin), a hipótese de que todos os conjuntos projetivos são Lebesque mensuráveis pode ser demonstrada (isso foi estabelecido como uma consequência do seu trabalho na assim denominada determinação projetiva, por Woodin, Martin e Steel; *vide* WOODIN, 1988; MARTIN; STEEL,1988,1989).

Eventualmente, inclui-se ainda nessa lista o celebrado resultado de Paul

[24] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/intuitionism/index.html#DesSetTheTopTopThe. Acesso em: 20 jan. 2022.

Cohen de que a Hipótese do Contínuo (CH) é independente de **ZFC** (COHEN, 1963, 1964); *vide* o verbete **Independence and Large Cardinals**[25] da SEP). No entanto, esse caso é muito diferente. Em todos os resultados de independência acima, as afirmações relevantes são ainda teoremas da matemática, tomadas como afirmações cujas verdades se demonstrou (o último caso, que requer axiomas de grandes cardinais que vão além de **ZFC** é mais controverso; ainda assim, ao menos muitos especialistas da teoria de conjuntos consideram tais axiomas plausíveis). E mesmo com o primeiro teorema da incompletude, a verdade da afirmação indemonstrável facilmente se segue, dado que a suposição de consistência do sistema é, de fato, correta. Todavia, no caso do resultado de Cohen, não existe absolutamente nenhuma indicação de se CH deveria ser considerada como verdadeira, falsa, ou mesmo como carente de valor-verdade.

5. A História e a Recepção Inicial dos Teoremas da Incompletude

Os resultados de Gödel foram certamente surpreendentes, mas algum tipo de fenômeno de incompletude não era completamente inesperado. A possibilidade de incompletude no contexto da teoria de conjuntos havia sido discutida por Bernays e Tarski em 1928, e von Neumann, em contraste com o espírito dominante do programa de Hilbert, considerava possível que a lógica e a matemática não fossem decidíveis. O próprio Gödel havia mencionado a possibilidade de um problema indecidível acerca dos números reais na sua tese em 1929 (*vide* DAWSON, 1985). Hilbert (1928), por outro lado, assumiu que a Aritmética de Peano e outras teorias padrão eram completas. Aparentemente, Gödel também se impressionou com Brouwer, que em sua palestra de 1928 em Viena sugeriu que a matemática não seria exaustível e que não poderia ser completamente formalizada (*vide* WANG, 1987, 84); e a seção sobre a visão de Brouwer do programa formalista no verbete **The Development of Intuitionistic**

[25] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/independence-large-cardinals/. Acesso em: 20 jan. 2022.

Logic[26] da SEP).

Seja como for, parece que Gödel realmente chegou às primeiras observações exatas sobre a incompletude por uma rota diferente, durante suas tentativas de **contribuir** para o programa de Hilbert, e não de miná-lo (*vide* DAWSON, 1997: Cap. IV). Ou seja, em 1930, Gödel fez um esforço para avançar o programa de Hilbert, tentando provar a consistência da análise (ou aritmética de segunda ordem) com os recursos da aritmética e, assim, reduzir a consistência da primeira à consistência da segunda. Em sua tentativa de prova, ele precisava da noção de verdade. Gödel logo se deparou com vários paradoxos (como o paradoxo do Mentiroso), e teve que concluir que a verdade aritmética não pode ser definida na aritmética. Assim, Gödel chegou pela primeira vez a uma versão do teorema da indefinibilidade da verdade, geralmente associada a Tarski (*vide* MURAWSKI, 1998). Isso também produz facilmente uma versão fraca do resultado de incompletude: o conjunto de sentenças demonstráveis na aritmética pode ser definido na linguagem da aritmética, mas o conjunto de sentenças aritméticas verdadeiras não pode, portanto os dois não podem coincidir. Além disso, sob a suposição de que todas as sentenças demonstráveis são verdadeiras, segue-se que deve haver sentenças verdadeiras que não podem ser demonstradas. Essa abordagem, contudo, não exibe uma sentença específica.

No entanto, o ambiente intelectual de Gödel era aquele do Círculo de Viena, com sua atitude radicalmente antimetafísica. Em particular, a própria noção de verdade era considerada como suspeita ou mesmo sem sentido à época, ao menos por alguns positivistas lógicos (por exemplo, Neurath, Hempel). Portanto, Gödel trabalhou duro para eliminar qualquer apelo à noção de verdade e na tentativa de dispensá-la. Ele então introduziu a noção de ω-consistência, que pode ser definida rigorosamente e em termos puramente sintáticos. Isso levou aos teoremas da incompletuda na forma que eles são hoje conhecidos.

Quanto ao **Lema da Diagonalização**, o próprio Gödel, de fato, demonstrou apenas um caso especial do lema, ou seja, apenas para o predicado de demonstrabilidade. Aparentemente, o lema geral foi primeiramente descoberto

[26] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/intuitionistic-logic-development/index.html#BrouViewFormProgGodeIncoTheo. Acesso em: 20 jan. 2022.

por Carnap 1934 (*vide* GÖDEL, 1934, 1935). Versões ainda mais gerais, para fórmulas com variáveis livres, foram apresentadas em Ehrenfeucht e Feferman (1960) e Montague (1962) (*vide* SMORYŃSKI, 1981).

A recepção dos resultados de Gödel foi mista. Algumas figuras importantes no campo da lógica e dos fundamentos da matemática assimilaram rapidamente os resultados e entenderam sua relevância, mas também houve muitos mal-entendidos e resistências (para relatos detalhados da recepção, *vide* DAWSON, 1985; MANCOSU, 1999).

Gödel revelou seus resultados para Carnap em Viena no dia 26 de agosto de 1930 e os anunciou (o primeiro teorema) em um comentário de discussão casual na famosa Conferência de Königsberg, no dia 7 de setembro de 1930. John von Neumann, que estava na audiência e trabalhava à época no contexto do programa de Hilbert, compreendeu imediatamente a grande importância do resultado. No dia 20 de novembro, ele escreveu uma carta para Gödel sobre um "notável" corolário do resultado de Gödel que ele teria descoberto: a indemonstrabilidade da consistência (o segundo teorema). Enquanto isso, entretanto, o próprio Gödel tinha já chegado à mesma ideia e enviado a versão final do seu artigo para publicação, o qual incluía também agora uma versão do segundo teorema da incompletude. O artigo foi publicado em janeiro de 1931 (GÖDEL, 1931; introduções úteis ao artigo original de Gödel são KLEENE, 1986, e ZACH, 2005). A notícia desses resultados de importância aparentemente grande para os fundamentos da matemática rapidamente começou a espalhar - embora as opiniões variassem sobre a moral de tais resultados. Paul Bernays, que talvez fosse o mais importante colaborador de Hilbert, mostrou grande interesse nos resultados, embora ele tivesse, primeiramente, dificuldades em compreedê-los apropriadamente. A sua correspondência com Gödel também mostra que Gödel já estava à época completamente ciente da indefinibilidade da verdade.

Na medida em que a abordagem original de Gödel estava focada no seu específico, embora muito abrangente, sistema **P** e em suas (primitivo recursivas) extensões, duvidou-se da generalidade de seus resultados. Alonzo Church, por exemplo, em uma carta endereçada a Gödel em julho de 1932, sugeriu que os resultados de Gödel não se aplicariam ao seu sistema de λ-conversão (Kleene

e Rosser provaram mais tarde que esse sistema é inconsistente). Gödel estava ansioso para generalizar as suas descobertas e estendeu os resultados para uma classe mais ampla de sistemas nos artigos de 1932 e 1934. Ele também sugeriu que esses métodos seriam aplicáveis a sistemas padrão da teoria de conjuntos (no entanto, foi apenas depois da caracterização satisfatória de decidibilidade e da tese Church-Turing, alguns anos depois, que se tornou possível dar uma formulação geral dos teoremas da incompletude (*vide* acima); isso foi feito primeiramente em KLEENE, 1936). O eminente especialista em teoria dos conjuntos, Ernst Zermelo, dirigiu algumas críticas bastante duras ao trabalho de Gödel, mas os dois também se corresponderam sobre o assunto. Zermelo parece ter tido sérias dificuldades em compreender os conceitos e resultados relevantes.

Em março de 1933, Gödel recebeu uma carta de Paul Finsler, de Zurique, que sugeria que ele já havia feito antes (em FINSLER, 1926) um trabalho intimamente relacionado, mas com uma relevância mais geral. Gödel respondeu que o sistema de Finsler não estava realmente definido. Em sua resposta acalorada, Finsler afirmou que não era necessário que um sistema fosse definido com precisão para que para que se pudesse estudá-lo, e que não havia diferença de princípio entre suas ideias e as de Gödel. Retrospectivamente, fica bastante claro que as abordagens de Finsler e Gödel eram muito diferentes: para o trabalho de Gödel, a noção de sistema formalizado era essencial, enquanto Finsler rejeitou a própria noção como artificialmente restritiva. Na verdade, está longe de ser claro que as ideias de Finsler façam algum sentido – quaisquer que sejam as analogias vagas que possam existir entre elas e a prova de Gödel.

É justo dizer, por outro lado, que Emil Post antecipou em alguns aspectos as descobertas de Gödel. Ele obteve, aparentemente já em 1922, versões abstratas dos resultados de incompletude. Em particular, ele observou que seus métodos forneceriam afirmações indecidíveis nos *Principia Mathematica*. Esses resultados eram, entretanto, baseados na versão própria de Post da "tese Church-Turing", com a qual ele estava insatisfeito, e seu trabalho não foi publicado. Isso foi relatado muito mais tarde em POST, 1941.

A correção dos teoremas de Gödel permaneceu um debate vívido ao longo da década de 1930 (*vide* DAWSON, 1985). Em 1939, veio à tona o segundo volume de *Die Grundlagen der Mathematik* de Hilbert e Bernay, incluindo uma

prova detalhada do segundo teorema da incompletude. Depois disso, oposições sérias às conclusões de Gödel desapareceram, ao menos entre aqueles que estavam trabalhando ativamente em lógica matemática e nos fundamentos da matemática. No entanto, em círculos mais filosóficos, permaneceu alguma resistência. Em um famoso exemplo, Wittgenstein fez alguns comentários críticos em relação ao toerema de Gödel em sua obra publicada póstumamente, *Observações sobre a Filosofia da Matemática*. A reação inicial predominante foi a de que Wittgenstein teria simplesmente falhado em compreender o resultado. Interpretações mais caridosas emergiram e esse debate permanece ainda muito vivo (*vide* a seção sobre Gödel e as proposições indecidíveis no verbete **Wittgenstein's Philosophy of Mathematics**[27] da SEP).

6. Implicações Filosóficas - Reais e Alegadas

6.1. Filosofia da Matemática

De todos os campos da filosofia, os teoremas de Gödel são, obviamente, mais imediatamente relevantes para a filosofia da matemática. Para começar, eles colocam sérios problemas para o programa de Hilbert, ao menos *prima facie* (essa questão é discutida em algum detalhe na seção sobre o impacto da incompletude no verbete **Hibert's Program**[28] da SEP). Mais uma vez, os resultados de Gödel tem também importantes consequências para o intuicionismo (*vide* o verbete **Intuitionism in the Philosophy of Mathematics**[29] da SEP) (*vide* também GÖDEL, 1933, 1941; RAATIKAINEN, 2005).

Houve alguma disputa sobre se os teoremas de Gödel refutam conclusivamente o logicismo (*vide* o verbete **Logicism and Neologicism**[30] da

[27] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/wittgenstein-mathematics/index.html#WittGodeUndeMathProp. Acesso em: 20 jan. 2022.

[28] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/hilbert-program/#1.4. Acesso em: 20 jan. 2022.

[29] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/intuitionism/. Acesso em: 20 jan. 2022.

[30] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/logicism/. Acesso em: 20 jan. 2022.

SEP). Henkin (1962) e Musgrave (1977), por exemplo, argumentam que sim; Sternfield (1976) e Rodriguez-Consuegra (1993) discordam (*vide* também HELLMAN, 1981; RAATIKAINEN, 2005).

O próprio Gödel desenvolveu um argumento, baseado nos resultados de incompletude, contra a filosofia convencionalista da matemática do positivismo lógico e a de Carnap em particular (GÖDEL, 1953/59). Esse argumento é discutido em GOLDFARB; RICKETTS, 1992; RICKETTS, 1995; GOLDFARB, 1995; CROCCO, 2003; AWODEY; CARUS, 2003, 2004; TENNANT, 2008.

6.2. Verdades analíticas e autoevidentes

Pode-se dar ainda interpretações epistemológicas mais gerais dos teoremas de Gödel. Quine e Ullian (1978), por exemplo, consideram a imagem filosófica tradicional segundo a qual todas as verdades poderiam ser demonstradas por passos autoevidentes a partir de verdades autoevidentes e da observação. Eles então apontam que mesmo as verdades da teoria elementar dos números presumivelmente não são em geral deriváveis por passos autoevidentes a partir de verdades autoevidentes (QUINE; ULLIAN 1978: 64-65.) Hilary Putnam (1975), por sua vez, sustenta que, sob uma certa compreensão natural de "analítico", deve haver, pelos teoremas de Gödel, verdades sintéticas na matemática. De fato, o próprio Gödel fez observações semelhantes em espírito, segundo as quais mesmo a teoria dos inteiros é demonstravelmente não-analítica (GÖDEL, 1944).

6.3. Argumentos Gödelianos contra o mecanicismo

Houve repetidas tentativas de se aplicar os teoremas de Gödel para demonstrar que os poderes da mente humana superam qualquer mecanismo ou sistema formal. Um destes argumentos Gödelianos contra o mecanicismo foi já considerado, apenas no intuito de refutá-lo, por Turing no final da década de 1940 (*vide* PICCININI, 2003). Uma conclusão antimecanicista não qualificada foi retirada dos teoremas da incompletude em uma exposição popular amplamente lida, **A Prova de Gödel** de Nagel e Newman (1958). Pouco depois, J.R. Lucas

famosamente proclamou que o teorema da incompletude de Gödel

> prova que o Mecanicismo é falso, ou seja, que mentes não podem ser explicadas como máquinas.

Ele afirma que

> dada uma máquina qualquer que seja consistente e capaz de executar aritmética simples, existe uma fórmula que ela é incapaz de produzir como verdadeira... mas cuja verdade nós somos capazes de ver.

Mais recentemente, alegações muito semelhantes foram apresentadas por Roger Penrose (1989, 1994). John Searle (1997) juntou-se à discussão e parcialmente defendeu Penrose contra seus críticos. Crispin Wright (1994, 1995) endossou ideias relacionadas de um ponto de vista intuicionista (para críticas, *vide* DETLEFSEN, 1995). Todos eles insistem que os teoremas de Gödel implicam que a mente humana supera infinitamente o poder de qualquer máquina finita ou sistema formal.

Esses argumentos Gödelianos antimecanicistas são, no entanto, problemáticos, e há um amplo consenso de que eles falham. A resposta padrão a este argumento segue as seguintes linhas (esta objeção remonta a PUTNAM, 1960; *vide* também BOOLOS, 1968, SHAPIRO, 1998): o argumento assume que para qualquer sistema formalizado, ou uma máquina finita, existe a sentença de Gödel que é indemonstrável naquele sistema, mas que a mente humana pode ver como verdadeira. No entanto, o teorema de Gödel tem na realidade uma forma condicional, e a suposta verdade da sentença de Gödel de um sistema depende da suposição da consistência do sistema. O argumento do antimecanicista, portanto, também requer que a mente humana sempre possa ver se uma determinada teoria formalizada é consistente ou não. No entanto, isso é altamente implausível (cf. DAVIS, 1990). Lucas, Penrose e outros tentaram responder a essas críticas (*vide*, por exemplo, LUCAS, 1996; PENROSE, 1995, 1997). Para críticas detalhadas de Penrose, *vide* BOOLOS, 1990; DAVIS, 1990, 1993; FEFERMAN, 1995; LINDSTRÖM,2001; PUDLÁK, 1999; SHAPIRO, 2003; muitas dessas considerações também são relevantes para o que diz Lucas.

6.4. Gödel e Benacerraf sobre Platonismo e Mecanicismo

Curiosamente, o próprio Gödel apresentou um argumento antimecanicista, embora mais cauteloso e publicado apenas postumamente (em seu **Collected Works**, Vol. III, em 1995). Assim, em sua Gibbs *lecture* de 1951, Gödel tirou a seguinte conclusão disjuntiva dos teoremas da incompletude:

> ou ... a mente humana (mesmo dentro do reino da matemática) supera infinitamente os poderes de qualquer máquina finita, ou existem problemas diofantinos absolutamente insolúveis.

Gödel fala sobre esta afirmação como um "fato matematicamente estabelecido" (GÖDEL, 1951); para mais discussão sobre a disjunção de Gödel, *vide*, por exemplo, SHAPIRO, 1998). De acordo com Gödel, a segunda alternativa

> parece refutar a visão de que a matemática é apenas nossa própria criação ... que objetos e fatos matemáticos ... existem objetivamente e independentemente de nossos atos e decisões mentais.

Gödel, no entanto, estava inclinado a negar a possibilidade de problemas absolutamente insolúveis e, embora acreditasse no Platonismo matemático, suas razões para essa convicção eram diferentes, e ele não sustentou que os teoremas da incompletude sozinhos estabelecem o Platonismo. Assim Gödel acreditava na primeira disjunção, que a mente humana supera infinitamente o poder de qualquer máquina finita. Ainda assim, essa conclusão de Gödel se sustenta, como o próprio Gödel explica claramente, apenas se alguém negar, como Gödel, a possibilidade de problemas humanamente insolúveis. Isso não é uma consequência necessária dos teoremas da incompletude.

Gödel era, ao contrário dos defensores posteriores do assim chamado argumento antimecanicista Gödeliano, razoável o suficiente para admitir que tanto o mecanicismo quanto a alternativa de que existem problemas humanamente absolutamente insolúveis são consistentes com seus teoremas da incompletude. Suas razões fundamentais para não gostar da última alternativa são muito mais filosóficas. Gödel pensava, de uma maneira um tanto kantiana, que a razão

humana seria fatalmente irracional se fizesse perguntas que não pudesse responder (para uma discussão crítica, *vide* KREISEL, 1967; BOOLOS, 1995; RAATIKAINEN, 2005).

Como reação ao argumento de Lucas, mas antes da publicação da Gibbs *lecture* de Gödel, Paul Benacerraf (1967) apresentou conclusões mais qualificadas que curiosamente se assemelham a algumas das ideias de Gödel. Ele argumentou que é consistente com todos os fatos que eu sou de fato uma máquina de Turing, mas que não posso determinar qual. Para alguma discussão crítica, *vide* CHIHARA, 1972, e HANSON, 1971.

6.5. Misticismo e a existência de Deus?

Às vezes, conclusões bastante fantásticas são tiradas dos teoremas de Gödel. Foi até sugerido que os teoremas de Gödel, se não provam exatamente, pelo menos dão um forte apoio ao misticismo ou à existência de Deus. Essas interpretações parecem assumir um ou mais mal-entendidos que já foram discutidos acima: ou é assumido que Gödel forneceu uma sentença absolutamente indemonstrável, ou que os teoremas de Gödel implicam o Platonismo, ou o antimecanicismo, ou ambos.

Para mais discussão sobre os aspectos filosóficos dos teoremas da incompletude, *vide* RAATIKAINEN, 2005, e FRANZÉN, 2005.

Leitura adicional

Uma referência padrão para os teoremas da incompletude é:

SMORYŃSKI, C. "The incompleteness theorems", in **Handbook of Mathematical Logic**, BARWISE, J. (ed.), Amsterdam: North-Holland, 1977, pp. 821–866 [https://www2.karlin.mff.cuni.cz/~krajicek/smorynski.pdf]

Existem vários livros introdutórios em lógica matemática que dão uma boa exposição dos teoremas da incompletude e tópicos relacionados; por exemplo:

BOOLOS, G.; JEFFREY, R. **Computability and Logic**, 3a Edição Revisada, Cambridge: Cambridge University Press, 1989.

ENDERTON, H. **A Mathematical Introduction to Logic**, New York: Academic Press, 1972.
VAN DALEN, D. **Logic and Structure**, 4th edition, Berlin: Springer, 2004.

Dois livros que são dedicados aos teoremas da incompletude são:

SMULLYAN, R. **Gödel's Incompleteness Theorems**, Oxford: Oxford University Press, 1991.
SMITH, P. **An Introduction to Gödel's Theorems**, Cambridge: Cambridge University Press, 2007.

Outro livro útil sobre os teoremas da incompletude e tópicos relacionados é:

MURAWSKI, R. **Recursive Functions and Metamathematics: Problems of Completeness and Decidability, Gödel's Theorems**. Dordrecht: Kluwer, 1999.

Um livro abrangente e mais avançado sobre esses temas é:

HÁJEK, P.; PUDLÁK, P. **Metamathematics of First-Order Arithmetic**, Berlin: Springer, 1993.

Outro livro útil, incluindo também alguns tópicos avançados é:

FRANZÉN, T. **Inexhaustibility: A Non-Exhaustive Treatment**, Lecture Notes in Logic 16, ASL, Wellesley: A.K. Peters, 2004.

Os aspectos mais filosóficos em torno dos teoremas da incompletude são examinados nas duas fontes a seguir (Franzén é uma explicação acessível, informal e ainda confiável dos teoremas da incompletude):

RAATIKAINEN, P. "On the Philosophical Relevance of Gödel's Incompleteness Theorems", **Revue Internationale de Philosophie**, 59: 513–534, 2005 [https://www.cairn.info/revue-internationale-de-philosophie-2005-4-page-513.htm]

FRANZÉN, T. **Gödel's Theorem: An Incomplete Guide to its Use and Abuse**, Wellesley: A.K. Peters, 2005.

Os dois artigos a seguir examinam várias questões em torno do primeiro teorema da incompletude:

BEKLEMISHEV, L. D. "Gödel incompleteness theorems and the limits of their applicability. I", **Russian Mathematical Surveys**, 65: 857–898, 2010.

BULDT, B. "The scope of Gödel's first incompleteness theorem", **Logica Universalis**, 8: 499–552, 2004.

Por fim, há um e-book de código aberto que contém uma apresentação dos teoremas da incompletude:

ZACH, R. **Incompleteness and Computability**, e-Book publicado pelo Open Logic Project. [https://ic.openlogicproject.org/], 2019.

Bibliografia

AUERBACH, D.J. "Intensionality and the Gödel theorems", **Philosophical Studies**, 48 (3):337-51, 1985.

AUERBACH, D.J. "How to say things with formalisms", in **Proof, Logic, and Formalization**, M. Detlefsen (ed.), London: Routledge, 77–93, 1992 [disponível em: https://www.academia.edu/1861652/How_to_Say_Things_With_Formalisms].

AWODEY, S.; CARUS, A.W. "Carnap versus Gödel on Syntax and Tolerance," in **Logical Empiricism: Historical and Contemporary Perspectives**, P. Parrini et al. (eds.), Pittsburgh: University of Pittsburgh Press, pp. 57–64, 2003 [disponível em: https://www.cmu.edu/dietrich/philosophy/docs/tech-reports/106_Awodey.pdf].

AWODEY, S.; CARUS, A.W. "How Carnap Could Have Replied to Gödel," in S. Awodey and C. Klein (eds.), **Carnap Brought Home: The View from Jena**, LaSalle, IL: Open Court, pp. 203–223, 2004 [disponível em: https://www.cmu.edu/dietrich/philosophy/docs/tech-reports/123_Awodey.pdf].

BARZIN, M. “Sur la portée du théorème de M. Gödel”, **Académie Royale de Belgique, Bulletin de la Classe des Sciences**, Series 5, 26: 230–39, 1940.

BENACERRAF, P. “God, the Devil, and Gödel,” **The Monist**, 51: 9–32, 1967 [disponível em: http://www2.units.it/etica/2003_1/3_monographica.htm].

BEZBORUAH, A.; SHEPHERDSON, J.C. “Gödel’s Second Incompleteness Theorem for Q,” **The Journal of Symbolic Logic**, 41: 503–512, 1976.

BOOLOS, G. “Review of ‘Minds, Machines and Gödel’, by J.R. Lucas, and ‘God, the Devil, and Gödel’,” **Journal of Symbolic Logic**, 33: 613–15, 1968.

BOOLOS, G. “On ‘Seeing’ the Truth of Gödel Sentence”, **Behavioral and Brain Sciences**, 13: 655–656, 1990.

BOOLOS, G. “Introductory Note to *1951”, in **Gödel**: 290–304, 1995.

BOOLOS, G.; JEFFREY, R. **Computability and logic**, 3rd revised edition, Cambridge: Cambridge University Press, 1989.

CARNAP, R. **Logische Syntax der Sprache**, Vienna: Julius Springer, 1934.

CHIHARA, C. “On Alleged Refutations of Mechanism Using Gödel’s Incompleteness Results,” **Journal of Philosophy**, 69: 507–26, 1972.

CHURCH, A. “An Unsolvable Problem of Elementary Number Theory”, **American Journal of Mathematics**, 58: 354–363, 1936a. Republished in Davis 1965, 89–107.

CHURCH, A. “A Note on Entscheidungsproblem,” **Journal of Symbolic Logic**, 1: 40–41, 1936b; correction, ibid., 101–102. Republished in Davis 1965, 110–115.

COHEN, P.J. “The Independence of the Continuum Hypothesis I”, **Proceedings of the National Academy of Sciences**, (U.S.A.), 50(6): 1143–48, 1963.

COHEN, P.J. “The Independence of the Continuum Hypothesis II”, **Proceedings of the National Academy of Sciences**, (U.S.A.), 51(1): 105–110, 1964.

CROCCO, G. “Gödel, Carnap, and the Fregean Heritage”, **Synthese**, 137: 21–41, 2003.

DAVIS, M. **The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions**, Hewlett, NY: Raven Press, 1965.

DAVIS, M. “Hilbert’s Tenth Problem is Unsolvable”, **The American Mathematical**

Monthly, 80: 233–269, 1973.

DAVIS, M. "Unsolvable Problems", in **Handbook of Mathematical Logic, J. Barwise** (ed.), Amsterdam: North-Holland, pp. 567–594, 1977.

DAVIS, M. "Is Mathematical Insight Algorithmic?", **Behavioral and Brain Sciences**, 13: 659–660, 1990.

DAVIS, M. "How Subtle is Gödel's Theorem? More on Roger Penrose", **Behavioral and Brain Sciences**, 16: 611–612, 1993.

DAVIS, M. & PUTNAM, H. & ROBINSON, J. "The decision problem for exponential diophantine equations", **Annals of Mathematics** (2), 74(3): 425–436, 1961.

DAWSON, J. "The Reception of Gödel's Incompleteness Theorems", **PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1984**, vol. II, pp. 253–271, 1985.

DAWSON, J. **Logical Dilemmas: The Life and Work of Kurt Gödel**, Natick, MA: A. K. Peters, 1997.

DETLEFSEN, M. "On Interpreting Gödel's Second Theorem", **Journal of Philosophical Logic**, 8(1): 297–313, 1979.

DETLEFSEN, M. **Hilbert's Program: An Essay in Mathematical Instrumentalism**, Dordrecht: Reidel, 1986.

DETLEFSEN, M. "On an Alleged Refutation of Hilbert's Program Using Gödel's First Incompleteness Theorem", **Journal of Philosophical Logic**, 19(4): 343–377, 1990.

DETLEFSEN, M. "Wright on the Non-mechanizability of Intuitionist Reasoning", **Philosophia Mathematica**, 3(1): 103–118, 1995.

DETLEFSEN, M. "What Does Gödel's Second Theorem Say?", **Philosophia Mathematica**, 9: 37–71, 2001.

DYSON, V.; JONES, J.P. & SHEPHERDSON, J.C. "Some Diophantine Forms of Gödel's Theorem", **Archiv für Mathematische Logik und Grundlagenforschung**, 22: 51–60, 1982.

EHRENFEUCHT, A.; FEFERMAN, S. "Representability of recursively enumerable sets in formal theories", **Arch. Math. Logik Grundlag.**, 5(1–2), 37–41, 1960.

FEFERMAN, S. "Arithmetization of Metamathematics in a General Setting", **Fundamenta Mathematicae**, 49: 35–92, 1960.

FEFERMAN, S. “Inductively Presented Systems and the Formalization of Meta-mathematics”, in **Logic Colloquium ’80**, D. van Dalen et al. (eds.), Amsterdam: North-Holland, pp. 95–128, 1982.

FEFERMAN, S. “Finitary Inductively Presented Logics,” in **Logic Colloquium ‘88**, R. Ferro, et al. (eds.), Amsterdam: North-Holland, pp. 191–220, 1989a. [disponível em: https://web.archive.org/web/20170311004001/http://math.stanford.edu/~feferman/papers/presentedlogics.pdf]

FEFERMAN, S. “Infinity in Mathematics: Is Cantor Necessary?”, **Philosophical Topics**, 17(2): 23–45, 1989b.

FEFERMAN, S. “Penrose’s Gödelian argument: A Review of Shadows of Mind, by Roger Penrose”, **Psyche**, 2 (7), 1995.

FEFERMAN, S. “My Route to Arithmetization”, **Theoria**, 63: 168–181, 1997.

FINSLER, P. “Formale Beweise und die Entscheidbarkeit”, **Mathematische Zeitschrift**, 25: 676–82, 1926.

FITTING, M. **Incompleteness in the land of sets**, London: College Publications. Series: Studies in logic ; v. 5, 2007.

FRANKS, C. **The Autonomy of Mathematical Knowledge. Hilbert’s Program Revisited**, Oxford: Oxford University Press, 2009.

GAIFMAN, H. “Naming and Diagonalization, From Cantor to Gödel to Kleene,” Logic Journal of the IGPL, 14: 709–728, 2006. [disponível em: https://haimgaifman.net/naming-diag.pdf]

GENTZEN, G. “Die Widerspruchsfreiheit der reinen Zahlentheorie”, **Mathematische Annalen**, 112: 493–565, 1936.

GÖDEL, K. “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”, **Monatshefte für Mathematik Physik**, 38: 173–198, 1931. English translation in van Heijenoort 1967, 596–616, and in Gödel 1986, 144–195.

GÖDEL, K. “Über Vollständigkeit und Widerspruchsfreiheit”, **Ergebnisse eines mathematischen Kolloquiums**, 3: 12–13, 1932. English translation “On Completeness and Consistency” in Gödel 1986: 235–7.

GÖDEL, K. “The Present Situation in Foundations of Mathematics”, in Gödel 1995: 45–53, 1933.

GÖDEL, K. “On Undecidable Propositions of Formal Mathematical Systems”,

(mimeographed lecture notes; taken by S. Kleene and J. Rosser), reprinted with corrections in Davis 1965, 41–81, and Gödel 1986, 346–371, 1934.

GÖDEL, K. "Review of Carnap 1934", in Gödel 1986: 389, 1935.

GÖDEL, K. "In What Sense is Intuitionistic Logic Constructive?" in Gödel 1995: 189–200, 1941.

GÖDEL, K. "Russell's Mathematical Logic," in **The Philosophy of Bertrand Russell**, P. A. Schilpp (ed.), Evanston, Il.: Northwestern University, pp. 125–153. Reprinted in Gödel 1990: 119–141, 1944.

GÖDEL, K. "Some Basic Theorems on the Foundations of Mathematics and their Implications" (Gibbs Lecture), in Gödel 1995: 304–323, 1951.

GÖDEL, K. "Is Mathematics a Syntax of Language?", lecture manuscript (two versions), in Gödel 1995: 334–362, 1953/9.

GÖDEL, K. "Note added 28 August 1963" (to Gödel 1931), in Gödel 1986: 195, 1963.

GÖDEL, K. **Collected Works I. Publications 1929–1936**, S. Feferman et al. (eds.), Oxford: Oxford University Press, 1986.

GÖDEL, K. **Collected Works II. Publications 1938–1974**, S. Feferman et al. (eds.), Oxford: Oxford University Press, 1990.

GÖDEL, K. **Collected Works III. Unpublished Essays and Lectures**, S. Feferman et al. (eds.), Oxford: Oxford University Press, 1995.

GOLDFARB, W. "Introductory Note to *1953/9," in Gödel 1995: 324–334, 1995.

GOLDFARB, W.; RICKETTS, T. "Carnap and the Philosophy of Mathematics," in **Science and Subjectivity**, D. Bell and W. Vossenkuhl (eds.), Berlin: Akademie Verlag, pp. 61–78, 1992.

GÓMEZ TORRENTE, M. "The Indefinability of Truth in the Wahrheitsbegriff", **Annals of Pure and Applied Logic**, 126(1–3): 27–37, 2004. [disponível em: https://www.sciencedirect.com/science/article/pii/S0168007203000903?via%3Dihub]

GOODSTEIN, R. "On the Restricted Ordinal Theorem", **The Journal of Symbolic Logic**, 9: 33–41, 1944.

GRELLING, K. "Gibt es eine Gödelsche Antinomie?", **Theoria**, 3: 297–306, 1937.

HANSON, W.H. "Mechanism and Gödel's theorems", **The British Journal for the Philosophy of Science**, 22: 9–16, 1971.

HELLMAN, G. "How to Gödel a Frege-Russell: Gödel's Incompleteness Theorems and Logicism", **Nous**, 15: 451–468, 1981.

HELMER, O. "Perelman versus Gödel", **Mind**, 46: 58–60, 1938.

HENKIN, L. "Problem", **The Journal of Symbolic Logic**, 17: 160, 1952.

HENKIN, L. "Are Mathematics and Logic Identical?", **Science**, 138: 788–794, 1962.

HILBERT, D. "Die Grundlagen der Mathematik", **Abhandlungen aus dem Mathematischen Seminar der Hamburgischen Universität**, 6: 65–85, 1928. English translation in van Heijenoort 1967.

HILBERT, D.; BERNAYS, P. **Grundlagen der Mathematik**, vol. 2, Berlin: Springer, 1939.

JEROSLOW, R. "Redundancies in the Hilbert-Bernays Derivability Conditions for Gödel's Second Incompleteness Theorem", **Journal of Symbolic Logic**, 38: 359–367, 1973.

KAYE, R. **Models of Peano Arithmetic**, (Oxford Logic Guides), Oxford: Clarendon Press, 1991.

KIRBY, L.; PARIS, J. "Accessible Independence Results for Peano Arithmetic", **Bull. London. Math. Soc.**, 14: 285–93, 1982.

KLEENE, S.C. "General recursive functions of natural numbers", **Mathematische Annalen** 112(1): 727–742, 1936.

KLEENE, S.C. "Review of Perelman 1936", **Journal of Symbolic Logic**, 2: 40–41, 1937a.

KLEENE, S.C. "Review of Helmer 1937", **Journal of Symbolic Logic**, 2: 48–49, 1937b.

KLEENE, S.C. "Introductory note to 1930b, 1931 and 1932b", in Gödel 1986, pp. 126–141, 1986.

KREISEL, G. "On a Problem of Henkin's", **Proc. Netherlands Acad. Sci.** 56: 405–406, 1953.

KREISEL, G. "Mathematical significance of consistency proofs", **The Journal of Symbolic Logic**, 23: 159–182, 1958.

KREISEL, G. "Mathematical Logic: What Has it Done For the Philosophy of Mathematics?" in **Bertrand Russell: Philosopher of the Century**, R. Schoenman (ed.), London: George Allen and Unwin, 1967.

KRUSKAL, J.B. “Well-quasi-ordering, the Tree Theorem, and Vazsonyi’s Conjecture”, **Transactions of the American Mathematical Society**, 95 (2): 210–225, 1960.

LINDSTRÖM, P. “Penrose’s New Argument”, **Journal of Philosophical Logic**, 30(3): 241–250, 2001.

LUCAS, J.R. “Minds, Machines, and Gödel”, **Philosophy**, 36(137): 112–137, 1961 [disponível em: https://web.archive.org/web/20200301170651/http://users.ox.ac.uk/~jrlucas/Godel/mmg.html].

LUCAS, J.R. “Minds, Machines, and Gödel: A Retrospect”, in **Machines and Thought. The Legacy of Alan Turing**, Vol. 1, P.J.R. Millican and A. Clark (eds.), Oxford: Oxford University Press, 103–124, 1996.

LÖB, M.H. “Solution of a Problem of Leon Henkin”, in **Journal of Symbolic Logic**, 20: 115–118, 1955.

MANCOSU, P. “Between Vienna and Berlin: The Immediate Reception of Gödel’s Incompleteness Theorems”, **History and Philosophy of Logic**, 20: 33–45, 1999.

MARTIN, D. “Descriptive Set Theory: Projective Sets,” in **Handbook of Mathematical Logic**, J. Barwise (ed.), Amsterdam: North-Holland, 783–815, 1977.

MARTIN, D.; STEEL, J. “Projective Determinacy”, **Proceedings of the National Academy of Sciences**, (U.S.A.), 85: 6582–86, 1988.

MARTIN, D.; STEEL, J. “A Proof of Projective Determinacy”, **Journal of the A.M.S.**, 2: 71–125, 1989.

MATIYASEVICH, Y. “Diofantovost’ perechislimykh mnozhestv”, **Dokl. Akad. Nauk SSSR**, 191(2): 297–282 (Russian), 1970. (English translation, 1970, “Enumerable sets are Diophantine”, Soviet Math. Dokl., 11(2): 354–358.)

MATIYASEVICH, Y. **Hilbert’s Tenth Problem**, Cambridge, MA: MIT Press, 1993.

MILNE, P. “On Gödel Sentences and What They Say”, **Philosophia Mathematica**, 15: 193–226, 2007.

MONTAGUE, R. “Theories Incomparable with Respect to Relative Interpretability”, **The Journal of Symbolic Logic**, 27: 195–211, 1962.

MURAWSKI, R. “Undefinability of Truth. The Problem of Priority: Tarski vs. Gödel”, **History and Philosophy of Logic**, 19: 153–160, 1998.

MURAWSKI, R. **Recursive Functions and Metamathematics: Problems of Completeness and Decidability, Gödel's Theorems**, Dordrecht: Kluwer, 1999.

MUSGRAVE, A. "Logicism Revisited", **British Journal for the Philosophy of Science**, 28: 99–127, 1977.

NAGEL, E.; NEWMAN, J.R. **Gödel's Proof**, New York: New York University Press, 1958.

PARIS, J.; HARRINGTON, L. "A Mathematical Incompleteness in Peano Arithmetic", in **Handbook of Mathematical Logic**, J. Barwise (ed.), Amsterdam: North-Holland, pp. 1133–1142, 1977 [disponível em: https://www2.karlin.mff.cuni.cz/~krajicek/ph.pdf].

PARIS, J.; KIRBY, L. "S_n Collection Schema in Arithmetic", in **Logic Colloquium '77**, A. McIntyre et al. (eds.), Amsterdam: North-Holland, pp. 199–209, 1978.

PARSONS, C. "On Number Choice Schema and its Relation to Induction", in **Intuitionism and Proof Theory**, Kino et al. (eds.), Amsterdam: North-Holland, pp. 459–473, 1970.

PENROSE, R. **The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics**, New York: Oxford University Press, 1989.

PENROSE, R. **Shadows of the Mind: A Search for the Missing Science of Consciousness**, New York: Oxford University Press, 1994.

PENROSE, R. "Beyond the Doubting of a Shadow: A Reply to Commentaries of Shadows of the Mind", **Psyche**, Vol 2, 1995.

PENROSE, R. "On understanding understanding", **International Studies in the Philosophy of Science**, 11: 7–20, 1997.

PERELMAN, C. "L'Antinomie de M. Gödel", **Académie Royale de Belgique. Bulletin de la Classe des Sciences** (Series 5), 22: 730–36, 1936.

PICCININI, G. "Alan Turing and the Mathematical Objection", **Minds and Machines**, 13: 23–48, 2003.

POST, E. "Absolutely Unsolvable Problems and Relatively Unsolvable Propositions: Account of an Anticipation", published in Davis 1965, 338–433, 1941.

PRESBURGER, M. "Über die Vollständigkeit eines gewissen Systems der

Arithmetik ganzer Zahlen, in welchem die Addition als einzige Operation hervortritt", **Sprawozdanie z I Kongresu Matematyków Krajów Slowiańskich**, (= Comptes-rendus du I Congrès Mathématiciens des Pays Slaves), Warsaw, pp. 92–101, 1929. English translation, 1991,"On the completeness of a certain system of arithmetic of whole numbers in which addition occurs as the only operation", **History and Philosophy of Logic**, 12(2): 225–232.

PUDLÁK, P. "On the Length of Proofs of Consistency", **Collegium Logicum, Annals of the Kurt-Gödel-Society**, 2: 65–86, 1996.

PUDLÁK, P. "A Note on Applicability of the Incompleteness Theorem to Human Mind", **Annals of Pure and Applied Logic**, 96: 335–342, 1999.

PUTNAM, H. "Minds and machines", in **Dimensions of Mind**, S. Hook (ed.), New York: New York University Press, 1960. Reprinted in H. Putnam, 1975, **Mind, Language, and Reality. Philosophical Papers**, Vol 2, Cambridge: Cambridge University Press, pp. 325–341.

PUTNAM, H. "What is Mathematical Truth?", **Historia Mathematica**, 2: 529–545, 1975. Reprinted in H. Putnam, 1975, **Mathematics, Matter and Method. Philosophical Papers**, Vol 1, Cambridge: Cambridge University Press, pp. 60–78.

QUINE, W.V.; ULLIAN, J.S. **The Web of Belief**, 2nd ed., New York: Random House, 1978.

RAATIKAINEN, P. "On the Philosophical Relevance of Gödel's Incompleteness Theorems", **Revue Internationale de Philosophie**, 59: 513–534, 2005.

RAMSEY, F.P. "On a Problem of Formal Logic", **Proceedings of the London Mathematical Society**, series 2, 30: 264–286, 1930.

RICKETTS, T. "Carnap's Principle of Tolerance, Empiricism, and Conventionalism", in **Reading Putnam**, P. Clark & B. Hale (eds.), Cambridge: Blackwell, pp. 176–200, 1995.

RODRÍGUEZ-CONSUEGRA, F. "Russell, Gödel and Logicism", in **Philosophy of Mathematics**, J. Czermak (ed.), Vienna: Hölder-Pichler-Tempsky, pp. 233–42, 1993. Reprinted in, 1998, **Bertrand Russell: Critical Assessments**, A. Irvine (ed.), vol. 2: **Logic and mathematics**, London: Routledge, pp. 320–29.

ROEPER, P. “Giving an Account of Provability within a Theory”, **Philosophia Mathematica**, 11: 332–340, 2003.

ROSSER, J.B. “Extensions of Some Theorems of Gödel and Church”, **Journal of Symbolic Logic**, 1: 87–91, 1936.

ROSSER, J.B. “Review: Kurt Grelling, Gibt es eine Godelsche Antinomie? [Grelling 1937/8]”, **Journal of Symbolic Logic**, 3(2): 86, 1938.

SEARLE, J. “Roger Penrose, Kurt Gödel, and the Cytoskeletons”, in J. Searle: **Mystery of Consciousness**, New York: New York Review of Books, pp. 55–93, 1997.

SHAPIRO, S. “Incompleteness, Mechanism, and Optimism”, **Bulletin of Symbolic Logic**, 4: 273–302, 1998.

SHAPIRO, S. “Mechanism, Truth and Penrose’s New Argument”, **Journal of Philosophical Logic**, 32(1): 19–42, 2003.

SIMPSON, S.G. “Nonprovability of Certain Combinatorial Properties of Finite Trees”, in **Harvey Friedman’s Research on the Foundations of Mathematics**, L. Harrington et al. (eds.), Studies in Logic and the Foundations of Mathematics, Amsterdam: North-Holland, pp. 87–117, 1985.

SIMPSON, S.G. **Subsystems of Second Order Arithmetic**, Berlin: Springer, 1999.

SKOLEM, T. “Über einige Satzfunktionen in der Arithmetik”, **Skrifter utgitt av Det Norske Videnskaps-Akademi i Oslo**, I, no. 7, 1–28, 1930. Reprinted in T. Skolem, 1970, **Selected Works in Logic**, (J. Fenstad, editor), Oslo: Universitetsforlaget, pp. 281–306.

SMORYŃSKI, C. “The Incompleteness Theorems”, in **Handbook of Mathematical Logic**, J. Barwise (ed.), Amsterdam: North-Holland, pp. 821–865, 1977.

SMORYŃSKI, C. “Fifty Years of Self-reference in Arithmetic”, **Notre Dame Journal of Formal Logic**, 22(4): 357–374, 1981.

SMORYŃSKI, C. “The Development of Self-reference: Löb’s Theorem”, in **Perspectives on the History of Mathematical Logic**, T. Drucker (ed.), Birkhauser, pp. 111–133, 1991.

SMULLYAN, R. **Gödel’s Incompleteness Theorems**, Oxford: Oxford University Press, 1992.

SOLOVAY, R.M. “A Model of Set Theory in which Every Set of Reals is Lebesgue Measurable”, **Annals of Mathematics**, 92: 1–56, 1970.

STERNFELD, R. “The Logistic Thesis”, in **Studien zu Frege/Studies on Frege I**, M. Schirn (ed.), Stuttgart-Bad Cannstatt: Frommann-Holzboog, pp. 139–160, 1976.

TARSKI, A. **A Decision Method for Elementary Algebra and Geometry, manuscript**. Santa Monica, CA: RAND Corp., 1948. Republished as **A Decision Method for Elementary Algebra and Geometry**, 2nd ed. Berkeley, CA: University of California Press, 1951.

TARSKI, A.; MOSTOWSKI, A. & ROBINSON, R.M. **Undecidable Theories**, Amsterdam: North-Holland, 1953.

TENNANT, N. “Carnap, Gödel, and the Analyticity of Arithmetic”, **Philosophia Mathematica**, 16: 100–112, 2008.

TURING, A.M. “On Computable Numbers, with an Application to the Entscheidungsproblem”, **Proceedings of the London Mathematical Society**, Series 2, 42: 230–265, 1936-7; correction, ibid., 43: 544–546. Republished in Davis 1965, 115–154.

VAN HEIJENOORT, J. (ed.) **From Frege to Gödel: A Source Book in Mathematical Logic**, 1879–1931, Cambridge, MA: Harvard University Press, 1967.

VISSER, A. “Can We Make the Second Incompleteness Theorem Coordinate Free”, **Journal on Logic and Computation**, 21(4): 543–560, 2011.

WOODIN, H. “Supercompact Cardinals, Sets of Reals, and Weakly Homogeneous Trees”, **Proceedings of the National Academy of Sciences**, (U.S.A.), 85: 6587–91, 1988.

WRIGHT, C. “About ‘The Philosophical Significance of Gödel’s Theorem’: Some Issues”, in **The Philosophy of Michael Dummett**, B. McGuinness and G. Oliveri (eds.) Dordrecht: Kluwer, pp. 167–202, 1994.

WRIGHT, C. “Intuitionists are not (Turing) Machines”, **Philosophia Mathematica**, 3: 86–102, 1995.

ZACH, R. “Paper on the Incompleteness Theorems”, in **Landmark Writings in Western Mathematics**, I. Grattan-Guinness (ed.), Amsterdam: Elsevier, pp. 917–25, 2005 [disponível em:

http://people.ucalgary.ca/~rzach/static/godel1931.pdf].

Complemento 1 - A Numeração de Gödel*

Autoria: Panu Raatikainen
Tradução: Guilherme A. Cardoso
Revisão: Sérgio R. N. Miranda

Um método central nas provas usuais do primeiro teorema da incompletude é a aritmetização da linguagem formal, ou **numeração de Gödel**: certos números naturais são atribuídos para termos, fórmulas e provas da teoria formal F. Há diferentes modos de realizar isso; uma abordagem padrão é esboçada aqui (para um método bem diferente de codificação, *vide*, por exemplo, BOOLOS; JEFFREY, 1989). Uma exigência essencial é que o método seja

*RAATIKAINEN, P. "Gödel Numbering", In: ZALTA, E. N. (ed.) **The Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/goedel-incompleteness/sup1.html. Acesso em: 20 jan. 2022.

The following is the translation of the supplement entry on on Gödel's Incompleteness Theorems by P. Raatikainen in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/goedel-incompleteness/sup1.html. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry.Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

efetivo, ou seja, uma rotina puramente mecânica.

O método procede em dois passos:

1. Números Símbolos

Para começar, a cada símbolo primitivo s da linguagem do sistema formal F em questão, um número natural $\#(s)$, chamado de **número símbolo** de s, é associado. Não importa como e em que ordem o procedimento é feito - isso é arbitrário -, mas, uma vez feito, ele é obviamente mantido fixo.

Por exemplo, no caso da linguagem padrão da aritmética, com os símbolos $0, ', +, \times, =, (,), \neg, \rightarrow, \forall$ (para simplificar, vamos assumir que $\neg, \rightarrow$ e $\forall$ são os únicos símbolos primitivos, e que $\wedge, \vee, \leftrightarrow$ e $\exists$ são definidos com a ajuda dos símbolos primitivos), podemos proceder, e.g., como se segue:

$$\begin{array}{llllllll} \#(\text{`0'}) & = & 1 & \#(\text{`='}) & = & 5 & \#(\text{`}\neg\text{'}) & = & 9 \\ \#(\text{`}'\text{'}) & = & 2 & \#(\text{`('}) & = & 6 & \#(\text{`}\forall\text{'}) & = & 10 \\ \#(\text{`+'}) & = & 3 & \#(\text{`)'}) & = & 7 & \#(\text{`}x_i\text{'}) & = & 11+i \\ \#(\text{`}\times\text{'}) & = & 4 & \#(\text{`}\rightarrow\text{'}) & = & 9 & & & \end{array}$$

2. Sequências Codificadoras

Algum modo de codificar sequências finitas de números por números simples é também fixado. Há indefinidamente muitos modos de fazer isso; uma abordagem comum (usada também pelo próprio Gödel) é baseada nos produtos de potências de números primos.

Lembre que um **número primo** é um número natural que é maior que 1 e pode ser dividido somente por 1 e pelo próprio número (todos os outros números maiores que 1 são chamados de "compostos"). Há infinitamente muitos números primos; o começo da sequência é 2, 3, 5, 7, 11, 13, 17 …

O Teorema Fundamental da Aritmética (ou o teorema da fatorização única em primos) afirma que qualquer número natural maior que 1 pode ser escrito como o produto único (até a ordenação dos fatores) de números primos.

Seja então p_1 o primeiro número primo, p_2 o segundo número primo, e

assim por diante. Em geral, p_n é o n-ésimo número primo, e

$$p_1, p_2, \ldots, p_n$$

é a sequência dos primeiros n números primos em ordem crescente.

Dada uma sequência finita arbitrária de números positivos (0 causaria complicações) com comprimento de $k + 1$, $(n_0, n_1, \ldots, n_k)$, ela pode ser univocamente condificada como um produto de potências de números primos $p_1, p_2, \ldots, p_{k+1}$ como se segue:

$$c = 2^{n_0} \times 3^{n_1} \times 5^{n_2} \times \cdots \times p_{k+1}^{n_k}$$

Por exemplo, a sequência $\langle 3, 1, 2 \rangle$ é codificada como $2^3 \times 3^1 \times 5^2$, ou seja, $8 \times 3 \times 25$, que é igual a 600.

Combinando os dois passos:

Dados esses dois métodos, é possível codificar uma expressão arbitrária da linguagem por um número simples: primeiramente, substitua cada símbolo s por um número símbolo #(s). Desse modo, uma sequência de símbolos se torna uma sequência de números. Em segundo lugar, usando a codificação por potências de números primos acima, associe a cada sequência de números um único número simples como o seu código - o seu "número de Gödel".

Por exemplo, considere a expressão simples: '$0 + 0$'. Como foi estipulado acima que os números símbolos dos símbolos '0' e '$=$' são 1 e 5, respectivamente, a sequência correspondente de números símbolos é $\langle 1, 5, 1 \rangle$. O código (ou seja, o número de Gödel) de '$0 + 0$' é então:

$$2^1 \times 3^5 \times 5^1 = 2 \times 243 \times 5 = 2430$$

De modo geral, o número de Gödel de uma fórmula (sentença, derivação) A é denotado por $\ulcorner A \urcorner$. Por exemplo, $\ulcorner 0 = 0 \urcorner$ refere, sob a codificação que fixamos,

a 2430.

Desse modo, podemos atribuir números de Gödel a fórmulas, sequências de fórmulas (desde que um método para distinguir quando uma fórmula termina e outra começa tenha sido adotado), e, mais notavelmente, a provas e derivações.

Podemos também ir na outra direção: é uma parte essencial do método que, se um número de código é dado (muitos números simplesmente não codificam qualquer coisa, mas pode-se decidir quais o fazem), é também possível decodificá-lo de uma maneira única, ou seja, reconstruir a expressão geral única (ou a derivação) que ele codifica.

Por exemplo, seja 18 o número de Gödel dado. A sua fatorização em primos, que é única (pelo teorema fundamental da aritmética; *vide* acima), pode ser determinada. Ela é:

$$2^1 \times 3^2$$

Focando nas potências, vemos que ela representa a sequência $\langle 1, 2 \rangle$, e relembrando os números símbolos relevantes:

$$1 = \#(\text{`0'})$$
$$2 = \#(\text{`}'\text{'})$$

pode-se ver que a sequência codificada de símbolos é $\langle 0, ' \rangle$, ou seja, a expressão que é codificada por 18 é '$0'$'.

3. Definindo propriedades sintáticas e operações

É então possível desenvolver as noções centrais da sintaxe exata na forma aritmetizada, imitando as suas definições ordinárias, embora definições rigorosas tendam a se tornar um pouco complicadas. Ou seja, é possível definir, em uma linguagem artimética, propriedades como:

$Const(x)$ x é (um número de Gödel de) uma constante.
$Var(x)$ x é (um número de Gödel de) uma variável.
$Term(x)$ x é (um número de Gödel de) um termo.
$Form(x)$ x é (um número de Gödel de) uma fórmula.

As seguinte operações sobre (números de Gödel de) fórmulas podem também ser facilmente definidas na linguagem da aritmética:

$neg(x)$ a função aritmética que leva do número de Gödel de uma formula ao número de Gödel de sua negação:

$$neg(\ulcorner A \urcorner) = (\ulcorner \neg A \urcorner);$$

$impl(x, y)$ a função que mapeia os números de Gödel de um par de fórmulas no número de Gödel da implicação de tais fórmulas:

$$impl(\ulcorner A \urcorner, \ulcorner B \urcorner) = \ulcorner A \to B \urcorner;$$

Dado que $y(z)$ é (um número de Gödel de) uma fórmula, as seguintes relações sintáticas podem ser também definidas na aritmética:

$Free(x, y)$ x é variável livre [*free*] em y.
$FreeFor(x, y, z)$ x é livre para [*free for*] y em z.

A seguinte operação sobre os números de Gödel tem um papel particularmente importante:

$subst(x, y) = z$ é a operação que mapeia o par com o número de Gödel de um fórmula com um variável livre e o número de um numeral ao número de Gödel

de uma fórmula fechada que resulta na fórmula original quando o dado numeral é substituído pela variável livre:

$$subst(\ulcorner A(x) \urcorner, \ulcorner \underline{n} \urcorner) = \ulcorner A(\underline{n}) \urcorner$$

Uma vez fixada a lógica de fundo, os seus axiomas e regras de inferência, o seguinte pode ser definido:

$LogAx(x)$ x é o número de Gödel de um axioma lógico.

Regras de inferência podem também ser expressas na forma aritmetizada. Por exemplo, há uma fórmula aritmética $M(x, y, z)$ que é verdadeira exatamente quando se tem em mãos uma aplicação de uma regra de inferência padrão "Modus Ponens"; ou seja, para algumas fórmulas A e B, $x = \ulcorner A \urcorner, y = \ulcorner A \to B \urcorner$ e $z = \ulcorner B \urcorner$.

Quando o sistema formal específico F em questão tenha sido fixado, as seguintes propriedades e relações podem também ser definidas:

$Axiom_F(x)$	x é o número de Gödel de um axioma não lógico de F.
$Prf_F(x, y)$	x é o número de Gödel de uma derivação (em F) da fórmula com o número de Gödel y.
$Prov_F(y)$	$\exists x Prf_F(x, y)$, ou seja, a fórmula (com o número de Gödel) y é demonstrável (derivável) em F.

Todas as propriedades e relações acima, exceto a última, demonstrabilidade, são decidíveis e podem ser, não só definidas, mas fortemente representadas em qualquer sistema suficientemente forte F.

Complemento 2 - O Lema da Diagonalização*

Autoria: Panu Raatikainen
Tradução: Guilherme A. Cardoso
Revisão: Sérgio R. N. Miranda

A prova do Lema da Diagonalização centra-se na **operação de substituição** (de um numeral por uma variável na fórmula): se uma fórmula com uma variável livre $A(x)$ e um número **n** são dados, a operação de construir uma fórmula em que o numeral para **n** substitui a (livre ocorrênica da) variável x, ou seja, $A(\underline{n})$, é puramente mecânica. Assim, ela é o análogo da operação aritmética que produz, dado o número de Gödel de uma fórmula (com uma

*RAATIKAINEN, P. "The Diagonalization Lemma", In: ZALTA, E. N. (ed.) **The Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/goedel-incompleteness/sup2.html. Acesso em: 20 jan. 2022.

The following is the translation of the supplement entry on Gödel's Incompleteness Theorems by P. Raatikainen in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/goedel-incompleteness/sup2.html. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry.Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

variável livre) $\ulcorner A(x) \urcorner$ e um número **n**, o número de Gödel da fórmula em que o numeral $\underline{n}$ substitui a variável na fórmula original, ou seja, $\ulcorner A(\underline{n}) \urcorner$. Esta última operação pode ser expressa na linguagem da aritmética. Note, em particular, que nada impede **n** de ser o próprio número de Gödel de $A(x)$, ou seja, $\ulcorner A(x) \urcorner$ (contudo, no esquema de codificação usual **n** não pode ser $\ulcorner A(\underline{n}) \urcorner$). Essa operação de substituição é aplicada aqui repetidamente.

Vamos nos referir à função de substituição aritmetizada com $subst n(\ulcorner A(x) \urcorner, \mathbf{n}) = \ulcorner A(\underline{n}) \urcorner$, e seja $S(x, y, z)$ uma fórmula que fortemente representa essa operação, como uma relação, na linguagem da teoria F. Em outros termos, S é verdadeira de x, y e z se, e somente se:

$$x = \ulcorner A(x) \urcorner, y = \mathbf{n} \text{ e } z = \ulcorner A(\underline{n}) \urcorner.$$

Novamente, nada impede-nos de considerar $subst n(\ulcorner A(x) \urcorner, \ulcorner A(x) \urcorner)$, ou, analogamente, $S(x, x, y)$.

Dada qualquer fórmula $A(x)$, podemos agora construir outra fórmula $\exists y[A(y) \wedge S(x, x, y)]$ com uma variável x. Vamos abreviá-la com $B(x)$.

Essa fórmula tem um número de Gödel, digamos, **k** = $\ulcorner B(x) \urcorner$. Ao substituirmos x em $B(x)$ pelo numeral $\underline{k}$ que a denota, obtemos $B(\underline{k})$; vamos chamar essa sentença de D. Olhando em retrospecto para a cadeia de definições, constatamos que:

$$D := B(\underline{k}) := \exists y[A(y) \wedge S(\underline{k}, \underline{k}, y)]$$

.

Se **m** $= \ulcorner B(\underline{k}) \urcorner$, então $subst n(\mathbf{k}, \mathbf{k}) = \mathbf{m}$, e (assumindo que F contém uma quantidade suficiente de aritmética; F pode então provar que o resultado da função de substituição aritmetizada é único)

$$F \vdash \forall y[S(\underline{k}, \underline{k}, y) \leftrightarrow y = \underline{m}]$$

Como **k** era o número de Gödel da fórmula $B(x)$ e **m** é o número de Gödel da sentença que resulta quando $\underline{k}$ substitui x em $B(x)$, ou seja, **m** $= \ulcorner B(\underline{k}) \urcorner$, podemos escrever isso assim:

$$F \vdash \forall y[S(\underline{k}, \underline{k}, y) \leftrightarrow y = \ulcorner B(\underline{k}) \urcorner]$$

Com um pouco de lógica, temos:

$F \vdash D \leftrightarrow \exists y[A(y) \wedge y = \ulcorner B(\underline{k}) \urcorner]$, e disso
$F \vdash D \leftrightarrow A(\ulcorner B(\underline{k} \urcorner)$, i.e.,
$F \vdash D \leftrightarrow A(\ulcorner D \urcorner)$.

E isso completa a prova.

Para as relações do Lema de Diagonalização Gödeliano com o método de diagonalização de Cantor na Teoria dos Conjuntos, *vide* GAIFMAN, 2006.

Máquinas de Turing*

Autoria: Liesbeth de Mol
Tradução: Guilherme A. Cardoso
Revisão: Sérgio R. N. Miranda

As máquinas de Turing, primeiramente descritas por Alan Turing em Turing 1936-1937, são simples dispositivos computacionais abstratos destinados a ajudar na investigação da extensão e dos limites do que pode ser computado. As "máquinas automáticas" de Turing, como ele mesmo as denominou em 1936, foram especificamente projetadas para a computação de números reais. Elas foram denominadas como "máquinas de Turing" pela primeira vez por Alonzo Church em um parecer dado ao artigo de Turing (CHURCH, 1937). Hoje se considera que elas sejam um dos modelos fundacionais da computabilidade e da

*DE MOL, LIESBETH "Turing Machines", In: ZALTA, E. N. (ed.) **Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/turing-machine/. Acesso em: 20 jan. 2022.

The following is the translation of the entry on Turing Machines by Liesbeth De Mol in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/turing-machine/. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the Stanford Encyclopedia of Philosophy, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

ciência da computação (teórica).[31]

[31] A atualização desse verbete, publicada em setembro de 2018, foi iniciada em 2015 por S. Barry Cooper, que faleceu inesperadamente logo depois de começar o trabalho. À época, as suas mudanças mais importantes consistiram em esboçar alguns dos pontos que ele estava planejando discutir na atualização do verbete. Nós utilizamos esses pontos como diretrizes para dar forma à versão atualizada de 2018. Em seguida, listamos um sumário desses pontos:

i. A descrição da atividade algorítmica é explicitamente direcionada para a máquina, e não para um assistente humano. E esta atividade é reduzida por Turing aos seus elementos mais simples, tornando mais honesto o trabalho adequado de medição da complexidade computacional ou permitindo computações mais gerais de comprimento infinito.
ii. O papel dos dados a serem manipulados é claro e relativamente flexível dentro da estrutura lógica do algoritmo, um destaque adequado ao mundo informacional de hoje.
iii. O foco em uma 'máquina', baseado nos modelos que Turing oferece para elas a partir do computador **humano** seguindo instruções, torna mais claro que a implementação de computação abstrata depende do fornecimento de um hospedeiro físico. Conversamente, a descrição clara de Turing dos modelos e o fato de estes modelos incorporarem todas as possibilidades persuadiu Gödel e outros da validade da tese Church-Turing.
iv. A predileção de Turing por programas engenhosos surgiu de uma consciência precoce da flexibilidade do equilíbrio entre hardware e software. Mas foi sua descrição do programa usando uma linguagem representável como dados legíveis pela máquina que antecipou o paradigma de programa como dado [*program-as-data paradigm*.] E este último levou à **máquina Universal** de Turing (*vide* abaixo) e à base teórica da **arquitetura lógica** de John von Neumann, tão importante na história do computador com programa armazenado de hoje.
v. Em Turing 1936-1937, a insolubilidade do *Entscheidungsproblem* de Hilbert torna-se mais claramente associada à amostragem de dados computáveis agrupados – esclarecendo a associação de incomputabilidade com descrições envolvendo o uso de quantificadores. A relação de descrições usando linguagem natural com a computabilidade subjacente, ou a falta dela, é uma área contínua de pesquisa e especulação.

Observamos aqui que a **Seção 5** foi baseada no ponto iv.

1. Definições da Máquina de Turing

1.1. A Definição de Turing

Turing introduziu as máquinas de Turing no contexto da investigação acerca dos fundamentos da matemática. Mais particularmente, ele utilizou estes dispositivos abstratos para provar que não há um método geral efetivo ou procedimento para resolver, calcular ou computar todas as instâncias do seguinte problema:

> **Entscheidungsproblem** O problema de decidir, para toda afirmação na lógica de primeira ordem (o assim denominado cálculo funcional restrito, consulte o verbete **Classical Logic**[32] da SEP para uma introdução), se ela é derivável nesta lógica ou não.

Note que na sua forma original (HILBERT; ACKERMANN, 1928), o problema foi formulado em termos de validade, ao invés de derivabilidade. Dado o teorema da completude de Gödel (GÖDEL, 1929), provar que existe um procedimento efetivo (ou não) para a derivabilidade é também uma solução para o problema na forma da validade. Para lidar com esse problema, é preciso uma noção formalizada de "procedimento efetivo" e as máquinas de Turing foram destinadas a fazer exatamente isso.

Uma máquina de Turing, ou **máquina de computação** como Turing a denominava, na definição original de Turing é uma máquina capaz de um conjunto finito de configurações $q_1, ..., q_n$ (os estados da máquina, denominados m-configurações por Turing). A máquina é suprida com uma fita, unidimensional e unidirecionalmente infinita, dividida em quadrados, cada um dos quais capaz de carregar exatamente um símbolo. Em qualquer momento, a máquina está escaneando o conteúdo de **um** quadrado r que é **em branco** (simbolizado por S_0) ou contém um símbolo $S_1, ..., S_m$, com $S_1 = 0$ e $S_2 = 1$.

A máquina é uma máquina automática (máquina-a), o que significa que

[32] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/logic-classical/. Acesso em: 20 jan. 2022.

em qualquer momento dado, o comportamento da máquina é completamente determinado pelo seu estado corrente e pelo símbolo que está sendo escaneado (a isso denomina-se a **configuração**). Isto é conhecido como a **condição de determinação** (Seção 3). Tais máquinas-a são contrastadas com as, assim denominadas, máquinas de escolha, para as quais o próximo estado depende da decisão de um dispositivo externo ou operador (TURING, 1936-1937: 232). Uma máquina de Turing é capaz de três tipos de ação:

1. Imprimir S_i, mover um quadrado para a esquerda (L) e ir para o estado q_j
2. Imprimir S_i, mover um quadrado para a direita (R) e ir para o estado q_j
3. Imprimir S_i, não se mover (N) e ir para o estado q_j

O 'programa' de uma máquina de Turing pode ser escrito como um conjunto de quíntuplas da forma:

$$q_i S_j S_{i,j} M_{i,j} q_{i,j}$$

Em que q_i é o estado corrente, S_j é o conteúdo do quadrado que está sendo escaneado, $S_{i,j}$ é o novo conteúdo do quadrado; $M_{i,j}$ especifica se a máquina deve se mover um quadrado para a esquerda, para a direita ou permanecer no mesmo quadrado, e $q_{i,j}$ é o novo estado da máquina. Tais quíntuplas são também denominadas de regras de transição de uma dada máquina. A máquina de Turing $T_{Simples}$ que, quando iniciada com uma fita em branco, computa a sequência $S_0 S_1 S_0 S_1 \ldots$ é então dada pela Tabela 1.

Tabela 1: Representação em quíntuplas de $T_{Simples}$

$; q_1 S_0 S_0 R q_2$
$; q_1 S_1 S_0 R q_2$
$; q_2 S_0 S_1 R q_1$
$; q_2 S_1 S_1 R q_1$

Observe que $T_{Simples}$ nunca vai entrar em uma configuração na qual esteja escaneando S_1 e duas das quatro quíntuplas sejam redundantes. Outro formato típico para representar máquinas de Turing e que foi também utilizado por Turing é a **tabela de transição**. A Tabela 2 dá uma tabela de transição de $T_{Simples}$.

Tabela 2: Tabela de transição para $T_{Simples}$

	S_0	S_1
q_1	S_0Rq_2	S_0Rq_2
q_2	S_1Rq_1	S_1Rq_1

Nessa tabela, as definições correntes das máquinas de Turing usualmente tem apenas um tipo de símbolos (usualmente, apenas 0 e 1; foi demonstrado por Shannon que qualquer máquina de Turing pode ser reduzida a uma máquina de Turing binária (SHANNON, 1956)). Em sua definição original das, assim denominadas, **máquinas de computação**, Turing utilizou dois tipos de símbolos: as **figuras**, que consistiam inteiramente de 0's e 1's, e os, assim denominados, símbolos do segundo tipo. Estes são diferenciados na fita da máquina de Turing utilizando-se um sistema de alternância entre quadrados de figuras e símbolos do segundo tipo. Uma sequência de quadrados alternados contém as figuras e é denominada a sequência de quadrados-F. Ela contém a **sequência computada pela máquina**; a outra é denominada a sequência de quadrados-E. As últimas são utilizadas para marcar quadrados-F e estão ali para "auxiliar a memória" (Turing 1936-1937: 232). O conteúdo dos quadrados-E é passível de alteração. Os quadrados-F, entretanto, não podem ser alterados, o que significa que não se pode implementar algoritmos pelos quais digitos previamente computados precisem ser alterados. Além disso, a máquina nunca vai imprimir um símbolo em um quadrado-F, se o quadrado-F precedente a este não foi ainda computado. Este uso de quadrados-F e quadrados-E pode ser muito útil (*vide* **Seção 2.3**), mas, como foi demonstrado por Emil L. Post, isto produz muitas complicações (*vide* **Seção 1.2**).

Há duas coisas importantes a serem observadas sobre a configuração da máquina de Turing. A primeira diz respeito à própria definição da máquina, ou seja, que a fita da máquina é potencialmente infinita. Isso corresponde a uma suposição de que a memória da máquina é (potencialmente) infinita. A segunda diz respeito à definição de Turing computável, ou seja, que uma função será Turing computável se existir um conjunto de instruções que resultará em uma máquina de Turing computando a função, independentemente da quantidade de tempo. Pode-se pensar nisso como assumindo a disponibilidade de tempo potencialmente infinito

para completar a computação.

Com estas duas suposições pretende-se garantir que a definição de computação resultante não seja demasiado estreita. Isto é, garante que nenhuma função computável deixe de ser Turing computável apenas por não haver tempo ou memória suficientes para completar a computação. Segue-se que podem haver algumas funções Turing computáveis que não podem ser executadas pelos computadores existentes, talvez porque não exista uma máquina com memória suficiente para realizar a tarefa. Algumas funções Turing computáveis podem nunca ser computáveis na prática, pois elas podem requerer mais memória do que se pode construir utilizando todos os (em número finito) átomos no universo. Entretanto, se assumirmos que um computador físico é uma realização finita da máquina de Turing, e assim que a máquina de Turing funciona como um bom modelo formal para o computador, então um resultado que mostre que uma função não é Turing computável é muito forte, pois ele implica que nenhum computador que pudéssemos construir poderia realizar a computação. Na seção 2.4, é mostrado que existem funções que não são Turing computáveis.

1.2. A Definição de Post

A definição de Turing foi padronizada por meio de (algumas das) modificações feitas por Post em (POST, 1947). Nesse artigo, Post demonstra que um certo problema da matemática conhecido como problema de Thue ou problema da palavra para semigrupos não é Turing computável (ou, nas palavras de Post, recursivamente insolúvel). A estratégia principal de Post foi mostrar que se fosse decidível, então o seguinte problema de decisão de Turing 1936-1937 seria também decidível:

> **PRINT?** O problema de decidir, para qualquer máquina de Turing M, se ela vai, em algum momento, imprimir um determinado símbolo (por exemplo, 0) ou não.

Entretanto, foi demonstrado por Turing que **PRINT?** não é Turing computável, assim o mesmo é verdadeiro do problema de Thue.

Enquanto a incomputabilidade de **PRINT?** desempenha um papel central

na prova de Post, Post acreditava que a prova de Turing fora afetada pela "expúria convenção de Turing" (POST, 1947: 9), viz. o sistema de quadrados-F e quadrados-E. Assim, Post introduziu uma versão modificada da máquina de Turing. As diferenças mais importantes entre as definições de Post e Turing são as seguintes:

1. A máquina de Turing de Post, quando em um dado estado, ou imprime ou se move, e assim suas regras de transição são mais 'atômicas' (ela não contém a operação composta de mover e imprimir). Isto resulta em uma notação quádrupla das máquinas de Turing, em que cada quádrupla tem uma das três formas da Tabela 3:

Tabela 3: A notação Quádrupla de Post

$; q_i S_j S_{i,j} q_{i,j}$
$; q_i S_j L q_{i,j}$
$; q_i S_j R q_{i,j}$

2. A máquina de Turing de Post tem apenas um tipo de símbolo e, assim, não depende do sistema de quadrados-F e quadrados-E de Turing.
3. A máquina de Turing de Post tem uma fita bidirecionalmente infinita.
4. A máquina de Turing de Post para quando atinge um estado para o qual nenhuma ação é definida.

Observe que a reformulação de Post da máquina de Turing está muito enraizada em seu artigo (POST, 1936). (Algumas das) modificações de Post da definição de Turing tornaram-se parte da definição de máquina de Turing nos trabalhos padronizados, tais como Kleene (1952) e Davis (1958). Desde esta época, várias definições (logicamente equivalentes) foram introduzidas. Hoje, as definições padrão das máquinas de Turing são, em alguns aspectos, mais próximas das máquinas de Turing de Post do que das máquinas de Turing. No que segue, iremos utilizar uma variante da definição padrão de Minsky (1967), que usa a notação quíntupla, mas não tem quadrados-E, quadrados-F e inclui um estado especial de parada H. Ainda, teremos apenas duas operações de movimento, quais sejam, L e R, e assim a ação por meio da qual a máquina

meramente imprime não é utilizada. Quando a máquina é iniciada, a fita está completamente em branco, excepto por uma porção finita da fita. Note que o quadrado em branco pode também ser representado como um quadrado contendo o símbolo S_0 ou simplesmente 0. O conteúdo finito da finita será também denominado o **texto** da fita.

1.3. A Definição Formalizada

Toda essa conversa sobre "fita" e um "cabeçote de leitura-escrita" destina-se a auxiliar as intuições (e revela algo da época em que Turing estava escrevendo), mas não desempenha nenhum papel importante na definição das máquinas de Turing. Nas situações em que uma análise formal das máquinas de Turing é exigida, é apropriado exprimir as definições do maquinário e do programa em termos mais matemáticos. Em termos puramente formais, uma máquina de Turing pode ser especificada como uma quádrupla $T = (Q, \Sigma, s, \delta)$, na qual:

- Q é um conjunto finito de estados q.
- Σ é um conjunto finito de símbolos.
- s é o estado inicial $s \in Q$.
- δ é uma função de transição determinando o próximo movimento:

$$\delta : (Q \times \Sigma) \longrightarrow (\Sigma \times \{L, R\} \times Q)$$

A função de transição da máquina T é uma função que leva de estados de computação para estados de computação. Se $\delta(q_i, S_j) = (S_{i,j}, D, q_{i,j})$, então, quando a máquina está no estado q_i, lendo o símbolo S_j, T substitui S_j por $S_{i,j}$, ela se move na direção $D \in \{L, R\}$ e vai para o estado $q_{i,j}$.

1.4. Descrevendo o Comportamento de uma Máquina de Turing

Introduzimos uma representação que nos permite descrever o comportamento ou dinâmica de uma máquina de Turing T_n fiando-nos na notação da **configuração completa** (TURING, 1936-1937: 232), também conhecida hoje como **descrição instantânea** (DI) (DAVIS, 1982: 6). Em qualquer estágio da

computação de T_i, a sua DI é dada por:

(1) o conteúdo da fita, ou seja, seu texto.
(2) a localização do cabeçote de leitura.
(3) o estado interno da máquina.

Assim, dada uma máquina de Turing T no estado q_i escaneando o símbolo S_j, sua DI é dada por Pq_iS_jQ, em que P e Q são textos finitos à esquerda e à direita do quadrado contendo o símbolo S_j. A Figura 1 dá uma representação visual de uma DI de uma máquina de Turing T no estado q_i escaneando a fita.

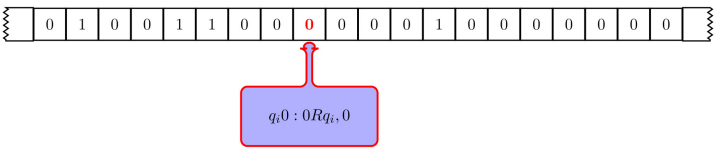


Figura 1: uma configuração completa de alguma máquina de Turing T

Assim, a notação nos permite capturar o desenvolvimento do comportamento da máquina de Turing e sua fita através de suas DI's consecutivas. A Figura 2 nos dá algumas das primeiras DI's consecuticas de $T_{Simples}$ utilizando uma representação gráfica.

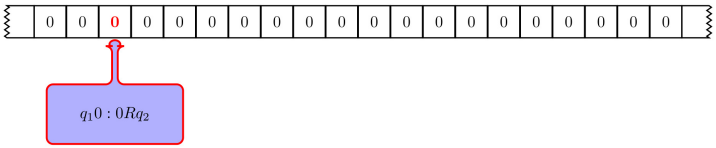


Figura 2: representação gráfica da dinâmica de $T_{Simples}$

A animação pode ser iniciada clicando-se na imagem[33]. Pode-se ainda explicitamente imprimir as DI's consecutivas, utilizando as suas representações

[33]N.T.: o leitor pode conferir a animação no endereço online do arquivo.

simbólicas. Esse procedimento resulta em um diagrama de estado espacial de uma máquina de Turing. Assim, para $T_{Simples}$ obtemos (note que $\overline{0}$ significa a repetição infinita de 0s):

$$\begin{gathered}\overline{0}q_1\mathbf{0}\overline{0}\\ \overline{0}0q_2\mathbf{0}\overline{0}\\ \overline{0}01q_1\mathbf{0}\overline{0}\\ \overline{0}010q_2\mathbf{0}\overline{0}\\ \overline{0}0101q_1\mathbf{0}\overline{0}\\ \overline{0}01010q_2\mathbf{0}\overline{0}\\ \vdots\end{gathered}$$

2. Computando com Máquinas de Turing

Como foi explicado na **Seção 1.1**, as máquinas de Turing foram originalmente destinadas a formalizar a noção de computabilidade para enfrentar um problema fundamental da matemática. Independentemente de Turing, Alonzo Church deu uma formulação diferente, mas logicamente equivalente (consulte a **Seção 4**). Hoje, os cientistas da computação, em sua grande maioria, concordam que a noção formal de Turing, ou qualquer outra noção logicamente equivalente, captura **todos** os problemas computáveis, ouseja, para qualquer problema computável, existe uma máquina de Turing que o computa. Isso é conhecido como a **tese Church-Turing**, **tese de Turing** (quando a referência é apenas ao trabalho de Turing) ou **tese de Church** (quando a referência é apenas ao trabalho de Church).

Esta tese, se aceita, implica que qualquer problema não computável por máquinas de Turing não é computável por nenhum meio finito. De fato, como era uma ambição de Turing capturar "[todos] os possíveis processos que podem ser realizados na computação de um número" (TURING, 1936-1937: 249), segue-se que, se aceitamos a análise de Turing:

- Qualquer problema que não seja computável por uma máquina de Turing não é "computável" em sentido absoluto (ao menos, absoluto relativamente aos humanos, consulte a **Seção 3**).

- Para qualquer problema que acreditamos que seja computável, deveríamos ser capazes de construir uma máquina de Turing que o computa. Colocando nas palavras de Turing:

> O que estou pontuando é que [as] operações [de uma máquina de computação] incluem todas aquelas que são utilizadas na computação de um número. (TURING, 1936-1937: 231)

Na seção a seguir serão dados exemplos que ilustram o poder computacional e os limites do modelo das máquinas de Turing. A Seção 3 então discute algumas questões filosóficas relacionadas à tese de Turing.

2.1. Alguns Exemplos (Simples)

Para falar sobre máquinas de Turing que fazem algo de útil da perspectiva humana, teremos de fornecer uma interpretação dos símbolos registrados na fita. Por exemplo, se queremos designar uma máquina que compute alguma função matemática, digamos a adição, então precisamos descrever como interpretar os uns e zeros que aparecem na fita como números.

Nos exemplos que se seguem, vamos representar o número n como um bloco de $n + 1$ cópias do símbolo '1' na fita. Assim, representamos o número 0 como um único '1' e o número 3 como um bloco de quatro '1's. Isto é denominado como **notação unária**.

Precisamos também fazer algumas suposições acerca da configuração da fita quando a máquina é iniciada e quando ela termina, de modo a interpretar a computação. Vamos assumir que se a função a ser computada requer n argumentos, então a máquina de Turing vai iniciar com seu cabeçote escaneando o '1' mais à esquerda de uma sequência de n blocos de '1's. Os blocos de '1's representando os argumentos devem ser separados por uma única ocorrência do símbolo '0'. Por exemplo, para computar a soma $3 + 4$, a máquina de Turing vai iniciar na configuração mostrada na Figura 3.

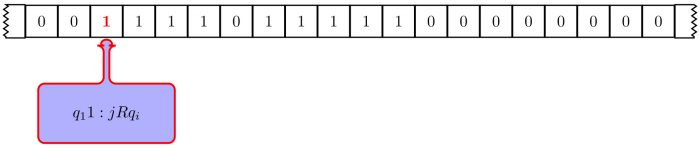


Figura 3: Configuração inicial para uma computação sobre dois números n e m

Aqui a suposta máquina da adição toma dois argumentos representando os números a serem adicionados, iniciando no 1 mais à esquerda do primeiro argumento. Os argumentos são separados por um único 0 como requerido, o primeiro bloco contém quatro '1's, representando o número 3, e o segundo bloco contém cinco '1's, representando o número 4.

Uma máquina deve terminar em uma configuração padrão também. Deve haver um único bloco de símbolos (uma sequência de 1s representando algum número ou um símbolo representando outro tipo de *output*) e a máquina deve estar escaneando o símbolo mais a esquerda da sequência. Se a máquina computa corretamente a função, então esse bloco deve representar a resposta correta.

Adotar esta convenção para a configuração de término de uma máquina de Turing significa que podemos compor máquinas, indetinficando o estado final de uma máquina com o estado inicial da próxima máquina.

A adição de dois números n e m.

A Tabela 4 nos dá a tabela de transição de uma máquina de Turing T_{Ad_2}, a qual adiciona dois números naturais n e m. Assumimos que as máquinas são iniciadas no estado q_1, escaneando o 1 mais à esquerda de $n + 1$.

Tabela 4: Tabela de transição para T_{Ad_2}

	0	1
q_1	/	$0Rq_2$
q_2	$1Lq_3$	$1Rq_2$
q_3	$0Rq_4$	$1Lq_3$
q_4	/	$0Rq_{halt}$

A ideia de fazer uma adição com máquinas de Turing por representação unária consiste em deslocar o número n mais à esquerda um quadrado para a direita. Isto é feito apagando-se o 1 mais à esquerda de $n+1$ (isso é feito no estado q_1) e então alterando o 0 entre $n+1$ e $m+1$ para 1 (estado q_2). Assim, obtemos $n+m+2$ e ainda precisamos apagar o 1 adicional. Isto é feito apagando-se o 1 mais à esquerda (estados q_3 e q_4). A Figura 4 mostra essa computação para $3+4$.

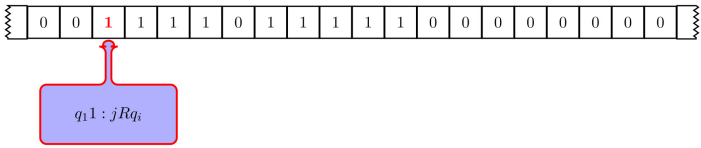


Figura 4: A computação de $3+4$ por T_{Ad_2}

Adição de n números

Podemos generalizar T_{Ad_2} para uma máquina de Turing T_{Ad_i} para a adição de um número arbitrário i de inteiros $n_1, n_2, ..., n_j$. Assumimos novamente que a máquina se inicia no estado q_1 escaneando o 1 mais à esquerda de n_1+1. A tabela de transição para uma tal máquina T_{Ad_i} é dada na Tabela 5.

Tabela 5: Tabela de transição para T_{Ad_i}

	0	1
q_1	/	$0Rq_2$
q_2	$1Rq_3$	$1Rq_2$
q_3	$0Lq_6$	$1Lq_4$
q_4	$0Rq_5$	$1Lq_4$
q_5	/	$0Rq_1$
q_6	$0Rq_{halt}$	$1Lq_6$

A máquina T_{Ad_i} utiliza aquele princípio de deslocamento dos adendos para a direita que foi também utilizado por T_{Ad_2}. Mais particularmente, T_{Ad_i} computa a soma de $n_1 + 1, n_2 + 1, ..., n_i + 1$, da esquerda para a direita, ou seja, ela computa esta soma da seguinte maneira:

$$\begin{aligned} N_1 &= n_1 + n_2 + 1 \\ N_2 &= N_1 + n_3 \\ N_3 &= N_2 + n_4 \\ &\vdots \\ N_i &= N_{i-1} + n_i + 1 \end{aligned}$$

A diferença mais importante entre T_{Ad_2} e T_{Ad_i} é que T_{Ad_i} precisa verificar se o adendo mais à esquerda N_j é igual a N_i (com $1 < j \leq i$). Isto é realizado checando se o primeiro 0 à direita de N_j é seguido por outro 0 ou não (estados q_2 e q_3). Se este não é o caso, então existe ao menos mais um adendo n_{j+1} a ser adicionado. Observe que, assim como foi o caso com T_{Ad_2}, a máquina precisa apagar um 1 adicional do adendo n_{j+1}, o que é feito no estado q_5. Ela então retorna ao estado estado q_1. Se, por outro lado, $N_j = N_i$, a máquina se desloca para o 1 mais à esquerda de $N_i = n_1 + n_2 + ... + n_i + 1$ e para.

2.2. Números e Problemas Computáveis

O artigo original de Turing se ocupa dos números (reais) computáveis. Um número (real) é Turing computável se existe uma máquina de Turing que computa uma aproximação arbitrariamente precisa desse número. Todos os números algébricos (raízes de polinômios com coeficientes algébricos) e muitas constantes matemáticas transcendentais, tais como e e π, são Turing-computáveis. Turing forneceu vários exemplos de classes de números computáveis por máquinas de Turing (*vide* a seção 10 de Turing (1936-1937), "Exemplos de Grandes Classes de números computáveis"), como um argumento

heurístico mostrando que uma ampla diversidade de classes de números podem ser computados por máquinas de Turing.

Podemos nos perguntar, entretanto, em que sentido a computação com números, ou seja, cálculo, captura os problemas computáveis não numéricos, e assim, como as máquinas de Turing capturam todos os procedimentos gerais e efetivos que determinam se algo é o caso ou não. Exemplos de tais problemas são:

- "decidir, para um x qualquer, se x denota um primo ou não".
- "decidir, para um x qualquer, se x descreve uma máquina de Turing ou não".

Em geral, esses problemas são da forma:

- "decidir, para um x qualquer, se x tem a propriedade X ou não".

Um importante desafio tanto nos avanços teóricos quanto concretos da computação (muitas vezes na interface com outras disciplinas) tornou-se o problema de fornecer uma interpretação de X, tal que, X possa ser abordado computacionalmente. Para dar apenas um exemplo concreto, nas práticas computacionais diárias pode ser importante ter um método para decidir, para qualquer "fonte" digital, se ela é confiável ou não. Portanto, é preciso uma interpretação computacional da confiança.

A "função característica" de um predicado é uma função que dá os valores VERDADEIRO ou FALSO quando recebe argumentos apropriados. Para que tais funções fossem computáveis, Turing se baseou na ideia de Gödel de que problemas desse tipo podem ser codificados como problemas sobre números (*vide* **Os teoremas da Incompletude de Gödel** e a próxima **Seção 2.3**). Nas palavras de Turing:

> A expressão "existe um processo geral para determinar..." foi usada [aqui] [...] como sendo equivalente a "existe uma máquina que determina...". Este uso pode ser justificado se, e somente se, pudermos justificar a nossa definição de "computável". Pois cada um desses problemas

> de “processo geral” pode ser expresso como um problema relacionado ao processo geral de determinar se um dado inteiro n tem uma propriedade $G(n)$ [por exemplo, $G(n)$ pode significar “n é satisfatório” ou “n é a representação de Gödel de uma fórmula demonstrável”], e isto é equivalente a computar um número cuja n-ésima figura é 1 se $G(n)$ é verdadeira e 0 se é falsa. (TURING, 1936-1937: 248)

A possibilidade de codificar os problemas de “processo geral” como problemas numéricos é essencial para a construção de Turing da máquina de Turing universal. Ademais, o seu uso dentro de uma prova mostra que existem problemas que não podem ser computados por uma máquina de Turing.

2.3. A Máquina de Turing Universal

A máquina de Turing universal, que foi construída para provar a incomputabilidade de certos problemas, é, grosso modo, uma máquina de Turing capaz de computar o que qualquer outra máquina de Turing computa. Assumindo que a noção de máquina de Turing captura completamente a computabilidade (e assim que a tese de Turing é válida), isto implica que qualquer coisa que possa ser “computada”, pode também ser computada por aquela máquina universal. Conversamente, qualquer problema que não seja computável pela máquina universal deve ser considerado incomputável.

Este é o poder retórico e teórico do conceito de máquina universal: que um dispositivo formal relativamente simples captura todos “os processos possíveis que podem ser realizados na computação de um número” (TURING, 1936-1937). Esta é também uma das principais razões pelas quais Turing tem sido retrospectivamente identificado como um dos fundadores da ciência da computação (*vide* a **Seção 5**).

Mas como construir uma máquina universal U a partir do conjunto de operações básicas que dispomos? A abordagem de Turing é a de construir uma máquina U que é capaz de (1) ‘entender’ o programa de **qualquer** outra máquina T_n e, baseado nesse ‘entendimento’, (2) ‘imitar’ o comportamento de T_n. Para

esta finalidade, é necessário um método que nos permita tratar o programa e o comportamento de T_n de forma intercambiável, já que ambos os aspectos são manipulados na mesma fita e pela mesma máquina. Isto é feito por Turing em dois passos básicos: o desenvolvimento de (1) um método notacional e de (2) um conjunto de funções elementares que tratam esta notação - independente de se ela está formalizando o programa ou o comportamento de T_n - como texto a ser comparado, copiado, apagado, etc. Em outras palavras, Turing desenvolve uma técnica que permite tratar programa e comportamento no mesmo nível.

Intercambialidade de programa e comportamento: uma notação

Dada uma máquina T_n, a ideia básica de Turing consiste em construir uma máquina T'_n que, ao invés de imprimir diretamente o resultado de T_n, imprime a sucessão de configurações completas ou descrições instantâneas de T_n. Para tal feito, T'_n:

> [...] poderia ser feita de modo a depender que as regras de operação [...] de [T_n] estejam escritas em algum lugar dela mesma [...] cada passo poderia ser realizado por referência a tais regras. (TURING, 1936–7: 242)

Em outras palavras, pelo fato de ter o programa de T_n escrito na sua própria fita, T'_n imprime a sucessão de configurações completas de T_n. Assim, Turing precisa de um método notacional que torne possível 'capturar' dois aspectos distintos de uma máquina de Turing em uma mesma fita, de tal modo que ambos possam ser tratados *pela mesma máquina*, quais sejam:

(1) sua descrição em termos **do que ela deveria fazer** - a notação quíntupla f
(2) sua descrição em termos **do que ela está fazendo** - a notação da configuração completa.

Assim, um passo inicial e possivelmente mais essencial na construção de U são as notações quíntupla e de configuração completa, juntamente com a ideia de colocá-los na mesma fita. Mais particularmente, a fita é dividida em duas regiões

que denominaremos aqui por regiões A e B. A região A contém uma notação do 'programa' de T_n; e a região B, uma notação para a sucessão de configurações completas de T_n. No artigo de Turing, elas eram separadas pelo símbolo adicional "::".

Para simplificar a construção de U e de modo a codificar qualquer máquina de Turing com um único número, Turing desenvolve uma terceira notação que permite expressar as quíntuplas a as configurações completas apenas com letras. Isto é determinado pelo seguinte [Note que nós estamos utilizando a codificação original de Turing. Claro, existe uma ampla variedade de codificações possíves, incluindo a codificação binária]:

- Substitua cada estado q_i em uma quíntupla de T_n por

$$D\underbrace{A\ldots A}_{i},$$

  Assim, por exemplo, q_3 torna-se $DAAA$.
- Substitua cada símbolo S_j em uma quíntupla de T_n por

$$D\underbrace{C\ldots C}_{j},$$

  Assim, por exemplo, S_1 torna-se DC.

Utilizando este método, cada quíntupla de alguma máquina de Turing T_n pode ser expressa em termos de uma sequência de letras maiúsculas e, assim, o 'programa' de qualquer máquina T_n pode ser expresso pelo conjunto de símbolos A, C, D, R, L, N e ;. Esta é a, assim denominada, Descrição Padrão (D.P.) de uma máquina de Turing. Assim, por exemplo, a D.P. de $T_{Simples}$ é:

$$;DADDRDAA;DADCDRDAA;DAADDCRDA;DAADCDCRDA$$

Isto é, essencialmente, a versão de Turing da **numeração de Gödel**. De fato, como mostra Turing, pode-se facilmente obter uma representação por

descrição numérica ou **Número de Descrição** (N.D.) de uma máquina de Turing T_n substituindo-se:

- "A" por "1"
- "C" por "2"
- "D" por "3"
- "L" por "4"
- "R" por "5"
- "N" por "6"
- ";" por "7"

Assim, o N.D. de $T_{Simples}$ é:

$$73133531173131353117311331531731131315 31$$

Note que toda máquina T_n tem um único N.D.; um N.D. representa uma, e apenas uma, máquina.

Claramente, o método utilizado para determinar a D.P. de alguma máquina T_n pode ser também utilizado para escrever a sucessão de configurações completas de T_n. Utilizando ":" como um separador entre as sucessivas configurações completas, as primeiras configurações completas de $T_{Simples}$ são:

$$: DAD : DDAAD : DDCDAD : DDCDDAAD : DDCDDCDAD$$

Intercambialidade de programa e comportamento:
um conjunto básico de funções

O primeiro passo na construção de Turing de U consiste em encontrar um método notacional para escrever o programa e as sucesivas configurações completas de alguma máquina T_n na mesma fita de alguma outra máquina T'_n. Entretanto, U também deveria ser capaz de "emular" o programa de T_n como escrito na região A, de tal modo que possa realmente escrever as suas

sucessivas configurações completas na região B. Além disso, deveria ser possível "retirar e trocar [...] [as regras de operações de alguma máquina de Turing] por outras" (TURING, 1936-1937: 242). Ou seja, U deveria ser capaz não apenas de calcular, mas também de computar. Essa é uma questão com a qual outros, com seus próprios dispositivos formais, também lidaram, como Church, Gödel e Post. Por exemplo, U deveria ser capaz de "reconhecer" se está na região A ou B, e deveria ser capaz de determinar se uma certa sequência de símbolos é o próximo estado q_i a ser executado ou não.

Isto é realizado por Turing através da construção de uma sequência de problemas Turing computáveis, tais como:

- encontrar a ocorrência mais à esquerda ou mais à direita de uma sequência de símbolos.
- marcar uma sequência de símbolos por meio de algum símbolo a (lembre-se que Turing utiliza dois tipos de quadrados alternantes).
- comparar duas sequências de símbolos.
- copiar uma sequência de símbolos.

Turing desenvolve uma técnica notacional, denominada **tabelas esqueleto**, para estas funções que servem como uma espécie de notação abreviativa para uma tabela completa de máquinas de Turing. Mas essas funções podem ser facilmente utilizadas para construir máquinas mais complicadas a partir das anteriores. A técnica é originada da técnica recursiva de composição (*vide* o verbete **Recursive Functions**[34] da SEP).

Para ilustrar como tais funções são Turing computáveis, vamos discutir uma delas em mais detalhes, precisamente a função de comparação. Ela é construída com base em várias outras funções Turing computáveis, que são construídas umas sobre as outras. Para entender como essas funções funcionam, lembre-se que Turing utilizava um sistema de alternância de quadrados-E e quadrados-F, em que os quadrados-F contêm as quíntuplas e as configurações completas, enquanto os quadrados-E são usados para marcar certas partes da fita da máquina. Para a comparação de duas sequências S_1 e

[34] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/recursive-functions/. Acesso em: 20 jan. 2022

S_2, cada símbolo de S_1 será marcado por algum símbolo a e cada símbolo de S_2 será marcado por algum símbolo b.

Turing definiu nove funções distintas para mostrar como a função de comparação pode ser computada por máquinas de Turing:

$FIND(q_i, q_j, a)$: essa função de máquina procura a ocorrência mais à esquerda de a. Se a é encontrado, a máquina se desloca para o estado q_i. De outro modo, se desloca para o estado q_j. Isso é feito da seguinte maneira: primeiro a máquina se desloca para o início da fita (indicado por uma marca especial). Então ela se desloca para a direita até encontrar a ou alcançar o símbolo mais à direita da fita.

$FINDL(q_i, q_j, a)$: o mesmo que $FIND$, mas depois de ter encontrado a, a máquina se desloca um quadrado para a esquerda. Isso é utilizado em funções que precisam computar sobre os símbolos nos quadrados-F que estão marcados por símbolos a nos quadrados-E.

$ERASE(q_i, q_j, a)$: a máquina computa $FIND$. Se a é encontrado, ela apaga a e vai para o estado q_i. De outro modo, ela vai para o estado q_j.

$ERASE\_ALL(q_j, a) = ERASE(ERASE\_ALL, q_j, a)$: a máquina computa $ERASE$ sobre a repetidamente, até que todas as ocorrências de a tenham sido apagadas. Então, ela se desloca para q_j.

$EQUAL(q_i, q_j, a)$: a máquina checa se o símbolo ocorrente é a ou não. Se sim, ela se desloca para o estado q_i. Caso contrário, ela se desloca para q_j.

$CMP\_XY(q_i, q_j, b) = FINDL(EQUAL(q_i, q_j, x), q_j, b)$: qualquer que seja o símbolo x ocorrente, a máquina computa $FINDL$ sobre b (e assim procura pelo símbolo marcado por b). Se há um símbolo y marcado com b, a máquina computa $EQUAL$ sobre x e y; de outro modo, ela vai para o estado q_j. Em outras palavras, $CMP\_XY(q_i, q_j, b)$ compara se o símbolo ocorrente é o mesmo que o símbolo mais à esquerda marcado com b.

$COMPARE\_MARKED(q_i, q_j, q_n, a, b)$: a máquina checa se os símbolos mais à esquerda marcados por a e b respectivamente são o mesmo. Se não há símbolos marcado por a nem por b, a máquina vai para o estado q_n: se há um símbolo marcado por a e um símbolo marcado por b

e eles são o mesmo, a máquina vai para o estado q_i, de outro modo, a máquina vai para o estado q_j. A função é computada como $FINDL(CMP\_XY(q_i, q_j, b), FIND(q_j, q_n, b), a)$.

$COMPARE\_ERASE(q_i, q_j, q_n, a, b)$: o mesmo que $COMPARE\_MARKED$, mas quando os símbolos marcados por a e b são o mesmo, as marcas a e b são apagadas. Isto é realizado computando $ERASE$ primeiramente em a depois em b.

$COMPARE\_ALL(q_j, q_n, a, b)$: a máquina compara as sequências A e B marcadas com a e b, respectivamente. Isto é feito repetindo-se a computação de $COMPARE\_ERASE$ sobre a e b. Se A e B são iguais, todos os a's e b's terão sido apagados e a máquina vai para o estado q_j; de outro modo, a máquina vai para o estado q_n. Esta máquina é computada por

$$COMPARE\_ERASE(COMPARE\_ALL(q_j, q_n, a, b), q_j, q_n, a, b)$$

Assim, por invocação recursiva, $COMPARE\_ALL$.

De modo semelhante, Turing define as seguintes funções:

$COPY(q_i, a)$: copie a sequência de símbolos marcada com a's à direita da última configuração completa e apague as marcas.

$COPY + n(q_i, a_1, a_2, ..., a_n)$ copie as sequências marcadas de a_1 a a_n à direita da última configuração completa e apague as marcas.

$REPLACE(q_i, a, b)$: substitua todas as letras a por b.

$MARK\_NEXT\_(q_i, a)$: marque a primeia configuração $q_i Sj$ à direita do cabeçote da máquina com a letra a.

$FIND\_RIGHT(q_i, a)$: encontre o símbolo a mais à direita.

Utilizando as funções básicas $COPY$, $REPLACE$ e $COMPARE$, Turing constrói a máquina de Turing universal.

Abaixo está um esboço da máquina de Turing universal, indicando como essas funções básicas de fato tornam possível a computação universal. Supõe-se que na inicialização, U tem na sua fita a D.P. de alguma máquina de Turing T_n. Lembre-se que Turing utiliza o princípio de alternância de quadrados-E e

quadrados-F e, assim, por exemplo, a D.P. de $T_{Simples}$ será escrita na fita de U como:

$$;\_D\_A\_D\_D\_R\_D\_A\_A\_;\_D\_A\_D\_C\_D\_R\_D\_A\_A\_;\_ \\ D\_A\_A\_D\_D\_C\_R\_D\_A\_;\_D\_A\_A\_D\_C\_D\_C\_R\_D\_A\_$$

em que "_" indica um quadrado-E não marcado.

- $INIT$: À direita da quíntupla mais à direita de T_n, U imprime :: _ : $\_D\_A\_$, em que _ indica um quadrado-E não marcado.
- $FIND\_NEXT\_STATE$: A máquina primeiramente marca (1) com um y a configuração $q_{CC,i}S_{CC,j}$ da configuração completa mais à direita (portanto, a última) computada por U na parte B da fita e (2) com um x a configuração $q_{q,m}S_{q,n}$ da quíntupla mais à esquerda que não é precedida por um ponto e vírgula marcado (com a letra z) na parte A da fita. As duas coonfigurações são comparadas. Se elas são idênticas a máquina vai para $MARK\_OPERATIONS$, se não, ela marca o ponto e vírgula que precede $q_{q,m}S_{q,n}$ com um z e vai para $FIND\_NEXT\_STATE$. Isso é facilmente realizado utilizando-se a função $COMPARE - ALL$, o que significa que, qualquer que seja o resultado da comparação, as marcas x e y serão apagadas. Por exemplo, suponha que $T_n = T_{Simples}$ e que a última configuração completa de $T_{Simples}$ conforme computada por U é:

(1) $$: \_\underbrace{D\_}_{S_0}\underbrace{D\_C\_}_{S_1}\underbrace{D\_}_{S_0}\underbrace{D\_A\_A\_}_{q_2}\underbrace{D\_}_{S_0}$$

Então, U move-se para a região A e determina que a quíntupla correspondente é:

(2) $$\underbrace{D\_A\_A\_}_{q_2}\underbrace{D\_}_{S_0}\underbrace{D\_C\_}_{S_1}\underbrace{R\_}\underbrace{D\_A\_}_{q_1}$$

- $MARK\_OPERATIONS$: A máquina U marca as operações que ela precisa executar para computar a próxima configuração completa de T_n. As operações de imprimir e mover-se (L, R, N) são marcadas com u e o próximo estado com y. Todas as marcas z são apagadas. Continuando com o nosso exemplo, U irá marcar (2) da seguinte maneira:

$$D\_A\_A\_D\_DuCuRuDyAy$$

- $MARK\_COMPCONFIG$: A última configuração completa de T_n conforme computada por U é marcada em quatro regiões: a configuração $q_{CC,i}S_{CC,j}$ ela mesma fica não marcada; o símbolo que a precede é marcado com um x e os símbolos restantes à esquerda são marcados com v. Finalmente, todos os símbolos à direita, se houver, são marcados com w e um ":" é impresso à direita do símbolo mais à direita para indicar o começo da próxima configuração completa de T_n a ser computada por U. Continuando com o nosso exemplo, (1) será marcada por U da seguinte maneira:

$$\underbrace{Dv}_{S_0}\underbrace{DvCv}_{S_1}\underbrace{Dx}_{S_0}\underbrace{D\_A\_A\_}_{q_2}\underbrace{D\_}_{S_0} : \_$$

  U então vai para $PRINT$
- $PRINT$: É determinado se, nas instruções que foram marcadas em $MARK\_OPERATIONS$, existe uma operação Print 0 ou Print 1. Se este for o caso, 0: ou respectivamente 1: é impresso à direita da última configuração completa. Isso não é uma função necessária, mas Turing insistia que U imprimisse não apenas a configuração completa (codificada) computada por T_n mas também o atual número real (binário) computado por T_n.
- $PRINT_COMPLETE_CONFIGURATION$L U imprime a próxima configuração completa e apaga todas as marcas u, v, w, x, y. Ela então retorna a $FIND\_NEXT\_STATE$. Primeiramente, U pela

letra u mais à direita, para checar qual movimento (R, L, N) é necessário, e apaga a marca u para R, L, N. Dependendo do valor L, R ou N, vai então escrever a próxima configuração completa, aplicando $COPY_5$ a u, v, w, x, y. A operação de movimento (L, R, N) é descrita pela combinação particular de u, v, w, x, y:

Quando $\sim L : COPY_5(FIND\_NEXT\_STATE, v, y, x, u, w)$
Quando $\sim R : COPY_5(FIND\_NEXT\_STATE, v, x, u, y, w)$
Quando $\sim N : COPY_5(FIND\_NEXT\_STATE, v, x, y, u, w)$

Seguindo o nosso exemplo, como $T_{Simples}$ precisa se mover para a direita, a nova configuração completa mais à direita de $T_{Simples}$ escrita na fita de U é:

$$\underbrace{D\_}_{S_0}\ \underbrace{D\_C\_}_{S_1}\ \underbrace{D\_}_{S_0}\ \underbrace{D\_C\_}_{S_1}\ \underbrace{D\_A\_}_{q_1}$$

Como, para essa configuração completa, o quadrado a ser escaneado por $T_{Simples}$ é um que não fora incluído em configurações completas anteriores (viz. $T_{Simples}$ foi além do ponto prévio mais à direita), a configuração completa como descrita por U é incompleta. Esse pequeno defeito foi corrigido por Post (POST 1947) pela inclusão de uma instrução adicional na função utilizada para marcar a configuração completa na próxima rodada.

Como é claro, a máquina de Turing universal requer, de fato, que o programa e 'dados' produzidos por este programa sejam manipulados de maneira intercambiável, ou seja, o programa e seus produtos são pareados e tratado da mesma maneira, como sequências de letras a serem copiadas, marcadas, apagadas e comparadas.

A construção particular de Turing é muito intrincada, com sua

dependência nos quadrados-E e quardados-F, o uso de uma enorme quantidade de símbolos e o uso de uma notação arcana para descrever as diferentes funções discutidas acima. Desde 1936, várias modificações e simplificações foram implementadas. A remoção da diferença entre quadrados-E e quadrados-F já foi discutida na **Seção 1.2** e foi demonstrado por Shannon que qualquer máquina de Turing, incluindo a máquina universal, pode ser reduzida a uma máquina de Turing binária (SHANNON 1956). Desde os anos de 1950, tem havido alguma procura por qual seria o menor dispositivo universal possível (com relação ao número de estados e símbolos) e algumas máquinas de Turing universais bem "pequenas" foram encontradas. Esses resultados são usualmente alcançados apelando-se a outros modelos equivalentes de computabilidade, tais como, por exemplo, as máquinas de Post. Para um panorama sobre a pesquisa de pequenos dispositivos universais, *vide* MARGENSTERN, 2000; WOODS; NEARY, 2009.

2.4. O Problema da Parada e o Entscheindungsproblem

Conforme explicado, o propósito do artigo de Turing era mostrar que o *Entscheidungsproblem* para a lógica de primeira ordem é incomputável. O mesmo resultado foi estabelecido independentemente por Church (1936a, 1936b), utilizando um tipo diferente de dispositivo formal que é logicamente equivalente a uma máquina de Turing (*vide* **Seção 4**). O resultado foi muito contra o que Hilbert esperava alcançar com o seu programa finitário e formalista. De fato, ao lado dos resultados de incompletude de Gödel, eles quebraram muito do sonho de Hilbert em esvaziar a matemática de todo *Ignorabimus*, o que foi explicitamente expresso nas seguintes palavras de Hilbert :

> A verdadeira razão pela qual Comte não conseguiu encontrar um problema insolúvel está, na minha opinião, na afirmação de que não existe problema insolúvel. Em vez do estúpido *Ignorabimus*, nossa solução deveria ser: nós devemos conhecer. Nós conheceremos. (1930: 963) [traduzido pelo autor]

Observe que a solucionabilidade à qual Hilbert está se referindo aqui diz respeito à solucionabilidade de problemas matemáticos em geral, não apenas os mecanicamente solucionáveis. Entretanto, mostra-se em Mancosu et al. (2009, p.94) que esse objetivo geral de resolver todos os problemas matemáticos sustenta-se em duas convições particulares de Hilbert, quais sejam, (1) que os axiomas da teoria dos números são completos e (2) que não existem problemas indecidíveis na matemática.

Provas diretas e indiretas de problemas de decisão incomputáveis

Como se pode demonstrar, para um problema de decisão particular D_i, que ele não é computável? Existem dois métodos principais:

Prova Indireta: tome algum problema D_{incomp} já reconhecido como sendo incomputável e mostre que o problema inicial se "reduz" a D_{incomp}.

Prova Direta: demonstre diretamente a incomputabilidade de D_i assumindo alguma versão da tese Church-Turing.

Hoje, o primeiro método é usualmente confiável, mas evidentemente na ausência de um problema D_{incom}, Turing mas também Church e Post (*vide* **Seção 4**) tiveram de se apoiar em uma abordagem direta.

A noção de redutibilidade teve a sua origem nos trabalhos de Turing e Post, que consideravam muitas variantes da computabilidade (POST, 1946; TURING, 1939). O conceito foi posteriormente apropriado na contexto da teoria da complexidade computacional e hoje é um dos conceitos básicos tanto da computabilidade quanto da teoria da complexidade computacional (ODIFREDDI, 1989; SIPSER, 1996). Grosso modo, uma redução de um problema D_i a um problema D_j consiste em fornecer um procedimento efetivo para traduzir toda instância $d_{i,m}$ do problema D_i para uma instância $d_{j,n}$ de D_j, de um tal modo que um procedimento efetivo para resolver $d_{j,n}$ também forneça um procedimento para resolver $d_{i,m}$. Em outras palavras, se D_i se reduz a D_j, então se D_i é incomputável, D_j também o é. Note que a redução de um problema a outro também pode ser utilizada em provas de decidibilidade: se D_i se reduz a D_j e D_j é reconhecidamente computável, então D_i também o é.

Na ausência de $\mathbf{D}_{\text{incomp}}$, uma abordagem muito diferente foi requerida. Church, Post e Turing utilizaram, cada um deles, mais ou menos a mesma abordagem para esse propósito (GANDY, 1988). Primeiramente, precisa-se de um formalismo que capture a noção de computabilidade. Turing propôs o formalismo da máquina de Turing para esse fim. Um segundo passo é mostrar que existem problemas que não são computáveis dentro desse formalismo. Para conseguir isso, um processo uniforme **U** que seja capaz de computar todos os números computáveis precisa ser estabelecido em relação ao formalismo. Pode-se então utilizar (alguma forma de) a diagonalização em combinação com **U** para derivar uma contradição. A diagonalização foi introduzida por Cantor para mostrar que o conjunto dos números reais é "incontável" ou não enumerável. Uma variante do método foi utilizada também por Gödel na prova do seu **primeiro teorema da incompletude**.

O problema básico de Turing CIRC?, PRINT? e o *Entscheidungsproblem*

Lembre-se que na versão original da máquina de Turing, as máquinas computam números reais. Isso implica que uma máquina de Turing "bem comportada" de fato nunca deveria parar e imprimir uma sequência infinita de figuras. Tais máquinas foram identificadas por Turing como **não-circulares**. Todas as outra máquinas são denominadas máquinas **circulares**. Um número n que é o N.D. de uma máquina não-circular é denominado **satisfatório**.

Essa diferença básica é utilizada na prova de Turing da incomputabilidade de:

> **CIRC?** O problema de decidir, para todo número n, se n é satisfatório ou não.

A prova da incomputabilidade de **CIRC?** utiliza a construção de uma máquina hipotética e não-circular T_{decide} que computa a sequência diagonal do conjunto de todos os números computáveis computados pelas máquinas não-circulares. Portanto, sua construção depende da máquina de Turing universal e de uma máquina hipotética que seja capaz de decidir **CIRC?** para cada número

n dado a ela. Mostra-se que a máquina T_{decide} se torna uma máquina circular quando é fornecido a ela o seu próprio número de descrição, logo a suposição de uma máquina que seja capaz de resolver **CIRC?** deve ser falsa.

Com base na incomputabilidade de **CIRC?**, Turing então mostra que **PRINT?** também é incomputável. Mais particularmente, ele mostra que se **PRINT?** fosse computável, **CIRC?** também seria decidível, viz. ele reformula **PRINT?** de um tal modo que este se torna o problema de decidir, para qualquer máquina de Turing, se ela vai imprimir uma infinidade de símbolos ou não, o que resultaria em decidir **CIRC?**.

Finalmente, com base na incomputabilidade de **PRINT?**, Turing mostra que o *Entscheidungsproblem* não é decidível. Isso é feito mostrando o seguinte:

(1) como é possível construir, para cada máquina de Turing T, uma fórmula correspondente **T** na lógica de primeira ordem e
(2) se existe um método geral para determinar se **T** é demonstrável, então existe um método geral para demonstrar que T vai imprimir 0. Este é o problema **PRINT?** e, portanto, não pode ser decidível (desde que aceitemos a tese de Turing).

Assim, segue-se da incomputabilidade de **PRINT?** que o *Entscheindungsproblem* não é computável.

O problema da parada

Dado o foco de Turing em números reais computáveis, seu problema de decisão base é determinar se alguma máquina de Turing **não** vai parar ou se todas param e, portanto, não é exatamente o mesmo que o problema da parada mais conhecido:

> **HALT?** O problema de decidir, para toda máquina de Turing T, se T vai parar ou não.

De fato, o problema **PRINT?** de Turing é muito próximo a **HALT?** (*vide* DAVIS, 1958: Capítulo 5, Teorema 2.3)

Uma prova popular de **HALT?** é como a seguinte. Suponha que **HALT?** é computável. Então deveria ser possível construir uma máquina de Turing que decide, para cada máquina T_i e alguma entrada w para T_i, se T_i vai parar com w ou não. Vamos denominar tal máquina T_H. Mais particularmente, temos:

$$T_H(T_i, w) = \begin{cases} \text{HALT} & \text{se } T_i \text{ para com } w \\ \text{LOOP} & \text{se } T_i \text{ entra em loop com } w \end{cases}$$

Agora definimos uma segunda máquina T_D que depende da suposição de que a máquina T_H possa ser construída. Mais particularmente, temos:

$$T_D(T_i, \text{ N.D. de } T_i) = \begin{cases} \text{HALT} & \text{se } T_i \text{ não para com o seu próprio N.D.} \\ \text{LOOP} & \text{se } T_i \text{ para com o seu próprio N.D.} \end{cases}$$

Se nós agora considerarmos o caso em que T_i é T_D, chegamos a uma contradição: se T_D para, isto significa que T_D não para, e vice-versa. Uma variante popular mas bastante informal dessa demonstração no contexto da programação foi dada por Christopher Strachey (STRACHEY, 1965).

2.5. Variações da máquina de Turing

Como fica claro a partir das **Seções 1.1** e **1.2**, há uma variedade de definições de máquinas de Turing. Pode-se utilizar uma notação quíntupla ou quádrupla; pode-se ter diferentes tipos de símbolos ou apenas um; pode-se ter uma fita unidirecionalmente ou bidirecionalmente infinita; etc. Várias outras modificações menos óbvias foram consideradas e utilizadas no passado. Estas modificações podem ser de dois tipos: generalizações ou restrições. Elas não resultam em modelos “mais fortes” ou “mais fracos”. Ou seja, essas máquinas modificadas não computam mais nem menos do que as funções Turing computáveis. Este ponto contribui para a robustez da definição de máquina de Turing.

Máquinas binárias

Em sua nota curta de 1936, Post considera máquinas que marcam ou desmarcam um quadrado, o que significa que temos apenas dois símbolos, S_0 e S_1, mas ele não provou que essa formulação captura exatamente as funções Turing computáveis. Foi Shannon quem provou que, para qualquer máquina de Turing T com n símbolos, existe uma máquina de Turing com dois símbolos que simula T (SHANNON, 1956). Ele também mostrou que, para qualquer máquina de Turing com m estados, existe uma máquina de Turing com apenas dois estados que a simula.

Máquinas que não apagam

Máquinas que não apagam são máquinas que podem apenas sobrescrever S_0. Em Moore (1952), é mencionado que Shannon demonstrou que máquinas que não apagam podem computar o que qualquer máquina de Turing computa. Este resultado foi dado no contexto dos computadores digitais da década de 50 que utilizavam fitas perfuradas (e, assim, não se podia apagá-las). Entretanto, o resultado de Shannon não foi publicado. Foi Wang quem publicou o resultado (WANG, 1957).

Máquinas que não escrevem

Foi Minsky quem mostrou que, para toda máquina de Turing, existe uma máquina de Turing, com duas fitas e que não escreve, que a simula.

Fitas múltiplas

Ao invés de uma fita, pode-se considerar uma máquina de Turing com fitas múltiplas. Isso acaba sendo muito útil em vários contextos distintos. Por exemplo, Minsky utilizou uma máquina de Turing com fita dupla que não escreve para demonstrar que um certo problema de decisão definido por Post (o problema de decisão para as máquinas de Post) não é Turing computável (MINSKY, 1961). Hartmanis e Stearns, em seu artigo inovador para a teoria da complexidade computacional, demonstraram que qualquer máquina de Turing com n-fitas se reduz a uma máquina de Turing com uma única fita e, assim, qualquer coisa que

pode ser computada por uma máquina de Turing com n-fitas, ou máquinas de Turing de fitas múltiplas, pode também ser computado por uma máquina de Turing de fita única, e também conversamente (HARTMANIS; STEARNS, 1965). Eles utilizaram máquinas com fitas múltiplas porque se considerava que elas eram mais próximas dos computadores digitais.

Máquinas de Turing n-dimensionais

Outra variante a ser considerada são as máquinas de Turing nas quais as fitas são não unidimensionais mas n-dimensionais. Essa variante também se reduz à variante unidimensional.

Máquinas indeterminísticas

Uma reformulação aparentemente mais radical da noção de máquina de Turing é a de máquina de Turing indeterminística. Conforme explicado em **1.1**, uma condição fundamental das máquinas de Turing é a, assim denominada, condição de determinação, ou seja, a ideia de que, em um dado momento qualquer, o comportamento da máquina é completamente determinado pela configuração ou estado em que ela se encontra e o símbolo que ela está escaneando. Próximo a isso, Turing também menciona a ideia de máquinas de escolha, para as quais o próximo estado não é completamente determinado pelo par símbolo e estado. Ao invés disso, algum dispositivo externo faz uma escolha aleatória do que fazer a seguir. Máquinas de Turing indeterminísticas são um tipo de máquinas de escolha: para cada par símbolo e estado, a máquina indeterminística faz uma escolha arbitrária entre um número finito (possivelmente, zero) de estados. Assim, diferentemente da computação de uma máquina de Turing determinística, a computação de uma máquina de Turing indeterminística é uma árvore de caminhos possíveis de configuração. Um modo de visualizar a computação de uma máquina de Turing indeterminística é que a máquina desova uma cópia exata de si mesma e da sua fita para cada transição alternativa disponível, e cada máquina continua a computação. Se qualquer uma das máquinas termina de modo bem sucedido, então a computação inteira termina e herda a fita resultante daquela máquina. Note a expressão “bem sucedido” na sentença precedente. Nesta formulação, alguns estados são designados como

aceitáveis e quando a máquina termina em um deles, a computação é bem sucedida, de outro modo, a computação é mal sucedida e qualquer outra máquina continua sua busca por um resultado bem sucedido. A adição de máquinas de Turing indeterminísticas não altera a extensão da Turing-computabilidade. O indeterminsimo foi introduzido para autômatos finitos no artigo de Rabin e Scott (1959), no qual se mostrou também que adicionar o indeterminismo não resulta em autômatos mais poderosos. Máquinas de Turing indeterminísticas são um importante modelo no contexto da teoria da complexidade computacional (*vide* o verbete **Computational Complexity Theory**[35] da SEP).

Máquinas fracas e semi-fracas

Máquinas de Turing fracas são máquinas em que alguma palavra do alfabeto é repetida infinitamente à esquerda e à direita da entrada. Máquinas semi-fracas são máquinas em que alguma palavra é repetida infinitamente à esquerda ou à direita da entrada. Essas máquinas são generalizações do modelo padrão em que a fita inicial contém alguma palavra finita (possivelmente nula). Elas foram introduzidas para determinar máquinas universais menores. Watanabe foi o primeiro a definir uma máquina universal semi-fraca com seis estados e cinco símbolos (WATANABE, 1961). Recentemente, vários pesquisadores determinaram várias pequenas máquinas de Turing universais fracas e semi-fracas (por exemplo, WOODS; NEARY, 2007; COOK, 2004).

Para além dessas variantes do modelo de máquinas de Turing, há ainda variantes que resultam em modelos que capturam, em algum sentido bem-definido, mais do que as funções (Turing)-computáveis. Exemplos de tais modelos são as máquinas oraculares (TURING, 1939), as máquinas de Turing com tempo infinito (HAMKINS; LEWIS, 2008) e as máquinas de Turing que aceleram (COPELAND, 2002). Existem várias razões para introduzir estes modelos mais fortes. Alguns modelos são bem conhecidos na computabilidade e na teoria da recursão e são utilizados na teoria da recursão de ordem superior e na computabilidade relativa (máquinas oraculares); outras, como as máquinas que aceleram, foram introduzidas no contexto das supertarefas (*vide* o verbete

[35]N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/computational-complexity/. Acesso em: 20 jan. 2022.

Supertasks[36] da SEP) e da ideia de fornecer modelos físicos que "computem" funções que não sejam Turing-computáveis.

3. Questões Filosóficas em Relação às Máquinas de Turing

3.1. Computação Humana e Computação de Máquinas

Em seu contexto original, a identificação de Turing entre números computáveis e máquinas de Turing se destinava a demonstrar que o *Entscheidungsproblem* não é um problema computável e, assim, não é um problema de, assim denominado, "processo geral" (TURING, 1936-1937: 248). A suposição básica que deve ser feita para este resultado é que a nossa noção intuitiva de computabilidade pode ser formalmente definida como computabilidade de Turing e que, portanto, não existem problemas "computáveis" que não sejam Turing computáveis. Mas o que era a noção "intuitiva" de computabilidade em Turing e como podemos garantir que ela cobre todos os problemas computáveis, e, de modo mais geral, todos os tipos de computação? Esta é uma questão básica na filosofia da ciência computacional (*vide* o verbete **Philosophy of Computer Science**[37] da SEP).

Na época em que Turing estava escrevendo seu artigo, o computador moderno ainda não havia sido desenvolvido e, portanto, reformulações da tese de Turing que identificam a computabilidade de Turing com a computabilidade por computadores modernos são apenas interpretações e não afirmações historicamente corretas da tese de Turing. As máquinas de computação existentes na época em que Turing escreveu seu artigo, como o analisador diferencial ou as calculadoras de mesa, eram bastante restritas no que podiam computar e eram usadas em um contexto de práticas computacionais humanas (GRIER, 2007). Portanto, não é surpeendente que Turing não tenha tentado formalizar a computação de máquinas, mas sim a computação humana e, assim,

[36] N.T.: Disponível em: //plato.stanford.edu/archives/win2021/entries/spacetime-supertasks/. Acesso em: 20 jan. 2022.

[37] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/computer-science/. Acesso em: 20 jan. 2022

os problemas computáveis, no artigo de Turing, se tornam computáveis por meios humanos. Isso é muito explícito na Seção 9 de Turing (1936-1937), na qual ele mostra, analizando o processo da computação humana, que as máquinas de Turing são um modelo "natural" da computação (humana). A análise resulta em uma espécie de "computador" humano abstrato que preenche um conjunto de diferentes condições que estão enraizadas no reconhecimento de Turing de um conjunto de limitações humanas que restringem o que podemos computar (de nosso aparato sensorial, mas também de nosso aparato mental). Esse 'computador' calcula números (reais) em uma fita infinita unidimensional dividida em quadrados [Nota: Turing assumiu que a redução do caráter bidimensional do papel em que um matemático humano geralmente trabalha "não é essencial para a computação" (TURING, 1936-7: 249)]. A análise impõe as seguintes restrições (GANDY, 1988; SIEG, 1994):

Condição de Determinação D "Em qualquer momento, o comportamento do computador é determinado pelos símbolos que ele está observando e pelo seu 'estado mental' naquele momento." (TURING, 1936-1937: 250)

Comdição de Limite B1 "existe um limite B para o número de símbolos ou quadrados que o computador pode observar a cada momento. Se ele deseja observar mais, ele deve então realizar observações sucessivas." (TURING, 1936-1937: 250)

Comdição de Limite B2 "o número de estados mentais que devem ser levado s em conta é finito" (TURING, 1936-1937: 250)

Comdição de Localidade L1 "podemos [...] assumir que os quadrados cujos símbolos são alterados são sempre quadrados 'observados'." (TURING, 1936-1937: 250)

Comdição de Localidade L2 "cada um dos novos quadrados observados está dentro de L quadrados de um quadrado imediatamente observado anteriormente." (TURING, 1936-1937: 250)

É este, assim denominado, "apelo direto à intuição" da análise de Turing e os modelos resultantes que explicam por que a máquina de Turing é hoje considerada por muitos como o melhor modelo padrão da computabilidade (para uma forte defesa deste ponto de vista, *vide* SOARE, 1996). De fato, do conjunto

de condições acima pode-se facilmente derivar as máquinas de Turing. Isto é feito basicamente analisando as condições restritivas em termos de "'operações simples', as quais são tão elementares que não é fácil imaginá-las ainda mais divididas" (TURING, 1936-1937: 250).

Observe que, enquanto a análise de Turing é focada na computação humana, a aplicação dessa indentificação entre computação (humana) e computação por máquina de Turing ao *Entscheindungsproblem* sugere que ele **não** considerou a possibilidade de um modelo de computação que de algum modo vá "além" da computação humana e seja capaz de fornecer um procedimento geral e efetivo que resolva o *Entscheindungsproblem*. Se este não fosse o caso, ele não teria considerado o *Entscheindungdproblem* como incomputável.

O foco na computação humana pela análise de computação de Turing levou pesquisadores a estender a análise de Turing para a computação por dispositivos físicos. Isso resulta na (versões da) tese física de Church-Turing. Robin Gandy se concentrou em estender a análise de Turing para dispositivos mecânicos discretos (observe que ele não considerou máquinas analógicas). Mais particularmente, como Turing, Gandy parte de um conjunto básico de restrições de computação por dispositivos mecânicos discretos e, com base nisso, desenvolve um novo modelo que provou ser redutível ao modelo de máquina de Turing. Este trabalho é continuado por Wilfried Sieg, que propôs o *framework* de Sistemas Dinâmicos Computáveis (SIEG, 2008). Outros consideraram a possibilidade de modelos "razoáveis" da física "computarem" algo que não seja Turing computável. Veja, por exemplo, AARONSON; BAVARIAN; GUELTIRNE, 2016, no qual é mostrado que, se existissem curvas fechadas de tipo temporal, o problema da parada se tornaria solucionável com recursos finitos. Outros propuseram modelos alternativos para a computação que são inspirados no modelo da máquina de Turing, mas capturam aspectos específicos das práticas atuais de computação para as quais o modelo da máquina de Turing é considerado menos adequado. Um exemplo aqui são as máquinas de Turing persistentes destinadas a capturar processos interativos. Observe, no entanto, que esses resultados não mostram que existem problemas "computáveis" que não sejam Turing computáveis. Essas e outras propostas relacionadas têm sido consideradas por alguns autores como modelos razoáveis de computação que, de

alguma forma, computam mais do que as máquinas de Turing. É o último tipo de declaração que se afiliou à pesquisa sobre a chamada hipercomputação, resultando no início dos anos 2000 em um debate bastante acirrado na comunidade da ciência da computação (para várias posições, *vide*, por exemplo, TEUSCHER, 2004).

3.2. Tese, Definição, Axiomas ou Teorema

Como é claro, estritamente falando, a tese de Turing não é demonstrável, pois, em sua forma original, é uma afirmação sobre a relação entre um conceito formal e um conceito vago ou intuitivo. Consequentemente, muitos a consideram como uma tese ou uma definição. A tese seria refutada se alguém fosse capaz de fornecer um procedimento eficaz intuitivamente aceitável para uma tarefa que não é Turing-computável. Até agora, nenhum contra-exemplo foi encontrado. Outras noções de computabilidade definidas independentemente com base em fundamentos alternativos, como funções recursivas e máquinas de ábaco, também se mostraram equivalentes à computabilidade de Turing. Essas equivalências entre formulações bastante diferentes indicam que há uma noção natural e robusta de computabilidade subjacente ao nosso entendimento. Dada essa aparente robustez de nossa noção de computabilidade, alguns propuseram evitar completamente a noção de tese e, em vez disso, propor um conjunto de axiomas utilizados para aguçar a noção informal. Existem várias abordagens, mais notavelmente, uma abordagem de axiomatização estrutural em que a própria computabilidade é axiomatizada (SIEG, 2008) e uma em que uma é dada uma axiomatização a partir da qual a tese de Church-Turing pode ser derivada (DERSHOWITZ; GUREVICH, 2008).

4. Modelos Alternativos Históricos da Computabilidade

Além da máquina de Turing, vários outros modelos foram introduzidos independentemente de Turing no contexto da pesquisa sobre os fundamentos da matemática que resultaram em teses que são logicamente equivalentes à tese de Turing. Para cada um desses modelos foi comprovado que eles capturam as

funções computáveis de Turing. Observe que o desenvolvimento do computador moderno estimulou o desenvolvimento de outros modelos, como máquinas de registradores ou algoritmos de Markov. Mais recentemente, abordagens computacionais em disciplinas como biologia ou física resultaram em modelos bioinspirados e inspirados na física, como redes de Petri ou máquinas de Turing quânticas. Uma discussão de tais modelos, no entanto, está além do escopo deste verbete.

4.1. Funções Recursivas Gerais

A formulação original de funções recursivas gerais pode ser encontrada em Gödel (1934), que se baseou em uma sugestão de Herbrand. Em Kleene (1936) foi dada uma definição mais simples e em Kleene (1943) foi introduzida a forma padrão que usa a chamada minimização ou operador-μ. Para mais informações, *vide* o verbete **Recursive Functions**[38] da SEP.

Church utilizou a noção de função recursiva geral para enunciar a sua tese:

A Tese de Church Toda função efetivamente calculável é recursiva geral.

No contexto das funções recursivas, usa-se a noção de solucionabilidade recursiva e insolucionabilidade ao invés de computabilidade de Turing e incomputabilidade. Esta terminologia é devida a Post (1944).

4.2. Definibilidade-λ

O Cálculo-λ de Church tem origem nos artigos (CHURCH, 1932, 1933) e que se destinavam a uma fundamentação lógica da matemática. À época, a convicção de Church era a de que esta abordagem formal distinta poderia evitar a incompletude de Gödel (SIEG, 1997: 177). Entretanto, a inconsistência do sistema lógico de Church foi demonstrada por seus dois alunos de doutorado,

[38] N.T.: Disponível em: https://plato.stanford.edu/archives/win2021/entries/recursive-functions/. Acesso em: 20 jan. 2022

Stephen C. Kleene e Barkley Rosser e, assim, eles começaram a focar em uma subparte desta lógica, que era basicamente o cálculo-λ. Church, Kleene e Rosser começar a λ-definir todas as funções calculáveis que eles pudessem pensar e logo em seguida Church propôs definir a computabilidade efetiva em termos de definibilidade-λ. Entretanto, foi somente depois de Church, Kleene e Rosser terem estabelecido que a recursividade geral e a definibilidade-λ são equivalentes é que Church anunciou a sua tese publicamente, e em termos das funções recursivas gerais ao invés da definibilidade-λ (DAVIS, 1982; SIEG, 1997).

No cálculo-λ há somente dois tipos de símbolos. Os três símbolos primitivos λ, (,), também denominados símbolos impróprios, e uma lista infinita de variáveis. Há três regras para definir as fórmulas bem formadas do cálculo-λ, denominadas fórmulas-λ.

1. Primeiramente, as variáveis são, elas mesmas, fórmulas-λ.
2. Se **P** é uma fórmula-λ contendo x como uma variável livre, então $\lambda x[\mathbf{P}]$ é também uma fórmula-λ. O operador-λ é utilizado para ligar variáveis, assim ele converte uma expressão contendo variáveis livres em uma expressão que denota uma função.
3. Se **M** e **N** são fórmulas-λ, então $\{\mathbf{M}\}(\mathbf{N})$ também o é, em que $\{\mathbf{M}\}(\mathbf{N})$ deve ser compreendida como a aplicação da função **M** a **N**.

As fórmulas-λ, ou fórmulas bem formadas do cálculo-λ, são todas e somente aquelas fórmulas de resultam de aplicações (repetidas) destas três regras.

Existem três operações ou regras de conversão. Vamos definir $S^x_{\mathbf{N}}\mathbf{M}|$ como a fórmula que resulta da substituição de **N** por x em **M**.

1. **Redução**. Substituir qualquer parte $\{\lambda x[\mathbf{M}]\}(\mathbf{N})$ de uma fórmula por $S^x_{\mathbf{N}}\mathbf{M}|$, desde que as variáveis ligadas de **M** sejam distintas, tanto de x quanto das variáveis livres de **N**. Por exemplo, $\{\lambda x[x^2]\}(2)$ reduz para 2^2.
2. **Expansão**. Substituir qualquer parte $S^x_{\mathbf{N}}\mathbf{M}|$ de uma fórmula por $\{\lambda x[\mathbf{M}]\}(\mathbf{N})$, desde que $((\lambda x\mathbf{M})\mathbf{N})$ seja bem formada e as variáveis ligadas de **M** sejam distintas tanto de x quanto das variáveis livres de **N**. Por exemplo, 2^2 pode ser expandida para $\{\lambda x[x^2]\}(2)$.

3. **Mudança de variáveis ligadas**. Substituir qualquer parte **M** de uma fórmula por $S^x_{\mathbf{y}}\mathbf{M}|$, desde que x não seja uma variável livre de **M** e y não ocorra em **M**. Por exemplo, a mudança de $\{\lambda x[x^2]\}$ para $\{\lambda y[y^2]\}$.

Church introduz as seguintes abreviações para definir os números naturais no cálculo-λ.

$$\begin{aligned} 1 &\rightarrow \lambda yx.yx, \\ 2 &\rightarrow \lambda yx.y(yx), \\ 3 &\rightarrow \lambda yx.y(y(yx)), \\ &\ldots \end{aligned}$$

Utilizando essa definição, é possivel λ-**definir** funções sobre os inteiros positivos. A função F de um inteiro positivo é λ-definível se podemos encontrar uma fórmula-λ **F**, tal que, se $F(m) = n$ e **m** e **n** são fórmulas-λ para os inteiros m e n, então a fórmula-λ $\{\mathbf{F}\}(\mathbf{m})$ pode ser **convertida** em **n** aplicando-se as regras de conversão do cálculo-λ. Assim, por exemplo, a função sucessor S, primeiramente introduzida por Church, pode ser λ-definida da seguinte maneira:

$$S \rightarrow \lambda abc.b(abc)$$

Para dar um exemplo, aplicando S à fórmula de 2, obtemos:

$$\begin{aligned} &\big(\lambda abc.b(abc)\big)\big(\lambda yx.y(yx)\big) \\ &\rightarrow \lambda bc.b\big(\big(\lambda yx.y(yx)\big)bc\big) \\ &\rightarrow \lambda bc.b\big(\big(\lambda x.b(bx)\big)c\big) \\ &\rightarrow \lambda bc.b(b(bc)) \end{aligned}$$

Hoje, o cálculo-λ é considerado um modelo básico na teoria da programação.

4.3. Sistemas de Produção de Post

Em meados de 1920-1921, Emil Post desenvolveu tipos diferentes, mas relacionados, de sistemas de produção, a fim de desenvolver uma forma sintática que lhe permitisse enfrentar o problema de decisão para a lógica de primeira ordem. Uma dessas formas são os sistemas canônicos de Post C, que mais tarde ficaram conhecidos como sistemas de produção de Post.

Um sistema canônico consiste de um alfabeto finito Σ, um conjunto finito de palavras iniciais $W_{0,0}, W_{0,1}, ..., W_{0,n}$ e um conjunto finito de regras de produção da seguinte forma:

$$g_{11}P_{i_1^1}g_{12}P_{i_2^1}\dots g_{1m_1}P_{i_{m_1}^1}g_{1(m_1+1)}$$
$$g_{21}P_{i_1^2}g_{22}P_{i_2^2}\dots g_{2m_2}P_{i_{m_2}^2}g_{2(m_2+1)}$$
$$\dots\dots\dots\dots\dots\dots\dots\dots$$
$$g_{k1}P_{i_1^k}g_{k2}P_{i_2^k}\dots g_{km_k}P_{i_{m_k}^k}g_{k(m_k+1)}$$
produz
$$g_1P_{i_1}g_2P_{i_2}\dots g_mP_{i_m}g_{(m+1)}$$

Os símbolos g são um tipo de metasímbolos: eles correspondem às sequências de letras atuais nas produções atuais. Os símbolos P são as variáveis operacionais e, assim, podem representar qualquer sequência de letras em uma produção. Assim, por exemplo, considere um sistema de produção sobre o alfabeto $\Sigma = \{a, b\}$ com a palavra inicial:

$$W_0 = ababaaabbaabbaabbaba$$

e a seguinte regra de produção:

$$P_{1,1}bbP_{1,2}$$
produz
$$P_{1,3}aaP_{1,4}$$

Então, começando com W_0, há três caminhos possíveis para aplicar a regra de produção e, em cada aplicação, as variáveis $P_{1,i}$ terão valores distintos, mas os valores de g estarão fixados. Qualquer conjunto de sequências finitas de

palavras que podem ser produzidas por um sistema canônico é denominado um **conjunto canônico**.

Uma classe especial de formas canônicas definidas por Post são os sistemas normais. Um sistema normal N cosiste de um alfabeto finito Σ, uma palavra inicial $W_0 \in \Sigma^*$ e um conjunto finito de regras de produção, cada uma da seguinte forma:

$$g_i P$$
produz
$$P g_i'$$

Qualquer conjunto finito de sequências de palavras que pode ser produzido por um sistema normal é denominado um **conjunto normal**. Post foi capaz de mostrar que, para conjunto canônico C sobre algum alfabeto Σ, existe um conjunto normal N sobre um alfabeto Δ com $\Sigma \subseteq \Delta$, tal que, $C = N \cap \Sigma^*$. Era uma convicção de Post que (1) qualquer conjunto de sequências finitas que possa ser gerado por meios finitos pode ser gerado por sistemas canônicos e (2) da prova de que para todo conjunto canônico existe um conjunto normal que o contém, o que resultava na tese de Post I:

> **Tese de Post I** (DAVIS, 1982) Todo conjunto de sequências finitas de letras que pode ser gerado por um processo finito pode também ser gerado por sistemas normais. Mais particularmente, qualquer conjunto de palavras em um alfabeto Σ que possa ser gerado por um processo finito é da forma $N \cap \Sigma^*$, em que N é um conjunto normal.

Post percebeu que "[para a validade de sua tese em plena generalidade] uma análise completa teria que ser feita de todas as maneiras possíveis pelas quais a mente humana poderia estabelecer processos finitos para gerar sequências" (POST, 1965: 408) e é bastante provável que a formulação 1 dada em Post (1936) e que é quase idêntica às máquinas de Turing seja o resultado de tal análise.

Os sistemas de produção de Post se tornaram um importante dispositivo formal na ciência da computação e, mais especificamente, na teoria da linguagem

formal (DAVIS, 1989; PULLUM, 2011).

4.4. A Formulação 1

Em 1936, Post publicou uma nota curta a partir da qual se pode derivar a segunda tese de Post (DE MOL, 2013):

> **Tese de Post II** A solucionabilidade de um problema no sentido intuitivo coincide com a solucionabilidade pela formulação 1.

A formulação 1 é muito semelhante às máquinas de Turing, mas o "programa" é dado como uma lista de instruções que um trabalhador humano precisa seguir. Em vez de uma fita unidirecionalmente infinita, a "máquina" de Post consiste em um espaço de símbolos bidirecionalmente infinito e dividido em caixas. A ideia é que um trabalhador esteja trabalhando neste espaço simbólico, sendo capaz de um conjunto de cinco atos primitivos (O_1 marque uma caixa, O_2 desmarque uma caixa, O_3 desloque-se uma caixa para esquerda, O_4 desloque-se uma caixa para a direita, O_5 determine se a caixa em que ele se encontra está marcada ou desmarcada), seguindo um conjunto finito de direções $d_1, ..., d_n$, em que cada direção d_i tem sempre uma das seguintes formas:

A. Execute uma das operações $(O_1 - O_4)$ e vá para a direção d_j.
B. Execute a operação O_5 e, conforme a caixa na qual se encontra o trabalhador esteja marcada ou desmarcada, siga a direção $d_{j'}$ ou $d_{j''}$.
C. Pare.

Post também definiu uma terminologia específica para sua formulação 1 a fim de definir a solucionabilidade de um problema em termos da formulação 1. Essas noções são aplicabilidade, processo-1-finito, solução-1 e dado-1. Grosso modo, essas noções asseguram que um problema de decisão é solucionável com a formulação 1 sob a condição de que a solução dada no formalismo sempre termine com uma solução correta.

5. Impacto das Máquinas de Turing na Ciência da Computação

Turing é hoje uma das mais célebres figuras da ciência da computação. Muitos o consideram o pai da ciência da computação e o fato de que o principal prêmio na comunidade da ciência da computação é chamado de "Prêmio Turing" é uma clara indicação disso (DAYLIGHT, 2015). Isso foi reforçado pelas celebrações do centenário Turing de 2012, que foram em grande parte coordenadas pro S. Barry Cooper. Isso resultou não só em um número enorme de eventos científicos em torno de Turing, mas também em numerosas iniciativas que levaram à ideia de Turing como pai da ciência da computação até o grande público (BULLYNCK, DAYLIGHT; DE MOL, 2015). Entre as contribuições de Turing que são hoje consideradas pioneiras, o artigo de 1936 sobre as máquinas de Turing sobressai-se como o que teve maior impacto na ciência da computação. No entanto, a investigação histórica recente mostra também que se deve tratar o impacto das máquinas de Turing com muito cuidado e que se deve ser cauteloso para não distorcer o passado pelo presente.

5.1. Impacto na Ciência da Computação

Hoje, a máquina de Turing e a sua teoria são parte dos fundamentos teóricos da ciência da computação. Trata-se de uma referência padrão na pesquisa sobre questões fundamentais como:

- O que é um algorítmo?
- O que é computação?
- O que é computação física?
- O que é computação eficiente?
- etc.

Trata-se também de um dos principais modelos para pesquisa em muitas subdisciplinas na ciência da computação teórica tais como: modelos de computabilidade variante e minimal, computabilidade de ordens superiores, teoria da complexidade computacional, teoria algorítmica da informação, etc. Essa importância d modelo da máquina de Turing para a ciência da computação teórica

tem ao menos duas raízes.

Primeiramente, há a continuação do trabalho em lógica matemática dos anos 1920 e 1930 por pessoas como Martin Davis - que foi aluno de Post e Church - e Kleene. Nessa tradição, a obra de Turing era obviamente bem conhecida e a máquina de Turing foi considerada o melhor modelo de computabilidade já dado. Davis e Kleene publicaram livros nos anos de 1950 sobre tópicos (KLEENE, 1952; DAVIS, 1958) que rapidamente se tornaram referências standard não só para a teoria da computabilidade inicial, mas também para mais reflexões teóricas sobre computação nos finais dos anos de 1950 e 1960.

Em segundo lugar, podemos ver que havia nos anos 1950 a necessidade de modelos teóricos para refletir sobre as novas máquinas computacionais, as suas capacidades e limitações e isso de maneira sistemática. É nesse contexto que o trabalho teórico já realizado foi recebido. Um desenvolvimento importante é a teoria dos automata em que se pode situar, entre outros, o desenvolvimento de outros modelos de máquina como o modelo de máquina registradora ou o modelo de máquina de Wang B que são, basicamente, enraizados nas máquinas de Turing e de Post; há projetos de máquina minimais discutidos na **Seção 5.2**; e há o uso das máquinas de Turing no contexto do que se tornaria as origens da teoria da linguagem formal, especificamente o estudo de classes diferentes de máquinas com relação a diferentes "linguagens" que elas podem reconhecer e assim também as suas limitações e forças. Foram esses desenvolvimentos mais teóricos que contribuíram para o estabelecimento da teoria da complexidade computacional nos anos de 1960. Obviamente, ao lado das máquinas de Turing, outros modelos também desempenharam e desempenham um papel importante nesses desenvolvimentos. Mesmo assim, na ciência da computação teórica é a máquina de Turing que permanece o modelo, ainda hoje. De fato, quando em 1965 um dos artigos fundacionais da teoria da complexidade computacional (HARTMANIS; STEARNS, 1965) foi publicado, foi a máquina de Turing de fitas múltiplas introduzida como o modelo padrão para o computador.

5.2. As Máquinas de Turing e Computador Moderno

Em muitas narrativas, Turing foi identificado não apenas como o pai da ciência da computação, mas também como o pai do computador moderno. A história clássica para isso é mais ou menos a seguinte: a planta do computador moderno pode ser encontrado no projeto EDVAC de von Neumann e hoje os computadores clássicos são geralmente descritos como tendo a chamada arquitetura von Neumann. Uma ideia fundamental do projeto EDVAC é a chamada ideia de programa armazenado. Grosso modo, isso significa o armazenamento de instruções e dados na mesma memória, permitindo a manipulação de programas como dados. Há boas razões para supor que von Neumann conhecia os principais resultados do artigo de Turing (DAVIS, 1988). Assim, pode-se argumentar que o conceito de programa armazenado se origina na noção de Turing da máquina de Turing universal e, destacando isso como a característica definidora do computador moderno, alguns podem alegar que Turing é o pai do computador moderno. Outro argumento relacionado é que Turing foi o primeiro a "capturar" a ideia de uma máquina de uso geral por meio de sua noção de máquina universal e que, nesse sentido, ele também "inventou" o computador moderno (COPELAND; PROUDFOOT, 2011). Este argumento é então reforçado pelo fato de que Turing também esteve envolvido com a construção de uma importante classe de dispositivos de computação (a *Bombe*) usada para decifrar o código alemão Enigma e, posteriormente, propôs o projeto do ACE (*Automatic Computing Engine*) que foi explicitamente identificado como uma espécie de realização física da máquina universal pelo próprio Turing:

> Alguns anos atrás, eu estava pesquisando naquilo que pode ser agora descrito como uma investigação das possibilidades teóricas e limitações das máquinas de computação digitais. [...] Máquinas como a ACE podem ser consideradas versões práticas desse mesmo tipo de máquina. (TURING, 1947)

Note, porém, que Turing já conhecia os projetos ENIAC e EDVAC e

propôs o ACE como uma espécie de melhoria nesse projeto (entre outros, tinha uma arquitetura de *hardware* mais simples).

Essas afirmações sobre Turing como o inventor e/ou pai do computador foram examinadas por alguns historiadores da computação (DAYLIGHT, 2014; HAIGH, 2013; HAIGH, 2014; MOUNIER-KUHN, 2012), principalmente na sequência do centenário de Turing e isso de vários perspectivas. Com base nessa pesquisa, fica claro que as alegações sobre Turing ser o inventor do computador moderno dão uma imagem distorcida e tendenciosa do desenvolvimento do computador moderno. Na melhor das hipóteses, ele é um dos muitos que contribuíram para um dos vários desenvolvimentos históricos (científicos, políticos, tecnológicos, sociais e industriais) que resultaram, em última análise, no (nosso conceito de) computador moderno. De fato, os "primeiros" computadores são o resultado de um grande número de inovações e, portanto, estão enraizados no trabalho de não apenas uma, mas várias pessoas com diversas origens e pontos de vista.

Na década de 1950, a máquina de Turing (universal) começa a se tornar um modelo aceito em relação aos computadores reais e é usada como ferramenta para refletir sobre os limites e o potencial dos computadores de uso geral por engenheiros, matemáticos e lógicos. Mais particularmente, no que diz respeito aos projetos de máquinas, foi a percepção de que apenas um pequeno número de operações era necessário para construir uma máquina de uso geral que inspirou as reflexões dos anos 1950 sobre arquiteturas de máquinas mínimas. Frankel, que (parcialmente) construiu o MINAC declarou o seguinte:

> Um resultado notável da investigação de Turing é que ele foi capaz de descrever um único computador que é capaz de computar qualquer número computável. Ele chamou esta máquina de computador universal. Ela é portanto o "melhor computador possível" que já foi mencionado.
>
> [...] Este resultado surpreendente mostra que, ao examinar a questão de quais problemas são, em princípio, solucionáveis por máquinas de computação, nós não precisamos considerar uma

> série infinita de computadores de complexidade cada vez maior, mas podemos apenas pensar em uma única máquina.
>
> Ainda mais surpreendente do que a possibilidade teórica de um tal "melhor computador possível" é o fato de que ele não precisa ser muito complexo. A descrição dada por Turing de um computador universal não é única. Muitos computadores, alguns de complexidade muito modesta, satisfazem os requisitos para um computador universal. (FRANKEL, 1956: 635)

O resultado foi uma série de máquinas experimentais como o MINAC, TX-0 (Lincoln Lab) ou a máquina ZERO (van der Poel), que por sua vez se tornaram predecessoras de várias máquinas comerciais. Vale ressaltar que também o design da máquina ACE de Turing se encaixa nessa filosofia. Também foi comercializado como a máquina BENDIX G15 (DE MOL; NULLYNCK; DAYLIGHT, 2018).

É claro que, ao minimizar as instruções da máquina, a codificação ou programação tornou-se uma tarefa muito mais complicada. Para colocar nas palavras de Turing, que percebeu claramente essa troca entre código e instruções (*hard-wired*) ao projetar o ACE: "Muitas vezes simplificamos o circuito às custas do código" (TURING, 1947). E, de fato, pode-se ver que, com esses *designs* iniciais mínimos, muito esforço é dedicado ao desenvolvimento de estratégias de codificação mais eficientes. Aqui também se pode situar uma raiz histórica da conexão entre a máquina de Turing universal e o importante princípio da intercambialidade entre *hardware* e programa.

Hoje, a máquina de Turing universal ainda é considerada por muitos como o principal modelo teórico do computador moderno especialmente em relação à chamada arquitetura de von Neumann. Obviamente, outros modelos foram introduzidos para outras arquiteturas, como o modelo paralelo síncrono em massa para máquinas paralelas ou a máquina de Turing persistente para modelagem de problemas interativos.

5.3. Teorias da Programação

A ideia de que qualquer máquina de uso geral pode, em princípio, ser modelada como uma máquina de Turing universal também se tornou um princípio importante no contexto da programação automática na década de 1950 (DAYLIGHT, 2015). No contexto do projeto de máquina, foi a minimização das instruções de máquina que foi a consequência mais importante desse ponto de vista. No contexto da programação, então, tratava-se da ideia de que se pode construir uma máquina que seja capaz de 'imitar' o comportamento de qualquer outra máquina e, portanto, em última análise, a intercambialidade entre o *hardware* da máquina e as implementações de linguagem. Isso foi introduzido de várias formas na década de 1950 por pessoas como John W. Carr III e Saul Gorn - que também estiveram ativamente envolvidos na formação da *Association for Computing Machinery* (ACM) - como a ideia teórica unificadora para a programação automática que de fato é sobre a "tradução" (automática) de ordem superior para o nível inferior e, em última análise, o código de máquina. Assim, também no contexto da programação, a máquina de Turing universal começa a assumir seu papel fundacional na década de 1950 (DAYLIGHT, 2015).

Enquanto a máquina de Turing foi e ainda é um modelo teórico fundamental delimitando o que é possível e não é possível no nível geral, ela não teve um impacto real na sintaxe e semântica das linguagens de programação. Nesse contexto, foram os sistemas de cálculo-λ e de produção de Post que tiveram efeito (embora também aqui se deva ter cuidado ao exagerar a influência de um modelo formal em uma prática de programação). Na verdade, as máquinas de Turing eram muitas vezes consideradas como modelos de máquina e não como um modelo para a programação:

> As máquinas de Turing não são conceitualmente distintas dos computadores automáticos no seu uso geral, mas elas são muito pobres no seu controle de estrutura. [...] Claro, muito da teoria da computabilidade lida com questões que não estão relacionadas com os modos particulares pelos quais as computações são representadas. É

> suficiente que as funções computáveis sejam representadas de algum modo por expressões simbólicas, por exemplo, números, e que as funções computáveis em termos das funções dadas sejam de algum modo representadas por expressões computáveis em termos das expressões que representam as funções originais. Entretanto, uma teoria da computação prática deve ser aplicável a algoritmos particulares. (MCCARTHY, 1963: 37)

Assim, vê-se que o papel da máquina de Turing para a ciência da computação deve situar-se mais no nível teórico: a máquina universal é hoje por muitos ainda considerada como o modelo para o computador moderno, enquanto sua capacidade de imitar máquinas através de sua manipulação de programas-como-dados é um dos princípios básicos da computação moderna. Além disso, sua robustez e naturalidade como modelo de computabilidade a tornaram o principal modelo a ser desafiado se alguém estiver atacando versões da chamada tese (física) de Church-Turing.

Bibliografia

BARWISE, J.; ETCHEMENDY, J. **Turing's World**, Stanford, CA: CSLI Publications, 1993.

BOOLOS, G.; JEFFREY, R. **Computability and Logic**, Cambridge: Cambridge University Press; 5a edição, 2007 [1974]

BROMLEY, A. "Stored Program Concept. The Origin of the Stored Program Concept", **Technical Report 274**, Basser Department of Computer Science, Nov. 1985 [disponível em: https://web.archive.org/web/20171002030127/http://sydney.edu.au/engineering/it/research/tr/tr274.pdf].

BULLYNCK, M. , DAYLIGHT, E. ; DE MOL, L. "Why Did Computer Science Make a Hero Out of Turing?", **Communications of the ACM**, 58(3): 37–39. 2015.

CHURCH, A. "A Set of Postulates for the Foundation of Logic", **Annals of**

Mathematics, 33(2): 346–366, 1932.

CHURCH, A. "A Set of Postulates for the Foundation of Logic (Second Paper)", **Annals of Mathematics**, 34(4): 839–864, 1933.

CHURCH, A. "An Unsolvable Problem of Elementary Number Theory", **American Journal of Mathematics**, 58(2): 345–363, 1936a.

CHURCH, A. "A Note on the Entscheidungsproblem", **Journal of Symbolic Logic**, 1(1): 40–41, 1936b.

CHURCH, A. "Review of: On Computable Numbers with An Application to the Entscheidungsproblem by A.M. Turing", **Journal of Symbolic Logic**, 2(1): 42–43, 1937.

COOK, M. "Universality in Elementary Cellular Automata", **Complex Systems**, 15(1): 1–40, 2004.

COOPER, S., LEEUWEN, J. V. **Alan Turing: His Work and Impact**, Amsterdam: Elsevier, 2013.

COPELAND, B. "Accelerating Turing Machines", **Minds and Machines**, 12(2): 281–301, 2002.

COPELAND, B., PROUDFOOT, J. , PROUDFOOT, D. "Alan Turing: Father of the Modern Computer", **The Rutherford Journal,** 4: 1, 2011 [disponível em: http://www.rutherfordjournal.org/article040101.html].

DAVIS, M. **Computability and Unsolvability**, Nova York: McGraw-Hill. Reimpressão Dover, 1982[1958].

DAVIS, M. **The Undecidable. Basic papers on undecidable propositions, unsolvable problems and computable functions**, Nova York: Raven Press. 1965.

DAVIS, M. "What is a Computation?", in: STEEN (ed.), **Mathematics Today: Twelve Informal Essays,** Nova York: Springer, p. 241–267, 1978.

DAVIS, M. "Why Gödel Didn't Have Church's Thesis", **Information and Control**, 54:(1–2): 3–24., 1982.

DAVIS, M. "Mathematical Logic and the Origin of the Modern Computer", in: HERKEN 1988: 149–174, 1988.

DAVIS, M. "Emil Post's Contribution to Computer Science", **Proceedings of the Fourth Annual Symposium on Logic in Computer Science**, IEEE Computer Society Press, p. 134–137, 1989.

DAVIS, M.; SIEG, W. “Conceptual Confluence in 1936: Post and Turing”, in: SOMMARUGA; STRAHM (ed.) **Turing’s Revolution: The Impact of His Ideas about Computability**, Cham: Springer, 2015.

DAYLIGHT, E. “A Turing Tale”, **Communications of the ACM**, 57(10): 36–38, 2014.

DAYLIGHT, E. “Towards a Historical Notion of ‘Turing—The Father of Computer Science’”, **History and Philosophy of Logic**, . 36(3): 205–228, 2015.

DE MOL, L. “Generating, Solving and the Mathematics of Homo Sapiens. Emil Post’s Views On computation”, in: ZENIL (ed.) **A Computable Universe. Understanding Computation & Exploring Nature As Computation**, Hackensack, NJ: World Scientific, p. 45–62, 2013 [disponível em: https://hal.univ-lille.fr/hal-01396500/document].

DE MOL, L., BULLYNCK, M.; DAYLIGHT, E. “Less is More in the Fifties: Encounters between Logical Minimalism and Computer Design during the 1950s”, **IEEE Annals of the History of Computing**, 40(1): 19–45. 2018.

DEUTSCH, D. “Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer”, **Proceedings of the Royal Society A**, 400(1818): 97–117, 1985.

DERSHOWITZ, N.; GUREVICH, Y. “ A Natural Axiomatization of Computability and Proof of Church’s Thesis”, **Bulletin of Symbolic Logic**, 14(3): 299–350, 2008.

FRANKEL, S. “Useful Applications of a Magnetic-Drum Computer”, **Electrical Engineering**, 75(7): 634–39, 1956.

GANDY, R. “Church’s Thesis and Principles for Mechanism”, in: BARWISE, KEISLER; KUNEN (eds.) **The Kleene Symposium: Proceedings of the Symposium Held June 18–24, 1978 at Madison, Wisconsin, U.S.A.**, (Studies in Logic and the Foundations of Mathematics, 101), Amsterdam: North-Holland, 1980, p. 123–148.

GANDY, R. “The Confluence of Ideas in 1936”, in: HERKEN, *op. cit.* p. 55–111, 1988.

GÖDEL, K. “Die Vollständigkeit der Axiome des logischen Funktionenkalkül”, **Monatshefte für Mathematik und Physik**, 37: 349–360, 1929.

GÖDEL, K. “On Undecidable Propositions of Formal Mathematical Systems,

mimeographed lecture notes by S. C. Kleene and J. B. Rosser", Institute for Advanced Study, Princeton, NJ, 1934; corrigidas e ampliadas em DAVIS, M. *op.cit.* 41–74, 1965.

GRIER, D. **When Computers Were Human**, Princeton, NJ: Princeton University Press, 2007.

HAIGH, T. "'Stored Program Concept' Considered Harmful: History and Historiography", in: BONIZZONI; BRATTKA; LÖWE, **The Nature of Computation. Logic, Algorithms, Applications: 9th Conference on Computability in Europe, CiE 2013, Milan, Italy, July 1–5, 2013 Proceedings** (Lecture Notes in Computer Science, 7921), Berlin: Springer, p. 241–251, 2013.

HAIGH, T. "Actually, Turing Did Not Invent the Computer", **Communications of the ACM**, 57(1): 36–41., 2014.

HAMKINS, J.; LEWIS, A. "Infinite Time Turing Machines", **Journal of Symbolic Logic**, 65(2): 567–604, 2000

HARTMANIS, J.; STEARNS, R. E. "On the Computational Complexity of Algorithms" Transactions of the American Mathematical Society, 117: 285–306, 1965.

HERKEN, R. (ed.) **The Universal Turing Machine: A Half-Century Survey**, Nova York: Oxford University Press, 1988.

HILBERT, D. "Naturerkennen und Logik", **Naturwissenschaften**, 18(47–49): 959–963, 1930.

HILBERT, D.; ACKERMAN, W. **Grundzüge der Theoretischen Logik**, Berlin: Springer, 1928.

HODGES, A. **Alan Turing: The Enigma**, Nova York: Simon and Schuster, 1983.

KLEENE, S. C. "General Recursive Functions of Natural Numbers", **Mathematische Annalen**, 112: 727–742, 1936.

KLEENE, S. C. "Recursive predicates and quantifiers", **Transactions of the American Mathematical Society**, 53(1): 41–73, 1936, 1943.

KLEENE, S. C. **Introduction to Metamathematics**, Amsterdam: North Holland, 1952.

LAMBECK, J. "How to Program an Infinite Abacus", **Canadian Mathematical Bulletin**, 4: 295–302, 1961.

LEWIS, H. ; PAPADIMITRIOUS, C. **Elements of the Theory of Computation**, Englewood Cliffs, NJ: Prentice-Hall, 1981.

LIN, S.; TIBOR, R. "Computer Studies of Turing Machine Problems", **Journal of the Association for Computing Machinery**, 12(2): 196–212, 1965.

MANCOSU; ZACH; BADESA "The Development of Mathematical Logic from Russell to Tarski, 1900–1935", in: HAAPARANTA, L. (ed.) **The Development of Modern Logic**, Nova York: Oxford University Press, p. 318–470, 2009 [disponível em: https://people.ucalgary.ca/~rzach/papers/history.html].

MARGENSTERN, M. "Frontier Between Decidability and Undecidability: A Survey", **Theoretical Computer Science**, 231(2): 217–251, 2000.

McCARTHY, J. "A Basis for a Mathematical Theory of Computation", in: BRAFFORT; HIRSCHBERG, **Computer Programming and Formal Systems,** Amsterdam: North-Holland, p. 33–70, 1963 [disponível em: http://www-formal.stanford.edu/jmc/basis1/basis1.html].

MINSKY, M. "Recursive Unsolvability of Post's Problem of 'Tag' and other Topics in Theory of Turing Machines", **Annals of Mathematics**, 74(3): 437–455, 1961.

MINSKY, M. **Computation: Finite and Infinite Machines**, Englewood Cliffs, NJ: Prentice Hall, 1967.

MOORE, E.F. "A simplified universal Turing machine", **Proceedings of the Association of Computing Machinery** (meetings at Toronto, Ontario), Washington, DC: Sauls Lithograph, 50–55, 1952.

MOUNIER-KUHN, P. "Logic and Computing in France: A Late Convergence", in: AISB/IACAP World Congress 2012: History and Philosophy of Programming, University of Birmingham, 2-6 July 2012 [disponível em: https://www.academia.edu/5252629/Logic_and_Computing_in_France_A_Late_Convergence].

ODIFREDDI, P. **Classical Recursion Theory**, Amsterdam: Elsevier, 1989.

PETZOLD, C. **The Annotated Turing: A Guided Tour Through Alan Turing's Historic Paper on Computability and Turing Machines,** Indianapolis, In: WILEY (2008).

POST, E. "Finite Combinatory Processes-Formulation 1", **Journal of Symbolic**

Logic, 1(3): 103–105, 1936.

POST, E. “Recursively Enumerable Sets of Positive Integers and Their Decision Problems”, **Bulletin of the American Mathematical Society,** 50(5): 284–316., 1944 [disponível em: https://projecteuclid.org/journals/bulletin-of-the-american-mathematical-society-new-series/volume-50/issue-5/Recursively-enumerable-sets-of-positive-integers-and-their-decision-problems/bams/1183505800.full].

POST, E. “Recursive Unsolvability of a Problem of Thue”, **Journal of Symbolic Logic**, 12(1): 1–11, 1947.

POST, E. “Absolutely Unsolvable Problems and Relatively Undecidable Propositions—Account of an Anticipation”, in: DAVIS, M. (ed.) **The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions**, Nova York: Raven Press, 1965. Ediçãorevisada em 2004, Dover publications, Nova York, p. 340–433.

PULLUM, G. K. “On the Mathematical Foundations of Syntactic Structures”, **Journal of Logic, Language and Information,** 20(3): 277–296, 2011.

RABIN, M. O.; SCOTT, D. “Finite Automata and their Decision Problems”, I**BM Journal of Research and Development**, 3(2): 114–125, 1959..

RADÓ, T. “On Non-Computable Functions”, **Bell System Technical Journa**l, 41(3/Maio): 877–884, 1962.

SHANNON, C.E. “A Universal Turing Machine with Two Internal States”, in: SHANNON; McCARTHY, *op. cit.*, 157–165, 1956.

SHANNON, C.E.; McCARTHY, J. (eds.) **Automata Studies**, (Annals of Mathematics Studies, 34), Princeton: Princeton University Press, 1956.

SHAPIRO, S. “Computability, Proof, and Open-Texture”, in: OLSZEWSKI, J.; JANUSZ, R. (eds.) **Church’s Thesis After 70 years**, Berlin: Ontos Verlag, p. 420–455, 2007.

SIED, W. “Mechanical Procedures and Mathematical Experience”, in: GEORGE, A. (ed.) **Mathematics and Mind**, Oxford: Oxford University Press, p. 71–117, 1994.

SIED, W. “Step by Recursive Step: Church’s Analysis of Effective Calculability”, **The Bulletin of Symbolic Logic**, 3(2): 154–180, 1997.

SIED, W. "Church without Dogma: Axioms for Computability", in:COOPER, LÖWE; SORBI (eds.) **New Computational Paradigms: Changing Conceptions of What is Computable**, Nova York: Springer Verlag, p. 139–152, 2008.

SIPSER, M. **Introduction to the Theory of Computation**, Boston: PWS Publishing, 1996.

SOARE, R. "Computability and Recursion", Bulletin for Symbolic Logic, 2(3): 284–321, 1996.

STRACHEY, C. "An Impossible Program (letter to the editor)", **The Computer Journal**, 7(4): 313, 1965.

TEUSCHER, C. (ed.)**Alan Turing: Life and Legacy of a Great Thinker,** Berlin: Springer, 2004.

TURING, A.M. "On Computable Numbers, With an Application to the Entscheidungsproblem**", Proceedings of the London Mathematical Society,** 2-42: 230–265 (1936); correção *ibid*., 2-43: 544–546 (1937).

TURING, A.M. "Computability and λ-Definability", **Journal of Symbolic Logic**, 2(4): 153–163, 1937.

TURING, A.M. "Systems of Logic Based on Ordinals", **Proceedings of the London Mathematical Society**, s2-45: 161–228, 1939.

TURING, A.M. "Lecture to the London Mathematical Society on 20 February 1947", reimpresso em CAPERPENTER, B. E.; DORAN R. W. (eds.) **Turing's ACE Report of 1946 and Other Papers: Papers by Alan Turing and Michael Woodger**, Cambridge, MA: MIT Press, 1986.

TURING, A.M. "Solvable and Unsolvable Problems", **Science News**, (February, Penguin), 31: 7–23, 1954.

WANG, H. "A Variant to Turing's Theory of Computing Machines", **Journal of the ACM**, 4(1): 63–92, 1957.

WATANABE, S. "5-Symbol 8-State and 5-Symbol 6-State Universal Turing Machines", **Journal of the ACM**, 8(4): 476–483, 1961.

WOODS; D. NEARY, T. "Small Semi-Weakly Universal Turing Machines", in: DURAND-LOSE, J.; MORGENSTERN, M. (eds.) **Machines, Computations, and Universality: 5th International Conference, MCU 2007 Orléans, França, September 10–13, 2007**, (Lecture Notes in Computer Science, 4664), Berlin: Springer, p. 303–315, 2007.

WOODS; D. NEARY, T. “The Complexity of Small Universal Turing Machines: A Survey”, **Theoretical Computer Science**, 410(4–5): 443–450, 2009.

Diagramas*

Autoria: Sun-Joo Shin, Oliver Lemon e John Mumma
Tradução: Hulian Ferreira de Araujo & Sérgio R. N. Miranda
Revisão: Guilherme A. Cardoso

Todos nós fazemos uso de raciocínios válidos, embora o raciocínio que de fato realizamos difira em vários aspectos das inferências estudadas pela maioria dos lógicos (formais). Os raciocínios realizados pelos seres humanos usualmente envolvem informações obtidas através de mais de um meio. A Lógica formal, por sua vez, tem como principal ocupação o raciocínio válido baseado em informaçoes numa única forma: a forma sentencial. Recentemente, muitos filósofos, psicólogos, lógicos, matemáticos e cientistas da computação se atentaram para a importância dos raciocínios multimodais e, além disso, muitas pesquisas têm sido feitas na área de sistemas de representações não simbólicas,

*SHIN, SUN-JOO; LEMON, OLIVER; MUMMA, JOHN, "Diagrams", In: ZALTA, E. N. (ed.) **Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/diagrams/. Acesso em: 05 jan. 2022.

The following is the translation of the entry on Diagrams by Shin, Lemon and Mumma in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/diagrams/. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the **Stanford Encyclopedia of Philosophy**, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

especialmente as diagramáticas.[39] Esse verbete oferece um panorama das direções gerais dessa nova área de pesquisa e dá ênfase ao status lógico dos diagramas em provas, na sua função representacional e adequação, nos diferentes tipos de sistemas diagramáticos e no papel dos diagramas na cognição humana.

1. Introdução

Diagramas ou imagens provavelmente estão entre as mais antigas formas de comunicação humana. Eles são usadas não só para representar, mas também podem ser usados para executar certos tipos de raciocício; desse modo, desempenham um papel especial na lógica e na matemática. No entanto, sistemas de representação sentencial (como a lógica de primeira ordem) têm sido dominantes na história da lógica moderna, enquanto os diagramas têm sido largamente encarados como possuindo um interesse menor. Diagramas são usualmente adotados como ferramentas heurísticas ao se explorar uma prova, mas não como uma parte mesmo da prova.[40] É um movimento recente entre filósofos, lógicos, cientistas cognitivos e cientistas da computação dar ênfase aos diferentes tipos de sistemas representacionais, e muitos pesquisadores têm particularmente se interessado por sistemas de representação diagramáticas.

Desafiando assim um longo e insistente preconceito contra representações diagramáticas, aqueles que trabalham com raciocício multimodal têm adotado diferentes tipos de abordagens que podemos categorizar em três grupos distintos. Um desses desdobramentos pode ser encontrado na filosofia da mente e ciências cognitivas. Visto que os limites das formas linguísticas são claros para aqueles que trabalham com raciocínio e representações mentais,

[39] A título de ilustração, a mais recente conferência sobre o tópico, **Diagrams, 2018** [http://www.diagrams-conference.org/2018/.], mostra a dimensão interdisciplinar da pesquisa na área.

[40] É uma questão interessante o porquê disso ser assim. A possibilidade de engano decorrente do uso dos diagramas e seu limitado poder expressivo têm sido vistos como as razões principais para seu abandono. Mark Greaves (2002) explora as raízes históricas e filosóficas para a aceitação e a rejeição dos diagramas no contexto das provas lógicas e geométricas, e conclui que a questão da invenção ou uso de um sistema diagramático envolve se o inventor ou usuário aceita uma certa metafísica.

alguns filósofos e cientistas cognitivos adotaram essa nova direção dos raciocínios multimodais com entusiasmo e têm explorado o raciocínio humano e representação modal envolvendo formas não-linguísticas. (CUMMINS, 1996; CHANDRASEKARAN *et al.*, 1995). Outra frente de trabalho em raciocínios diagramáticos mostra que não há diferença intrínseca entre sistemas representacionais simbólicos e diagramáticos no que diz respeito ao seu status lógico. Alguns lógicos apresentaram estudos de caso a fim de provar que sistemas diagramáticos são completos e corretos do mesmo modo que sistemas simbólicos o são. Esse tipo de resultado refuta diretamente a difundida suposição de que diagramas são inerentemente enganadores, e extingue objeções teóricas a respeito do uso de diagramas em provas (SHIN, 1994; HAMMER, 1995a). Uma terceira direção dentro dos raciocínios multimodais foi tomada pelos cientistas da computação, que posssuem um interesse muito mais prático do que os outros grupos. De maneira não surpreendente, esses que se debruçam sobre as várias áreas da ciência da computação - por exemplo, representação de conhecimento, design de sistemas, programação visual, design GUI, entre outros - encontraram novas oportunidades empolgantes nesse conceito de "sistema heterogêneo" e têm implementado representações diagramáticas nas suas áreas de pesquisa.

Os nossos objetivos com este verbete são os seguintes. Antes de tudo, gostaríamos de familiarizar o leitor com os detalhes de alguns sistemas diagramáticos. Ao mesmo tempo, o verbete irá tratar de questões teóricas ao explorar a natureza dos sistemas diagramáticos e do raciocínio em termos de poder expressivo e correção. O estudo de caso da segunda seção irá não só safisfazer nosso primeiro objetivo, mas também fornecerá um material robusto para uma discussão mais geral e teórica na terceira seção. A quarta seção apresenta outro estudo de caso e o considera sob a luz da discussão geral da terceira seção. Conforme mencionado acima, o tópico dos diagramas tem atraído muita atenção com resultados importantes em vários domínios de pesquisa. Portanto, nossa quinta seção visa introduzir várias abordagens aos raciocínios diagramáticos presentes nesses diferentes domínios.

Para a discussão abaixo, precisamos clarificar dois usos distintos, embora relacionados, da palavra "diagrama": diagrama enquanto representações mentais internas e diagrama enquanto representações externas. Esta citação de

Chandrasekaran *et al.* (1995, p. xvii) expressa sucintamente a distinção entre representações diagramáticas internas e externas:

- **Representações diagramáticas externas**: Construídas pelo agente num meio no mundo externo (papel, etc.), mas tomadas como representações pelo agente.
- **Diagramas internos ou Imagens**: Incluem as (controversas) representações internas postuladas de modo a possuirem algumas propriedades pictóricas.

Como veremos, enquanto os lógicos direcionam a atenção para os sistemas diagramáticos externos, o debate acerca das imagens na filosofia da mente e nas ciências cognitivas é majoritariamente a respeito dos diagramas internos, e a pesquisa do papel cognitivo dos diagramas aborda ambas as formas.

2. Diagramas enquanto Sistemas Representacionais

A predominância dos sistemas de representação sentencial na história da lógica moderna ocultou vários fatos importantes sobre os sistemas diagramáticos. Um desses fatos é que vários sistemas diagramáticos bem conhecidos estavam disponíveis enquanto ferramentas heurísticas antes mesmo do surgimento da lógica moderna. Círculos de Euler, diagramas de Venn e os quadrados de Lewis Carroll eram amplamente adotados para certos tipos de argumentos silogísticos (EULER, 1768; VENN, 1881; CARROLL, 1896). Outro ponto intreressante da história, embora negligenciado, foi que um dos fundadores da lógica simbólica moderna, Charles Peirce, não apenas revisou os diagramas de Venn como também criou um sistema gráfico, chamado de Grafos Existênciais, que se provou ser equivalente à linguagem de predicados (PEIRCE, 1933; ROBERTS, 1973; ZEMAN, 1964).

Esses diagramas inspiraram alguns pesquisadores que recentemente chamaram a nossa atenção para as representações multimodais. Lógicos que participam desse projeto exploram o tópico em duas frentes distintas. A primeira delas é a ênfase nas representações externas, em detrimento das representações internas. A segunda consiste em estabelecer o status lógico dos sistemas

representacionais, em vez de explicar o seu poder heurístico, avaliando a correção e o poder expressivo de uma classe de sistemas representacionais. Caso um sistema falhe ao estabelecer a sua correção ou mesmo possua poder expressivo insatisfatório, o interesse do lógico nesse sistema em específico desaparecerá (SOWA, 1984; SHIN, 1994).

Nessa seção, oferecemos um panorama do desenvolvimento histórico dos diagramas de Euler e de Venn como estudo de caso com o propósito de evidenciar os seguintes aspectos. Primeiramente, esse processo mostra como a simples intuição dos matemáticos acerca dos raciocínios silogísticos diagramáticos foi gradualmente desenvolvida em direção aos sistemas de representações formais. Em segundo lugar, observamos diferentes ênfases dadas nos diferentes estágios de modificação dos sistemas diagramáticos. Em terceiro lugar, e não menos importante, esse desenrolar histórico ilustra uma tensão interessante num balanço entre o poder expressivo e a clareza visual desses sistemas. Mais importante ainda, o leitor verá que os lógicos colocam em questão se há alguma razão intrínseca por que os sistemas sentenciais, mas não os sistemas diagramáticos, fornecem provas rigorosas, e que eles com sucesso respondem negativamente a essa questão.

Desse modo, o leitor não ficará surpreso com a seguinte conclusão de Barwise e Etchementdy, os primeiros lógicos a se debruçarem numa investigação das provas diagramáticas na lógica,

> [que] não há a princípio uma distinção entre os formalismos inferenciais que adotam textos e aqueles que usam diagramas. Podemos ter sistemas formais rigorosos e logicamente corretos (e completos) baseados em diagramas (BARWISE; ETCHEMENDY, 1995, p.214)

Essa convicção foi necessária para o nascimento do inovador programa de computador *hyperproof*, que adota linguagens de primeira ordem e diagramas (num sistema multimodal) para ensinar lógica elementar (BARWISE & ETCHEMENDY, 1993 e 1994).

2.1. Diagramas de Euler

Leonhard Euler, um matemático do século XVIII, adotou curvas fechadas (círculos) para ilustrar o raciocínio silogístico (EULER, 1768). Os quatro tipos de sentenças categóricas são representados por ele conforme mostrado na figura 1.

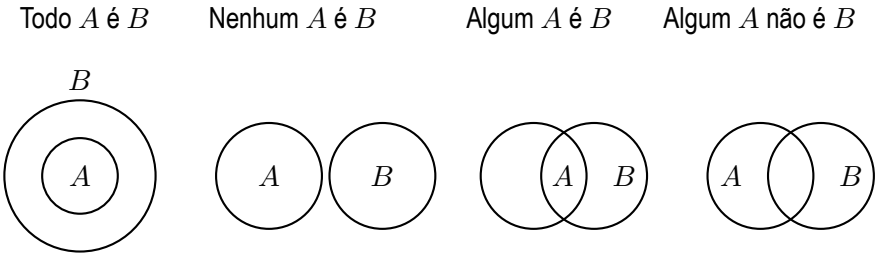


Figura 1: Diagramas de Euler

Para os enunciados universais, o sistema emprega relações espaciais entre os círculos de uma maneira bem intuitiva: se o círculo denominado *"A"* está **incluído** no círculo *"B"*, então o diagrama representa a informação de que todo *A* é *B*. Se não há **sobreposição** entre os dois círculos, então o diagrama transmite a informação de que nenhum *A* é *B*.

Essa representação segue a seguinte convenção:[41]

> Todo objeto x no domínio corresponde a uma única posição $l(x)$ no plano de tal modo que $l(x)$ está na região *R* se e somente se x é membro do conjunto que a região *R* representa.

O poder dessa representação reside no fato de que um objeto ser membro de um conjunto é facilmente conceitualizado como um objeto caindo dentro de um conjunto, tal como posições na página são consideradas como caindo dentro ou fora dos círculos desenhados. O poder do sistema consiste no fato de que não é

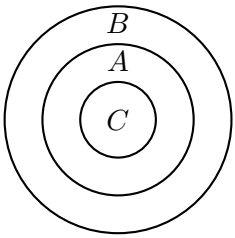
[41] Notemos que, embora a convenção soe natural, ainda assim ela é arbitrária. Por exemplo, os sistemas de Lambert e Englebretsen visualizam indivíduos como pontos e conjuntos como linhas (LAMBERT, 1764; ENGLEBRETSEN, 1992)

preciso convencões adicionais para estabelecer o significado dos diagramas que recorrem a mais que um círculo: relações entre conjuntos são asseridas por meio das mesmas relações presentes nos círculos que as representam. A representação das duas afirmações universais, "Todo *A* é *B*" e "Nenhum *A* é *B*" ilustra a força do sistema.

Essa clareza já não é mais preservada na afirmações existenciais. Euler justifica o diagrama de "Algum *A* é *B*" dizendo que podemos inferir **visualmente** que algo em *A* está contido em *B* porque parte da área *A* está contida na área *B* (EULER, 1768, p.233). Obviamente, o próprio Euler acreditava que o mesmo tipo de relação de contenção visual poderia ser nesse caso assim como era no caso dos enunciados universais. No entanto, a crença de Euler não é correta e essa representação coloca uma ambiguidade danosa. No diagrama, não apenas parte do círculo *B* está contido em *A* (como Euler mesmo descreve), mas o seguinte também é verdadeiro: (i) parte do círculo *B* está contido em *A*, (ii) parte do círculo *A* não está contido em *B*, (iii) parte do círculo *B* não está contido em *A*. Ou seja, o terceiro dos diagramas pode ser lido não só como "Algum *A* é *B*", mas também como "Algum *B* é *A*", "Algum *A* não é *B*" ou "Algum *B* não é *A*". Para evitar essa ambiguidade, precisamos estabelecer outras convenções.[42]

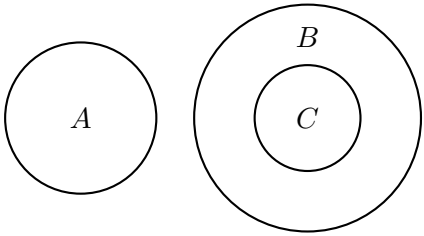
Os próprios exemplos de Euler ilustram bem as forças e as fraquezas de seu sistema.

Exemplo 1: Todo A é B. Todo C é A. Logo, todo C é B.



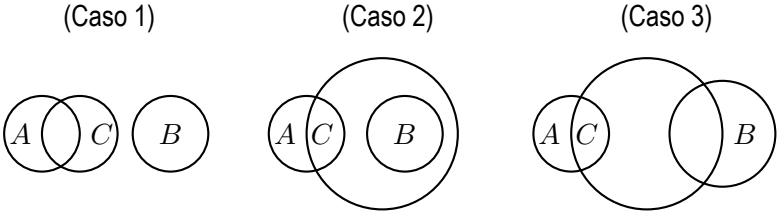
[42] Para mais detalhes, *vide* HAMMER; SHIN, 1998.

Exemplo 2: Nenhum A é B. Todo C é B. Logo, nenhum C é A.



Em ambos os exemplos, o leitor pode facilmente inferir a conclusão, e isso ilustra o poder visual dos diagramas de Euler. Todavia, quando os enunciados existenciais são representados, as coisas ficam bastante complicadas, conforme exposto acima. Por exemplo:

> **Exemplo 3**. Nenhum *A* é *B*. Algum *C* é *A*. Logo, algum *C* não é *B*.
> Não há um diagrama único que representa as duas premissas, uma vez que a relação entre os conjuntos *B* e *C* não pode ser inteiramente especificada com um único diagrama. Em vez disso, Euler sugere os três possíveis casos seguintes:



Euler afirma que a proposição “Algum *C* não é *B*” pode ser extraída de todos esses diagramas. No entanto, está longe de ser visualmente claro como o usuário pode extrair essa proposição nos dois primeiros casos, uma vez que pode extrair “Nenhum *C* é *B*” do caso 1 e “Todo *B* é *C*” do caso 2.

Desse modo, a representação de enunciados existenciais não só ofusca a clareza visual dos Círculos de Euler, mas também coloca sérios problemas

interpretativos no sistema. O próprio Euler parece reconhecer esse problema potencial e introduz um novo dispositivo sintático, "*****" (representando o não-vazio), como uma tentativa de reparo a esse problema (1768; carta 105). Não obstante, um problema mais sério se coloca quando esse sistema falha ao representar certos fragmentos compatíveis (isto é, consistentes) de informação num único diagrama. Por exemplo, o sistema de Euler impede, apenas recorrendo a um único diagrama, de representar o seguinte par de enunciados: (i) "Todo *A* é *B*" e "Nenhum *A* é *B*" (que é consistente caso *A* seja um conjunto vazio). (ii) "Todo *A* é *B*" e "Todo *B* é *A*" (que é consistente quando *A*=*B*). (iii) "Algum *A* é *B*" e "Todo *A* é *B*" (suponha que desenhamos um diagrama de Euler para o primeiro enunciado e tente adicionar um fragmento de informação, ou seja, o segundo enunciado, ao diagrama existente). Tal falha é intimamente relacionada com as motivações de Venn para a criação do seu próprio sistema diagramático (*vide* **seção 3.1** para outras falhas do sistema de Euler).

2.2. Diagramas de Venn

A crítica de Venn aos círculos de Euler pode ser resumidas na seguinte passagem:

> O ponto fraco nesse [diagrama de Euler] e em outros esquemas similares consiste no fato de que apenas ilustra estritamente as relações atuais das classes entre si, em vez do conhecimento imperfeito dessas relações que podemos ter ou querer transmitir por meio de uma proposição. (VENN, 1881; p. 510)

Devido ao seu rigor, o sistema de Euler por vezes falha em representar fragmentos consistentes de informações em um único diagrama, conforme mostrado acima. Além da limitação expressiva, o sistema de Euler também sofre de outros tipos de limitações expressivas com respeito aos conjuntos não vazios, por conta de restrições topológicas presentes nas figuras planas (*vide* **Seção 3.1**).

O sistema proposto por Venn (1881) superou essas limitações

expressivas de tal modo que essas informações parciais pudessem ser representadas. A solução foi sua ideia de “diagramas primários”. Um diagrama primário representa todas as relações conjuntistas entre um número de conjuntos sem fazer quaisquer comprometimentos existenciais sobre esses conjuntos. Por exemplo, a figura 2 mostra o diagrama primário sobre os conjuntos *A* e *B*.

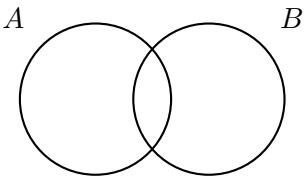


Figura 2: Diagramas Primários de Venn

No sistema de Venn, esse diagrama não transmite nenhuma informação específica sobre a relação desses dois conjuntos. Essa é uma diferença crucial entre os diagramas de Euler e de Venn.

Para a representação de enunciados universais, distintamente das visualmente claras relações de contenção dos diagramas de Euler, a solução de Venn é “sombrear [as áreas apropriadas]” (VENN, 1881; p. 122). Ao empregar esse dispositivo sintático, obtemos diagramas para enunciados universais conforme mostra a figura 3.

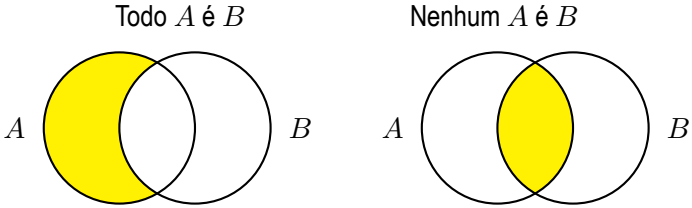
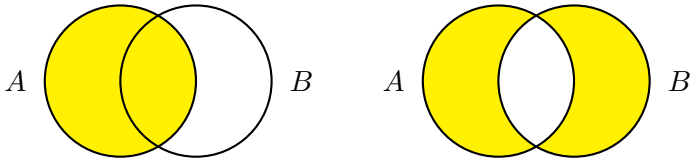


Figura 3: Sombreamento de Venn

A escolha por Venn pelo sombreamento não é inteiramente arbitrária se considerarmos que o sombreamento pode ser interpretado como a visualização da vacuidade de conjuntos. Contudo, deve-se notar que o sombreamento é

também um novo dispositivo sintático não adotado por Euler. Essa revisão deu flexibilidade para o sistema de tal modo que certos fragmentos de informações podem ser representados num único diagrama. No que se segue, o diagrama à esquerda combina dois fragmentos de informações, "Todo *A* é *B*" e "Nenhum *A* é *B*", para visualmente transmitir a informação de que "Nada é *A*". O diagrama à direita, que representa ambas "Todo *A* é *B*" e "Todo *B* é *A*", mostra claramente que *A* é igual a *B*.



O uso de diagramas primários acaba evitando outros problemas de expressividade (a respeito das propriedades espaciais dos objetos do diagrama) tratados abaixo na Seção 3. Surpreendentemente, Venn nada disse a respeito da representação dos enunciados existenciais, que foram uma dificuldade para os diagramas de Euler. Podemos imaginar que ele poderia ter introduzido outro tipo de objeto sintático representando o comprometimento existencial. Foi exatamente o que Charles Peirce fez vinte anos mais tarde.

2.3. Extensão de Peirce

Peirce ressalta que o sistema de Venn não tem recursos para representar os seguintes tipos de informação: enunciados existenciais, informação disjuntiva, probabilidades e relações. Ele procurou estender o sistema de Venn a fim de representar os dois primeiros tipos, a saber, enunciados existenciais e disjuntivos. A extensão foi feita através de três dispositivos. (i) Substituição da representação do vazio feita por meio do sombreamento do vazio com um novo símbolo, "o". (ii) Introdução do símbolo "x" com significação existencial. (iii) Para a informação disjuntiva, introdução de um símbolo linear "-" que conecta "o" e "x".

Por exemplo, a Figura 4 representa o enunciado, "Todo *A* é *B* ou algum

A é *B*", que nem Euler e nem Venn com seus sistemas poderiam representar num simples diagrama.

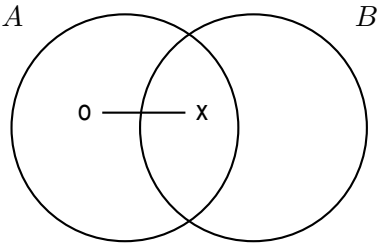
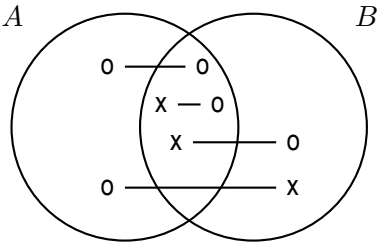


Figura 4: Diagrama de Peirce

A razão por que Peirce substituiu o sombreamento de Venn pelo símbolo "o" é clara: não seria fácil conectar sombreamentos ou sombreamentos com "x" para representar a informação disjuntiva. Desse modo, Peirce de fato ampliou o poder expressivo do sistema, mas essa mudança teve os seus custos.

Por exemplo, o diagrama abaixo representa o enunciado "**Ou** todo *A* é *B* e algum *A* é *B*, **ou** nenhum *A* é *B* e algum *B* não é *A*":



Ler esse diagrama requer mais do que notar a contenção visual dos círculos (como nos diagramas de Euler) ou sombreamentos (como nos diagramas de Venn), mas também demanda convenções extras na leitura das combinações dos símbolos "o", "x" e das linhas. De fato, essas novas convenções de Peirce ampliaram o poder expressivo dos diagramas, mas a arbitrariedade das suas convenções e representações mais confusas (como no diagrama acima)

sacrificaram a clareza visual presente no sistema original de Euler. Nesse ponto, o próprio Peirce confessa que "há uma grande complexidade na expressão que é essencial ao significado (...)"(PEIRCE, 1933, p. 4365). Assim, com a revisão de Peirce, grande parte das motivações de Euler quanto à visualização que esses diagramas forneceriam foi perdido, com a exceção de que um objeto geométrico (um círculo) é usado para representar conjuntos (possivelmente vazios).

Outra contribuição importante de Peirce para o estudo dos diagramas consiste na seguinte observação:

> "Regra" aqui é empregado no mesmo sentido que falamos das "regras" da álgebra; isto é, como uma permissão sob condições estritamente definidas (PEIRCE, 1933, p. 4361).

Peirce provavelmente foi o primeiro a propor a discussão sobre regras de transformação num sistema de representação não sentencial. Assim como efetivamente fazem as regras da álgebra, as regras de manipulação de diagramas deveriam nos dizer quais manipulações de símbolos são ou não permitidas. Algumas das seis regras propostas por Peirce demandam clarificação e acabam se mostrando incompletas - um problema que ele mesmo já antecipara. Todavia, e mais importante ainda, Peirce não tinha a disposição nenhuma ferramenta teórica - como por exemplo uma distinção clara entre sintaxe e semântica - para convencer o leitor que cada regra é correta ou para determinar se mais regras são necessárias. Em suma, a justificação da sua importante intuição (que pode haver regras de tranformação para diagramas) fica em aberto.

2.4. Diagramas enquanto sistemas formais

Shin (1994) seguiu os esforços de Peirce em duas direções. Uma foi desenvolver a versão Peirceana dos diagramas de Venn, a outra foi provar a correção e completude desse sistema.

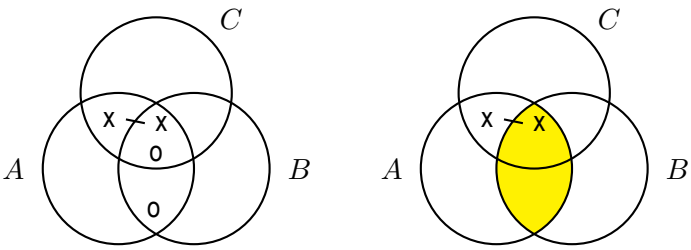
O trabalho de Shin ajusta as modificações feitas por Peirce nos diagramas de Venn a fim de obter um aumento no poder expressivo sem incorrer em uma grande perda de clareza visual. Tal revisão se deu em dois estágios: (i) Venn-I: preserva o sombreamento de Venn (para representar o vazio), o "x" proposto

por Peirce (com a significação existencial) e a linha que conecta os "x's" (para a informação disjuntiva). (ii) Venn-II: esse sistema, provado como equivalente ao cálculo de predicados monádico, é o mesmo que Venn-I, exceto pela introdução de uma linha que conecta diagramas para representar a informação disjuntiva.

Se voltarmos a um dos exemplos propostos por Euler, veremos o contraste entre essas diferentes versões:

Exemplo 3. Nenhum *A* é *B*. Algum *C* é *A*. Logo, algum *C* não é *B*.

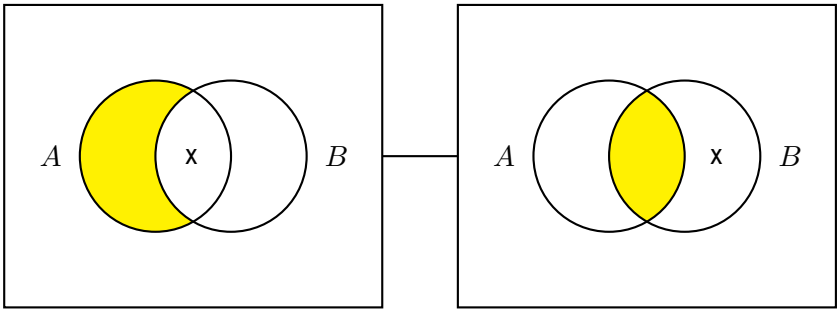
Euler mesmo admite que não pode ser desenhado um único diagrama que represente as duas premissas, mas três casos possíveis devem ser construídos. O sistema de Venn não diz nada sobre enunciados existenciais. Agora, os sistemas de Peirce e Shin representam as duas premissas num diagrama único do seguinte modo:



Peirce Shin

No caso do diagrama de Shin, a convenção de sombreamento para a vacuidade, diferentemente do "o" sugerido por Peirce, leva o leitor mais naturalmente à conclusão "Algum *C* não é *B*" do que no caso do diagrama de Peirce.

Entretanto, Venn-I não representa a informação disjuntiva entre enunciados universais ou entre enunciados universais e existenciais. Venn-II, por sua vez, preservando o poder expressivo de Venn-I, permite a conexão entre diagramas com uma linha. Desse modo, o diagrama de Peirce acima, de visualização confusa, equivale ao seguinte diagrama no sistema Venn-II:



Além dessa revisão, Shin (1994) apresentou cada um dos dois sistemas com uma representação formal munida de sua própria sintaxe e semântica. A sintaxe nos diz quais dos diagramas são aceitáveis, isto é, quais são bem-formados, e quais manipulações são permitidas em cada sistema. A semântica define as consequências lógicas entre diagramas. Com esses recursos, prova-se que os sistemas são corretos e completos, no mesmo sentido que algumas lógicas simbólicas o são.

Essa abordagem colocou desafios fundamentais a algumas suposições sobre sistemas representacionais. Dado o desenvolvimento da lógica moderna, conceitos importantes (como sintaxe, semântica, inferência, consequência lógica, validade e completude) foram empregados apenas em sistemas de representação sentenciais. No entanto, nenhum desses conceitos se mostrou como intrínseco a esses sistemas. Para qualquer sistema representacional, seja sentencial ou diagramático, podemos realizar discussões em dois níveis, um sintático e um semântico. O que as regras de inferência nos dizem é como manipular uma dada unidade, seja simbólica ou diagramática, em outra. A definição de consequência lógica é também independente de qualquer forma específica de sistema representacional. O mesmo argumento se segue para as provas de correção e completude. Quando se prova que um sistema é correto, devemos ser capazes de adotá-lo em provas. De fato, muitas pesquisas exploram atualmente o emprego de diagramas em provas automáticas de teoremas (BARKER-PLUMMER; BAYLIN, 1997; JAMNIK *et al.*, 1999).

2.5. Círculos de Euler revisitados

É interessante e importante notar que essas mudanças graduais feitas dos círculos de Euler até os sistemas de Shin compartilham de um tema: o aumento do poder expressivo e lógico do sistema de tal modo a torná-lo correto, completo e logicamente equivalente ao cálculo de predicados monádico. A principal revisão de Euler para Venn, ao introduzir os diagramas primários, nos permite representar o conhecimento parcial sobre as relações entre conjuntos. A extensão nos diagramas de Venn para os de Peirce é feita de tal modo que a informação existencial e a disjuntiva possam ser representadas mais eficientemente.

Ambos Venn e Peirce adotam o mesmo tipo de solução buscando aperfeiçoamentos em seus sistemas: introduzir novos objetos sintáticos, isto é, os sombreamentos de Venn e os "x's", "o's" e as linhas de Peirce. No entanto, de modo negativo, esses sistemas revisados sofrem pela perda da clareza visual, conforme vimos acima, exatamente devido a introdução de convenções um tanto quanto arbitrárias. As modificações nos diagramas de Peirce e Shin visam restaurar a clareza visual, mas sem perda do poder expressivo.

Hammer e Shin adotam um caminho diferente nessas revisões: reviver a relação homomórfica entre círculos e conjuntos - contenção entre círculos representa a relação de subconjunto entre conjuntos, e a não-sobreposição de regiões representa a relação de disjunção - e ao mesmo tempo adotar os diagramas primários como padrão. Por outro lado, esse sistema de Euler com acréscimos não é suficiente para expressar raciocínios silogísticos, pois não representa enunciados existenciais. Para mais detalhes sobre o sistema, *vide* HAMMER; SHIN, 1998.

Esse estudo coloca uma questão interessante para as futuras pesquisas sobre o raciocínio diagramático. No decorrer dos diferentes desenvolvimentos feitos a partir dos diagramas de Euler, acrescentar poder expressivo e aperfeiçoar a clareza visual parecem ser complementares entre si. A depender dos propósitos, precisamos priorizar um em relação ao outro. O sistema proposto por Hammer e Shin fornece um modelo simples para o desenvolvimento de outros sistemas de representação não sentencial, um tópico que tem recebido atenção crescente nas ciências da computação e nas ciências cognitivas.

3. Consequências das propriedades espaciais dos diagramas

Embora seja possível oferecer aos diagramas o mesmo status lógico das fórmulas (conforme defendido acima), ainda há diferenças importantes (que podem ter ramificações para a correção do sistema) entre diagramas e os tradicionais cálculos lineares. Um ponto importante a se notar acerca dos diagramas (comparar com RUSSELL, 1923) é que as relações espaciais entre os objetos do diagrama podem ser usadas para representar relações entre objetos de outra natureza. Linguagens sentenciais (lógica simbólica, linguagem natural), contudo, usam apenas a relação de concatenação para representar relações entre objetos. O uso representacional das relações espaciais no caso dos diagramas é direto e intuitivo, conforme visto no desenvolvimento dos diagramas de Euler, mas também possui seus riscos, como veremos. Pode-se esperar que as relações espaciais, peculiares aos sistemas diagramáticos, sejam fontes importantes tanto da força quanto de suas fraquezas. Considerações psicológicas acerca da capacidade humana para o processamento visual da informação e a habilidade na consecução de raciocínios espaciais qualitativos também têm ramificações relevantes para a eficácia de raciocínios com diagramas, mas não as abordaremos nesse verbete.

Uma característica peculiar de diagramas é que obedecem a certas restrições "nômicas" ou "intrínsecas" devidas ao uso de superfícies planas como meio de representação. O ponto é que as linguagens sentenciais são baseadas em sinais acústicos que são sequenciais por natureza, devendo haver, portanto, uma sintaxe complexa para que se possa representar certas relações - por sua vez, diagramas, por serem bi-dimensionais, podem apresentar algumas relações sem o intermédio de uma sintaxe complexa (STENNING & LEMON, 2001). Diagramas exploram essa possibilidade - o uso das relações espaciais para representar outras relações. A questão é: em que medida relações espaciais e objetos podem representar bem outros (possivelmente mais abstratos) objetos e relações?

O raciocínio lógico com diagramas é realizado com frequência em virtude de ser uma representação de todos os modelos possíveis de uma situação, até a equivalência topológica dos diagramas (isso, claro, depende do sistema

diagramático particular em questão). Um diagrama é muitas vezes uma abstração de uma classe de situações, e, uma vez construído o diagrama adequado, inferências podem simplesmente ser lidas na representação sem qualquer manipulação. Em alguns sistemas diagramáticos (por exemplo, círculos de Euler) inferências são feitas por meio da construção correta do diagrama e da extração das informações do diagrama construído. A complexidade presente no uso de regras de inferência na lógica simbólica é substituída em sistemas diagramáticos pelo problema de construir corretamente diagramas.[43] Por exemplo, um diagrama de Euler empreende capturar relações entre conjuntos com relações topológicas entre regiões do plano de tal modo que represente todas as maneiras possíveis que uma certa coleção de enunciados conjuntistas possa ser verdadeira. Isso tem duas consequências importantes: (1) se um certo diagrama não pode ser desenhado, então a situação descrita deve ser impossível (chamada de "auto-consistência"); (2) se uma certa relação entre objetos do diagrama deve ser desenhada, então a relação correspondente é uma relação logicamente válida (*vide* vários exemplos na seção 2). Esse fenômeno por vezes é chamado de "carona" (BARWISE; SHIMOJIMA, 1995). Esse modo de raciocínio diagramático é dependente de um uso representacional particular dos diagramas - que eles representam uma classe de modelos. Se uma dada classe de modelos não pode ser representada por um sistema diagramático, então tais casos não serão considerados nas inferências desse sistemas, e inferências incorretas poderiam ser feitas. Tal fato faz a adequação representacional dos diagramas, restritas pela sua natureza espacial, de importância primordial, como exploraremos abaixo.

3.1. Limitações da representação e do raciocinio diagramático

O uso representacional de relações espaciais num plano restringe de modo significativo a representação diagramática e, portanto, o raciocínio com diagramas. Em especial, há propriedades geométricas e topológicas (que agrupamos como "espaciais") que limitam o poder expressivo desses sistemas. Por exemplo, dentro da teoria dos grafos é um fato que algumas estruturas

[43] Esses problemas têm sido estudados com a nomenclatura de "inferência topológica" e na sua grande maioria são NP-difíceis (GRIGNI et al, 1995; LEMON & PRATT, 1997; LEMON, 2001).

simples não podem ser desenhadas no plano. Por exemplo, o grafo K_5 é um grafo com 5 nós, cada um dos quais está conectado com outro por um arco. Esse grafo não é planar, ou seja, ele não pode ser representado sem que ao menos dois arcos se cruzem. Isso é um exemplo do tipo de restrição para possíveis diagramas que limitam seu poder expressivo. Agora, uma vez que raciocínios diagramáticos podem ocorrer pela enumeração de todos os modelos possíveis de uma situação, essa inadequação (um tipo de incompletude) torna muitos sistemas incorretos para o emprego em raciocínios lógicos. (ver a crítica de ENGLEBRETSEN, 1992 em LEMON & PRATT, 1998).

Um dos exemplos mais simples desse tipo de impasse nos é dado por Lemon e Pratt (1997).[44] Consideremos diagramas de Euler - em que regiões do plano representam conjuntos, e a sobreposição dessas regiões representa interseção não vazia de conjuntos. Um resultado oriundo da topologia convexa conhecido como Teorema de Helly diz (para o caso de 2 dimensões) que se qualquer tripla de 4 regiões convexas possui uma interseção não vazia, então todas as quatro regiões devem ter uma interseção não vazia.

Para compreender as ramificações desse resultado, considere o seguinte problema:

Exemplo 4. Usando círculos de Euler, represente as premissas abaixo:

- $A \cap B \cap C \neq \emptyset$
- $B \cap C \cap D \neq \emptyset$
- $C \cap D \cap A \neq \emptyset$

Note que, em termos da teoria de conjuntos, apenas consequências triviais se seguem das premissas acima. Entretanto, como podemos notar no diagrama de Euler abaixo, os diagramas conduzem à conclusão incorreta de que que $A \cap B \cap C \cap D \neq \emptyset$ (devido à região sobreposta quatro vezes no centro).

[44] Atualmente, Ian Pratt-Hartmann.

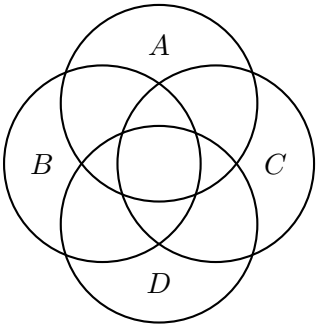


Figura 5: uma representação por círculos de Euler ilustrando o Teorema de Helly

Dito de outro modo, ao adotar os círculos de Euler para representar a relação entre conjuntos, somos forçados[45] a aceitar uma relação entre conjuntos que não é logicamente necessária. Isso significa que há situações logicamente possíveis que o sistema não pode representar e que o usuário viria a fazer inferências incorretas apoiando-se nesse sistema para raciocinar. De forma geral, resultados como esse podem ser gerados em diferentes tipos de sistemas diagramáticos, a depender das relações espaciais e os objetos assumidos no sistema - um programa de pesquisa ainda em desenvolvimento.

Por exemplo, usar regiões não convexas (por exemplo, "borrões" em vez de círculos) conduz a um problema similar, mas dessa vez envolvendo grafos não-planares em vez do Teorema de Helly. Um resultado relacionado versa sobre diagramas lineares para representação de silogismos (ENGLEBRETSEN, 1992), no qual linhas representam conjuntos, pontos representam indivíduos, interseção linha-ponto representa pertencimento e interseção de linhas representa interseção de conjuntos. Mais uma vez, restrições de planaridade restringem o poder expressivo do sistema e levam a inferências incorretas.

A "hipótese de restrição" de Atsushi Shimojima expressa bem o que está em questão:

> Representações são objetos no mundo e

[45] Como uma exemplificação prática do teorema de Helly.

> enquanto tais obedecem restrições estruturais no que diz respeito a sua construção. A variação no potencial inferencial dos diferentes modos de representação é atribuível aos diferentes modos que essas restrições estruturais nas representações coincidem com as restrições nos objetos de representação. (SHIMOJIMA, 1996a; SHIMOJIMA, 1999)

3.2. Eficácia dos diagramas

Conforme discutido acima, muito do interesse nos diagramas surgiu a partir da crença de que para certos tipos de tarefa eles são de algum modo mais "efetivos" em comparação com as representações lógicas tradicionais. Por exemplo, um mapa certamente auxilia mais a navegação do que uma descrição detalhada da paisagem. Não obstante, embora haja claras vantagens psicológicas obtidas através do uso de diagramas, eles são muitas vezes inficazes para representar objetos abstratos e relações (como no caso dos círculos de Euler). Antes uma noção puramente intuitiva, afirmações não psicológicas sobre a "eficácia" de sistemas diagramáticos podem ser examinadas em termos de propriedades formais de linguagens (LEMON *et al.*, 1999). Em particular, muitos sistemas diagramáticos são auto-consistentes, incorretos, incompletos e a complexidade das inferências com diagramas é NP-difícil. Em comparação, grande parte das lógicas sentenciais, embora capazes de expressar inconsistências, são completas e corretas.[46]

Por outro lado, não ser capaz de representar contradições pode nos prover com *insights* interessantes acerca da natureza da representação diagramática. Se um dos objetivos centrais de uma linguagem é representar o mundo ou um estado de coisas, então representar contradições e tautologias é um ponto em questão. Contradições e tautologias não fazem parte do mundo. Como podemos representar a imagem, ou ter uma imagem, da contradição "está

[46] Para estudos mais recentes desse tópico, remetemos aos artigos de Aiello e van Benthem, Fisler e Lemon, todos presentes em BARKER-PLUMMER *et al*, 2002.

chovendo e não está chovendo"? E quanto à imagem da informação disjuntiva "está chovendo ou não está chovendo"? Agora estamos mais perto da clássica concepção pictórica wittgensteiniana da linguagem (WITTGENSTEIN, 1921).

4. Sistemas diagramáticos na geometria

Matemáticos usaram, e seguem usando, diagramas amplamente. A comunicação de conceitos e provas matemáticas - em manuais, no quadro-negro - não é uniformemente sentencial. Imagens e figuras são comuns. Alinhados com a concepção prevalente da lógica sendo essencialmente sentencial, entretanto, matemáticos não os tomam como tendo papel essencial e rigoroso no raciocínio matemático. Seu uso é restrito a auxiliar a compreensão de uma prova e não são usualmente considerados como constituintes da prova.

Essa atitude é bem ilustrada pela avaliação padrão da metodologia de Euclides na obra **Os Elementos**. Não há área da matemática em que diagramas são tão proeminentes quanto na geometria elementar desenvolvida por Euclides nesse texto. Em certo sentido, as provas nessa área parecem ser sobre os diagramas de triângulos e círculos que aparecem com elas. Esse é o caso especialmente das provas geométricas em **Os Elementos**. Os diagramas para Euclides não eram meramente ilustrativos. Alguns passos inferenciais dependem de uma construção adequada de um diagrama. Na história oficial, esses passos indicam falhas nas provas. Eles mostram como Euclides não conseguiu completamente seguir com seu projeto de desenvolver axiomaticamente a geometria.

Ken Manders buscou explorar essa história no seu trabalho seminal "O Diagrama Euclidiano"(2008[1995]). Sua análise do método de prova diagramático de Euclides mostra que Euclides emprega os diagramas de maneira sistemática e controlada. Isso coloca em questão a avaliação negativa do rigor em **Os Elementos**. Além disso, especificidades da análise de Manders sugerem que as provas do texto podem ser entendidas como aderindo a uma lógica formal diagramática. Subsequentemente isso foi confirmado no desenvolvimento de sistemas diagramáticos formais para caracterizar tal lógica. O primeiro desses sistemas foi o sistema **FG** (apresentado em MILLER, 2007), seguido do sistema

Eu (Mumma, 2010).

Esta seção é dedicada à exposição da análise de Manders e aos sistemas formais que se seguiram. Após uma breve exposição de como os diagramas de Euclides foram vistos através dos séculos, a ideia de Manders sobre o papel do diagramas nas provas geométricas é apresentada. Então se segue uma descrição de como os sistemas **FG** e **Eu** apresentam essa ideia em termos formais e como eles caracterizam a lógica dos diagramas euclidianos.

4.1. Percurso nos diagramas euclidianos do século IV a.C. até o século XX d.C.

A geometria elementar presente em **Os Elementos** fora considerada fundacional para a matemática desde a sua origem na Grécia antiga até o século XIX. Desse modo, filósofos preocupados com a natureza da matemática se viram obrigados a comentar o estatuto das provas diagramáticas presentes na obra. Um tópico central, se não o tópico central, é o **problema da generalidade**. O diagrama presente numa prova euclidiana é uma instanciação **singular** do tipo de configuração geométrica de que se trata a prova. Ainda assim, as propriedades presentes no diagrama também são tomadas como presentes em quaisquer configuração de um dado tipo. O que justifica esse salto do particular para o geral?

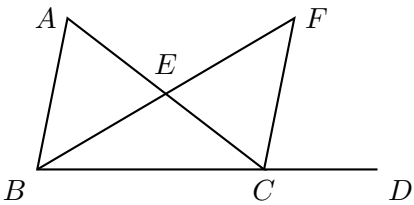
Como ilustração, considere a prova da proposição 16 do livro I de **Os Elementos**. A proposição é:

> Em um triângulo qualquer, se um de seus lados for prolongado, o ângulo exterior é maior que cada um dos ângulos interiores e opostos.

E sua prova é a seguinte:

- Seja ABC um triângulo, prolongue o lado BC até D.
- Podemos dizer que o ângulo ACD é maior que os ângulos internos e o ângulo oposto BAC.
- suponhamos que AC sofra uma bisseção em E [I, 10] e seja o lado BE prolongados por uma linha reta até F.

- Suponha que EF é igual à BE [I, 15] e conecte o lado FC.
- Então, uma vez que AE é igual à EC, e BE é igual à EF, os dois lados AE, EB são iguais aos dois lados CE, EF, respectivamente; e o ângulo AEB é igual ao ângulo FEC [I, 15]
- Logo, a base AB é igual a base FC, e o triângulo ABE é igual ao triângulo CFE [I, 4]; logo o ângulo BAE é igual ao ângulo ECF (que também é o ângulo ACF).
- Mas o ângulo ACD é maior que o ângulo ACF;
- Logo o ângulo ACD é maior que o BAE.



A prova parece fazer referência às partes do diagrama presente na prova. Contudo, ela não busca estabelecer um resultado sobre o diagrama específico, mas sim um resultado sobre todos os triângulos. Portanto, o diagrama visa representar, de algum modo, todos os triângulos.

O papel representacional dos diagramas é observado por Aristóteles no livro A, capítulo 10 dos **Analíticos Posteriores**:

> O geômetra não funda conclusões na linha em particular desenhada e descrita, mas [refere-se] ao que é ilustrado pela figura. (traduzido por T. Heath, em EUCLIDES, 1956, v.1, p. 119)

Aristóteles não enfrenta a questão do modo como os geômetras usam diagramas para raciocinar sobre o que é ilustrado. Todavia, alguns séculos mais tarde, Proclo o faz em seus comentários sobre **Os Elementos**. Ele afirma que passar de uma instância particular para uma conclusão universal é justificado porque os geômetras

> ...usam os objetos colocados no diagrama não enquanto figuras particulares, mas sim figuras similares a outras do mesmo tipo. Não é enquanto tem tal e tal comprimento que o ângulo diante de mim é bissectado, mas sendo reto e nada mais... Suponha que um dado ângulo é um ângulo reto... se não faço uso de sua retilinearidade e o considero apenas no seu caráter retilinear, a proposição se aplica para todos os ângulos com lados retilineares (**A Commentary on The First Book of Euclid's Elements**, MORROW, 1970; p. 207)

O lugar dos diagramas na geometria seguiu sendo um tópico no início da era moderna. Grandes personalidades filosóficas no século XVII e XVIII avançaram posições a esse respeito. Antecipando a posição predominante nesse período, Leibniz afirma:

> Não são as figuras que constituem a prova dos geômetras, embora seu estilo conduza a essa impressão. A força da demonstração é independente da figura presente na prova, ocorrendo apenas para facilitar o conhecimento (...) e fixar a atenção; são proposições universais, isto é, definições, axiomas e teoremas já demonstrados, que constituem a argumentação e se mostram suficientes mesmo com a exclusão das figuras (1974, p. 403 - **Novos Ensaios sobre o Entendimento Humano**)

Na introdução de seu **Tratado sobre os Princípios do Conhecimento Humano** (1710, seção 16), Berkeley reitera a questão colocada há 13 séculos atrás por Proclo sobre o problema da generalidade. Embora sempre deparemos com um triângulo em particular ao lidar com uma demonstração, não há "nem mesmo a menção" dos detalhes particulares desse triângulo. A demonstração prova, segundo Berkeley, uma proposição geral sobre triângulos.

Identificamos em Kant o mais desenvolvido e complexo tratamento dos diagramas geométricos presente no período moderno. Ele viu algo de profunda significância epistemológica no uso geométrico de diagramas em particular para raciocinar acerca dos conceitos geométricos. Ao raciocinar dessa maneira, o geômetra:

> considera o conceito *in concreto*, e embora não empiricamente, mas apenas e simplesmente como exposto *a priori*, ou seja, construído, e no qual aquilo que se segue das condições gerais da construção deva também se seguir universalmente do objeto do conceito construído(...)(1781, **Crítica da Razão Pura**, A716/B744).

Para visões díspares sobre o que essa passagem revela a respeito do lugar dos diagramas na filosofia da geometria de Kant, *vide* SHABEL, 2003, e FRIEDMAN, 2012.

No século XIX, a geometria e a matemática como um todo passaram por uma revolução. Conceitos com maior grau de abstração do que os que ocorrem em **Os Elementos** emergiram na matemática. Não apenas questões acerca do recurso aos Diagramas por Euclides perderam a sua significância, como também tal método foi mesmo tomado como matemáticamente falho. Esse ponto é expresso precisamente no trabalho inovador de Moritz Pasch, que forneceu a primeira axiomatização da geometria elementar em Pasch (1882). Nessa obra, Pasch mostrou como o tema poderia ser desenvolvido sem qualquer apelo aos diagramas ou mesmo conceitos geométricos que os diagramas instanciavam. A norma metodológica vigente nessa obra é bem expressa na seguinte passagem:

> Efetivamente, se a geometria é genuinamente dedutiva, o processo de dedução deve em todos os aspectos ser independente do **sentido** presente nos conceitos geométricos, assim como deve ser independente das figuras; apenas as **relações** estabelecidas entre os conceitos geométricos ocorrentes nas proposições (respectivamente, definições) devem ser levadas

> em conta. (PASCH, 1882, p.98; com ênfase no original. A tradução aqui se deve a SCHLIMM, 2010).

Tal norma desde então é arraigada tanto na matemática como também nas discussões filosóficas acerca da matemática. É contra esse arraigamento nas discussões filosóficas que Manders se opõe em Manders (2008)[1995]. Ele argumenta que o recurso aos diagramas na geometria antiga não evidencia uma falha dedutiva e, pelo contrário, que diagramas em conjunção com o texto formam provas matemáticas genuínamente dedutivas e rigorosas.

4.2. Distinção de Manders entre exato/co-exato e o problema da generalidade

4.2.1. A distinção entre propriedades exatas e co-exatas

Para tratar da divisão do trabalho entre o componente textual e o componente diagramático na geometria antiga, Manders (2008[1995]) distingue as propriedades de um diagrama geométrico entre **exatas** e **co-exatas**. Subjacente a essa distinção temos a noção de variação. As condições **co-exatas** realizadas por um diagrama são "as condições que não são afetadas por algum intervalo de toda variação contínua de um diagrama especificado". Condições **exatas**, por outro lado, são sensíveis a qualquer tipo de variação presente no diagrama. Grosso modo, as propriedades co-exatas de um diagrama compreendem os modos que as suas partes definem um conjunto finito de regiões no plano, como também as relações de contenção entre essas regiões. Uma relação exata proeminente é a igualdade entre duas grandezas (magnitudes) num diagrama. Por exemplo, uma mudança mínima na posição de CF no diagrama da proposição 16 faz com que os ângulos BAE e ECF se tornem diferentes.

O ponto central para Manders seria que os diagramas euclidianos contribuem para uma prova **apenas** através de suas propriedades co-exatas. Euclides nunca infere uma propriedade exata a não ser que ela se siga diretamente de uma propriedade co-exata. As relações de grandezas que não são apresentadas como relações de contenção são ou assumidas no ponto de partida ou são fornecidas no texto via uma cadeia de inferência. Isso pode ser

facilmente confirmado na prova da proposição 16: a única inferência que recorre ao diagrama é a penúltima, ou seja, que o ângulo ACD é maior que o ângulo ACF. Isso é crucialmente baseado no fato de o ângulo ACD **conter** o ângulo ACF. Há muitas outras relações presentes na prova e, embora o diagrama as instancie, elas estão devidamente justificadas no corpo do texto. E quanto às relações, os **relata** são também grandezas espacialmente separadas.

Não é difícil supor as razões de Euclides para se restringir às propriedades co-exatas, uma vez que é apenas pela capacidade de representar tais propriedades e relações que diagramas se mostram capazes de desempenhar um papel legítimo como símbolos de prova. As propriedades exatas, por sua vez, são muito específicas para serem reproduzidas e dar suporte a determinados julgamentos. De acordo com Manders,

> A prática possuí recursos para limitar o risco de desacordo sobre atribuições co-exatas (explícitas) de um diagrama; mas nela faltam esses mesmos recursos para as atribuições exatas, e portanto não poderia admití-las (...) sem se dissolver numa confusão de julgamentos insolúveis.

Seu *insight* conduz naturalmente à ideia de que os argumentos euclidianos podem ser formalizados do mesmo modo que os diagramas de Venn o foram em Shin (1994). As informações co-exatas nos diagramas de Euclides são discretas. Quando um diagrama é consultado para essas informações, o que importa é como suas linhas e circulos fazem a partição uma região plana delimitada em um conjunto finito de sub-regiões. Isso possibilita tomar os diagramas euclidianos como parte da **sintaxe** do método de prova de **Os Elementos**.

4.2.2. O problema da generalidade nas construções euclidianas

Conforme Shin (1994), realizar provas euclidianas dentro de um sistema formal equivale a especificar a sintaxe e semântica dos diagramas. Na dimensão sintática, isso significa definir precisamente os diagramas como objetos formais e fornecer regras pelas quais os diagramas figurarão nas derivações das

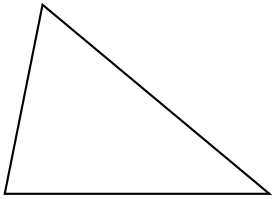
proposições euclidianas. Na semântica, equivale a especificar como as expressões derivadas serão interpretadas geometricamente, ou em outras palavras, como exatamente esses diagramas serão compreendidos como representando proposições euclidianas.

O situação da semântica dos diagramas de Euclides é, portanto, distinta da situação semântica dos diagramas de Venn. Os diagramas de Venn são empregados para provar resultados **lógicos**. As inferências feitas são neutras com relação ao conteúdo. Os diagramas euclidianos, por sua vez, serão empregados para a prova de resultados **geométricos**. As inferências são específicas de um tópico. Embora os objetos tratados na geometria euclidiana plana sejam abstratos (por exemplo, linhas geométricas não têm largura), ainda assim são espaciais. Consequentemente, questões acerca da espacialidade dos diagramas e seu escopo representacional não surgem em relação aos diagramas de Euclides como como surgiram em relação aos diagramas de Euler. No caso da geometria, a espacialidade dos diagramas conta a seu favor. Restrições espaciais do que é possível com configurações geométricas também restringem os diagramas euclidianos espaciais.

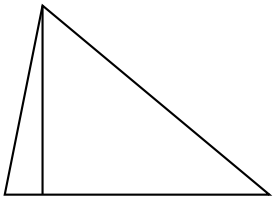
Não obstante, conforme reconhecido nos comentários filosóficos sobre a geometria euclidiana da antiguidade em diante, há questões a respeito do escopo representacional dos diagramas para as quais se deve atentar. Quais são as razões para se tomar as propriedades de um diagrama em particular como representativo de todas as configurações possíveis contidas no escopo da prova? Como um diagrama pode fornecer um resultado geral? A distinção entre propriedades exatas e co-exatas fornece uma resposta, ainda que parcial. As propriedades co-exatas são compartilhadas por todas as configurações possíveis no escopo da prova; em tais casos, portanto, parece ser justificado se basear nessas propriedades. Por exemplo, em uma prova acerca de triângulos, a variação entre as configurações no escopo da prova é uma variação de propriedades exatas - por exemplo, as medidas dos ângulos dos triângulos, as proporções entre seus lados. Todas elas compartilham das mesmas propriedades co-exatas - isto é, todas consistem em regiões circunscritas por três linhas que conjuntamente definem uma área.

Entretanto, a distinção não serve como uma resposta completa ao

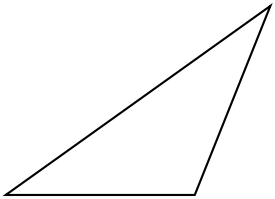
problema, pois as provas usualmente recorrem a construções sobre um certo tipo de configuração inicial. Na prova da proposição 16, por exemplo, uma construção sobre um triângulo com um lado prolongado é especificada. Nesses casos, um diagrama pode representar fielmente as propriedades co-exatas da configuração inicial. Mas não se pode assumir que o resultado de aplicar uma construção de prova ao diagrama represente as propriedades co-exatas de todas as configurações que resultam da construção. Nem precisamos considerar construções complexas a fim de ilustrar o ponto. Suponhamos uma prova que tenha como configuração inicial um **triângulo**. O diagrama



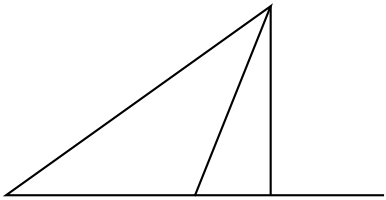
serve para representar todas as propriedades co-exatas desse tipo. Suponhamos também que o primeiro passo da prova é traçar uma semirreta perpendicular de um vértice até o lado oposto a esse vértice. O resultado desse primeiro passo no diagrama



deixa de ser representativo. O fato que uma perpendicular está contida dentro de um triângulo é uma característica co-exata. Mas há triângulos com propriedades exatas distintas do diagrama inicial em que o passo construtivo resulta em uma perpendicular fora do triângulo. Por exemplo, com o triângulo



o resultado da aplicação acima é



4.3. Os sistemas formais FG e Eu

Desse modo, realizar uma construção euclidiana num diagrama representativo pode, por vezes, resultar num diagrama não representativo. Uma tarefa central ao se formalizar as provas diagramáticas euclídianas é explicar o fato acima - ou seja, prover, através de regras, um método para distinguir as propriedades co-exatas entre gerais e não-gerais em representações diagramáticas de construções. Os sistemas **FG** e **Eu** seguem duas vias distintas para lidar com essa tarefa.

Ao se adotar o método de **FG**, temos de produzir com um diagrama **todo** caso que poderia se seguir da construção. Uma relação co-exata geral seria aquela que aparece em todos os casos. A demanda por parte de FG de que todo caso seja produzido seria pouco interessante se também não fornecesse um método para produzir todos eles. O método que **FG** oferece depende do fato de que linhas e círculos nos sistemas diagramáticos sejam definidas em termos puramente topológicos. A flexibilidade resultante torna possível formular e

implementar em um programa computacional um método para gerar casos.[47]

As linhas e círculos dos diagramas especificados em **Eu** já não são tão flexíveis quanto em **FG** e, por conseguinte, não pode lidar com o problema da generalidade por meio de uma mera análise de casos. A ideia central dessa abordagem é permitir que diagramas tenham certas informações parciais de início. Numa derivação de **Eu**, o diagrama produzido por uma construção possui um conteúdo inicial, consistindo em todas as relações qualitativas presentes no diagrama inicial. As relações qualitativas acerca dos objetos adicionadas na construção não podem ser extraídas imediatamente do diagrama, e sim devem ser derivadas a partir de certas regras do sistema.[48]

As diferenças identificadas entre as abordagens de **FG** e **Eu** na formalização das construções euclidianas podem ser entendidas como concepções distintas e gerais acerca do papel dos diagramas na matemática. **FG** incorpora a concepção na qual diagramas expõem concretamente uma miríade de possibilidades matemáticas, e seu suporte para as inferências matemáticas se dão no acesso direto a tais possibilidades. Já em **Eu** se encontra uma concepção na qual diagramas servem para representar, num único símbolo, vários componentes de uma situação matemática complexa e lidam com as inferências matemáticas permitindo que o matemático considere todos os componentes num único lugar, selecionando destes componentes os relevantes para a feitura da

[47] O nome do programa é **CDEG** (Geometria Elementar Diagramática Computadorizada) (*vide* MILLER, 2008). À medida que as linhas e círculos dos diagramas de **FG** não possuem as propriedades métricas de linhas e círculos euclidianos, as faixas das possibilidades matemáticas que os diagramas de **FG** realizam não se alinham com as faixas de possíveis configurações na geometria euclidiana. Em especial, há diagramas de **FG** cujas relações topológicas não podem ser realizadas numa configuração euclidiana. Determinar se um dado diagrama de **FG** possui essa propriedade é um problema decidível, embora seja NP-difícil (*vide* MILLER, 2006).

[48] Caso estejamos interessados apenas no modo que a informação dos diagramas se situa logicamente nas provas euclidianas segundo a abordagem oferecida por **Eu**, os diagramas não necessitam de serem formalizados como símbolos do sistema. Podemos simplesmente representar a informação extraída diretamente dos diagramas na forma sentencial. Isso foi feito na construção do sistema formal *E*, apresentado em Avigad et al. (2009), e é completo com relação às axiomatizações modernas da geometria euclidiana elementar. A inferência diagramática é capturada nesse sistema pela noção de **consequência diagramática direta**. Decidir se algo é uma relação diagramática direta em *E* é um problema de tempo polinomial.

prova.

5. Diagramas e cognição: aplicações

A despeito das limitações formais de alguns sistemas diagramáticos, como os tratados nas seções anteriores, vários sistemas são empregados numa ampla gama de contextos; ensino de lógica, raciocínio automatizado, especificação de programas computacionais, raciocinios acerca de situações na física, interfaces para programas de computador, e assim por diante. Em geral, não é bem conhecido o quão efetivo (no sentido explorado acima) são muitos desses sistemas. Oferecemos abaixo um breve panorama sobre outros sistemas diagramáticos e suas aplicações, assim como também abordaremos questões filosóficas suscitadas pelo debate acerca do estatuto do raciocínio diagramático.

5.1. Outros sistemas diagramáticos

Vale notar que vários matemáticos e filósofos propuseram sistemas diagramáticos, muitas vezes com motivações didáticas. Alguns deles, como o sistema de Lewis Carroll em "The Game of Logic" (1896), são apenas variações dos sistemas de Euler e Venn. Outros, como Frege (1879), recorrem a linhas em vez de regiões planas na construção dos diagramas (para uma descrição da notação Fregeana, *vide* a seção "Afirmações complexas e Generalidade" no verbete **Gottlob Frege**[49] na SEP, e também ENGLEBRETSEN, 1992). O sistema de Carroll supera o de Venn porque acrescenta recursos para a expressão da relação de complemento entre conjuntos e, como resultado disso, é capaz de construir inferências acerca de relações com complemento de propriedades, com o custo de representar certas propriedades como regiões disjuntas (isto é, não conectadas). Essa mudança reflete de perto a mudança na lógica da argumentação baseada em sujeito-predicado para uma representação funcional, em termos de função e argumento (STENNING, 1999).

Um dos fundadores da lógica quantificada, Peirce, também criou um sistema gráfico, nomeado Grafos Existenciais, o qual se prova como logicamente

[49] N.T.: Disponível em: https://plato.stanford.edu/entries/frege/. Acesso em: 05 jan. 2022.

equivalente com a lógica de predicados. Com os trabalhos de Don Roberts nos Grafos Existenciais e a aplicação inovadora desses Grafos por John Sowa, recentemente um grupo de pesquisadores nessa área desenvolveu várias abordagens para esse sistema representacional em contextos teóricos mais amplos (SHIN, 2003).

Guiados por questões práticas, pesquisadores de inteligência artificial (AI), com especial atenção ao poder heurístico dos sistemas em conjunção com poder expressivo, têm discutido acerca das diferentes formas representacionais por décadas (SLOMAN, 1971, 1985, 1995). Naturalmente, se interessam em discussões a respeito do papel do raciocínio visual e se engajam em simpósios interdisciplinares sobre raciocínio diagramático em conferências de IA.[50] Sob o mesmo aspecto, ao considerar que humanos adotam diferentes formas representacionais a depender do tipo de problema com o qual se lida, alguns pesquisadores na área de IA e teóricos do design se dedicam a abordagens específicas ao domínio com o propósito de fornecer formas representacionais bem adaptadas a problemas particulares.[51]

Por exemplo, Harel (1988) inventou os *Higraphs* para representar especificações de sistemas na ciência da computação. Essa ideia é adotada em aplicações na área industrial (por exemplo, UML, em BOOCH, *et al.*, 1998). Barker-Plummer e Bailin (1997) apresentam um estudo de caso no desenvolvimento de computadores capazes de realizar raciocínios analógicos do tipo que os humanos normalmente realizam na prova de certos teoremas matemáticos. Recentemente, um resultado interessante foi apresentado por Mateja Jamnik, membro do grupo de Raciocínio Matemático Alan Bundy, em Edimburgo (JAMNIK, 2001). Ele mostra como um sistema de provas formais

[50] Como exemplo de conferências, temos: Reasoning with Diagrammatic Representations: 1992 AAAI Spring Symposium; Cognitive and Computational Models of Spatial Representation: 1996 AAAI Spring Symposium; Reasoning with Diagrammatic Representations II: 1997 AAAI Fall Symposium e Formalizing Reasoning with Visual and Diagrammatic Representations: 1998 AAAI F all Symposium (*vide* também NARAYANAN, 1993).

[51] As seguintes conferências são boas evidências desse esforço: VISUAL '98: Visualization Issues in Formal Methods(Lisboa); International Roundtable Conference on Visual and Spatial Reasoning in Design (MIT, 1999) e Theories of Visual Languages - Track of VL '99: 1999 IEEE Symposium on Visual Languages.

semi-automático pode executar algumas das inferências perceptuais naturalmente feitas por humanos. Por exemplo, que a soma dos primeiros n números ímpares consiste em n ao quadrado é facilmente visualizado ao se decompor uma tabela $n \times n$ em muitos "L's" (JAMNIK *et al.*, 1999).

Pesquisadores da Universidade de Brighton realizam projetos interessantes tanto no desenvolvimento de sistemas diagramáticos quanto na sua aplicação em sistemas visuais e desenvolvimento de softwares.

Por fim, vale mencionar que cientistas como químicos e físicos também recorrem a diagramas para a realização de certas computações. Diagramas de Feynman, por exemplo, são utilizados para a execução de cálculos na física sub-atômica. Recentemente, raciocínios diagramáticos formais estão sendo desenvolvidos para a teoria quântica (COECKE; KISSINGER, 2017). Na teoria dos nós (que possuí aplicações na física, em KAUFFMAN, 1991) os três movimentos de Reidemeister são operações diagramáticas que se figuram num cálculo completo para provar equivalência de nós. De maneira não surpreendente, diagramas de nós têm despertado o interesse de pesquisadores (DE TOFFOLI; GIARDINO, 2014). O papel que diagramas e raciocínios diagramáticos desempenham na matemática abstrata da teoria das categorias também vêm sendo investigado (HALIMI, 2012; DE TOFFOLI, 2017).

5.2. Diagramas como representações mentais

Nossas representações mentais possuem como componentes entidades análogas a diagramas ou a imagens? Tal questão possui um longo histórico, em certo sentido independente, na filosofia e na psicologia. Recentemente, entretanto, filósofos têm participado do "debate sobre as imagens", uma das questões mais duradouras e controversas na psicologia, com alguns cientistas cognitivos tomando certas teorias epistemológicas na filosofia como úteis na sustentação de suas teses em relação a essa questão.

A natureza das representações mentais é uma questão perene na filosofia, e podemos facilmente rastrear discussões sobre imagens e representações mentais na era antiga.[52] Vemos como um tópico central nos

[52] *vide* **Da Alma** e **Da Memória e da Revocação** de Aristóteles.

escritos de Hobbes, Locke, Berkeley e Hume uma discussão sobre o discurso mental, o significado das palavras, imagens mentais, ideias particulares e abstratas, impressões, e assim por diante. A distinção cartesiana entre imaginar e conceber algo gerou um volume grande de discussão sobre o papel das imagens visuais nas representações mentais . O desenvolvimento das ciências cognitivas no século XX naturalmente aproximou certo grupo de filósofos e psicólogos e identificamos uma série de autores cujos trabalhos são de ambas as áreas (BLOCK, 1983; DENNETT, 1981; FODOR, 1981).

A concepção imagética baseada na introspecção foi o foco no período inicial do desenvolvimento da psicologia até o momento da predominância da abordagem behaviorista nessa disciplina. Nesse período, qualquer noção que remetia à introspecção mental, incluindo imagens, era excluído de qualquer agenda de pesquisa. Finalmente, em meados dos anos 1960, o tópico das imagens mentais ressurgiu na psicologia com uma agenda um tanto quanto humilde em relação ao que se tinha anteriormente: nem todos os tipos de representações mentais são imagéticas, e representação via imagens é um dos vários modos de tratamento da informação na mente. Além disso, graças à influência do Behaviorismo, é reconhecido que a mera introspecção não é suficiente para investigar imagens, e uma afirmação sobre imagens mentais deve ser confirmavel experimentalmente para mostrar a nossa externalização dos eventos mentais é bem sucedida. Ou seja, se o que uma certa introspecção mental diz é genuíno, então deve haver consequências experimentais observáveis desse estado mental.

Desse modo, o debate sobre as imagens entre os cientistas cognitivos versa sobre a afirmação de que nossas representações mentais são como imagens e sobre como interpretamos certos experimentos a esse respeito.[53]

Kosslyn (1980, 1994) e outros pictorialistas (SHEPARD; METZLER, 1971) apresentam dados experimentais na sustentação da afirmação de que algumas imagens mentais são mais similares a fotografias do que a uma

[53] BLOCK, 1981, é uma das melhores coletâneas de artigos sobre o debate e em BLOCK, 1983, é apresentado um resumo sucinto da controvérsia e coloca questões elucidativas acerca do debate. Nos capítulos 1-4 de TYE, 1991, temos uma boa visão geral tanto do que dizem os filósofos quanto do que dizem os cientistas cognitivos sobre a questão.

linguagem linear (como são linguagens naturais e linguagens simbólicas artificiais) em importantes aspectos, ainda que nem todas essas imagens mentais sejam exatamente do mesmo tipo. Em contraste, Pylyshin (1981) e outros descritivistas (DENNETT, 1981) colocam questões acerca do traço imagético das imagens mentais e argumentam que imagens mentais são constituídas de descrições estruturadas. Para eles, imagens mentais representam mais ao modo de uma linguagem do que uma fotografia e, consequentemente, não há imagens mentais visuais fotográficas.

Ambos os lados do debate por vezes recorrem a teorias filosóficas como um suporte. Por exemplo, pictorialistas nesse debate simpatizam com as teorias modernas dos *sense-data* e, contrariamente, os críticos de teorias do *sense-data* argumentam que o erro da concepção pictórica das imagens mentais decorre principalmente da nossa confusão acerca da linguagem ordinária e afirmam que imagens mentais são meros epifenômenos.

5.3. O papel cognitivo dos diagramas

Sem um envolvimento direto no debate sobre as imagens, alguns pesquisadores têm dado ênfase ao papel distintivo que diagramas e imagens - e em oposição às formas sentenciais - desempenham em nossas atividades cognitivas (SHIN, 2015; HAMAMI; MUMMA, 2013). Partindo da conjectura que humanos adotam representações internas, sejam diagramáticas ou espaciais, na realização de raciocínios sobre situações concretas e abstratas (conferir HOWELL, 1976; SOBER, 1976), alguns cientistas cognitivos têm se concentrado no papel das imagens e diagramas em atividades cognitivas diversas, por exemplo, na memória, imaginação, percepção, navegação, inferência, resolução de problemas, e assim por diante. Aqui, a natureza da "informação visual", obtida seja por imagens mentais internas seja por diagramas construídos externamente, tem sido um tópico crescente de pesquisa. Embora grande parte desses desenvolvimentos assumam a existência de imagens mentais (ou seja, aceitem a tese pictorialista), num sentido estrito não precisam se comprometer com a posição de que essas imagens existem como unidades básicas da nossa cognição. Descritivistas não descartam discussões acerca do papel das imagens,

mas apenas se vêem obrigados a afirmar que essas imagens não são unidades primitivas armazenadas em nossa memória, mas sim formadas a partir de descrições estruturadas de maneira similar às sentenças de uma linguagem (*vide* PYLYSHYN, 1981).

A busca pelo papel distintivo dos diagramas levou muitos pesquisadores a explorar as diferenças entre as formas diversas de representações externas e internas, e majoritariamente entre representações sentenciais e diagramáticas. Muitos resultados a esse respeito foram obtido nas ciências cognitivas. Partindo do clássico estudo de caso de Larkin e Simon (1987), que estabeleceram a diferença entre equivalência informacional e equivalência computacional em sistemas representacionais, o trabalho de Lindsay localiza o ponto em que reside a diferença computacional, denominado por ele de um método "não-dedutivo". Como brevemente apontado acima, esse processo inferencial é chamado de "carona" por Barwise e Shimojiima (1995), ou seja, o tipo de inferência na qual as conclusões parecem ser lidas automáticamente a partir da representação das premissas. Em Gurr, Lee e Stenning (1998), e em Stenning e Lemon(2001), temos uma explicação da singularidade da inferência diagramática em termos de graus de "instantaneidade" da interpretação, e é argumentado que é uma propriedade relativa, e desse modo "algumas caronas saem mais baratas do que outras". Ciente do papel dos gráficos na nossa mente, Wang e Lee (1993) apresentam um *framework* formal como um guia para linguagens visuais corretas. Neste ponto, estamos bem perto das aplicações da pesquisa em raciocínios multimodais - teoria do design e pesquisa em IA - e provendo essas disciplinas de suporte computacional para o raciocínio visual.

Também relacionado com a questão das representações mentais imagéticas é o exame da semântica de vários sistemas diagramáticos e o que nos ensinam a respeito da natureza geral das linguagens (por exemplo, GOODMAN, 1968). Por exemplo, Robert Cummins (1996) e outros argumentam que a pouca atenção dada para representações diagramáticas e o foco na noção de "representação estrutural" podem ajudar a explicar a natureza da representação ela mesma. Acreditamos que as considerações apresentadas acima nos forneçam um tratamento empírico para esse tipo de afirmação - a depender dos objetos imagéticos e das relações adotadas, padrões de inferências incorretas

devem ser detectáveis e predizíveis. Um artigo importante, embora pouco reconhecido, sobre o tema é Malinas (1991). No artigo, Malinas explora os conceitos de representação pictórica e "verdade em" uma figura via a noção de semelhança, e desse ponto considera alguns puzzles semânticos sobre representação pictórica. Ele desenvolve a "Tese Central" da representação de Peacocke (PEACOCKE, 1987), em que similaridades entre as propriedades de objetos pictóricos e seus referentes no campo visual dão origem à relação de representação. A partir disso, ele fornece uma semântica formal para figuras "análoga a uma semântica para uma linguagem ideal".

Sumário

Começamos por motivar o interesse filosófico nos diagramas por meio do seu papel no raciocínio humano e sua relação com o estudo da linguagem em geral, e no processamento de informações multimodais. Então explicamos o balanço entre poder expressivo e clareza visual nos sistemas diagramáticos ao examinar o desenvolvimento histórico dos sistemas diagramáticos, de Euler e Venn, seguindo o trabalho feito por Peirce até os desenvolvimentos recentes de Shin e Hammer. Foi argumentado que podemos oferecer o mesmo status lógico dos sistemas lineares de prova para sistemas diagramáticos. Expomos também algumas dificuldades dos sistemas diagramáticos ao examinar algumas restrições espaciais nesses sistemas e como afetam sua correção e seu poder expressivo. Fechamos a exposição oferecendo um panorama sobre outros sistemas diagramáticos, como também o interesse nos diagramas pelas áreas de ciências da computação e ciências cognitivas, e uma breve incursão no debate acerca das imagens na filosofia da mente.

Bibliografia

Referências

ALLWEIN, G; BARWISE, J (eds.) **Logical Reasoning with Diagrams**. Oxford: Oxford University Press, 1964.

AVIGAD, J; DEAN, E; MUMMA, J. "A Formal System for Euclid's Elements". **Review of Symbolic Logic**, v.2, p. 700-768, 2009.

BARKER-PLUMMER, D; BAILIN, S. "The role of Diagrams in Mathematical Proofs". **Machine GRAPHICS and VISION**, v.6(1), p. 25-56, 1997 (Special issue on Diagrammatic Representation and Reasoning)

BARKER-PLUMMER, D; BEAVER, D; VAN BENTHEM, J; SCOTTO DI LUZIO, P. **Words, Proofs and Diagrams**. Stanford: CSLI Publications, 2002.

BARWISE, J. "Heterogeneous Reasoning" em MINEAU, G; MOULIN, B; SOWA, J (eds). **ICCS 1993: Conceptual Graphs for Knowledge representation** (Lecture Notes in Artificial Intelligence: Volume 699). Berlim: Springer Verlag, p. 64-74, 1993.

BARWISE, J; ETCHEMENDY, J. "Information, Infons and Inference" em COOPER, R; MUKAI, K; PERRY, J (eds). **Situation Theory and its Applications**, volume 1. Stanford: CSLI Publications, 1989.

BARWISE, J; ETCHEMENDY, J. "Visual Information and Valid Reasoning", em ZIMMERMAN, W; CUNNINGHAM, S (eds). **Visualization in Teaching and Learning Mathematics**. Washington: Mathematical Association of America, p. 9-24, 1991.

BARWISE, J; ETCHEMENDY, J. **The Language of First Order Logic**. Stanford: CSLI Publications, 1993.

BARWISE, J; ETCHEMENDY, J. **Hyperproof**. Stanford: CSLI Publications, 1994.

BARWISE, J; ETCHEMENDY, J. "Heterogeneous Logic" em GLASGOW, J; HARI NARAYANAN, N; CHANDRASEKARAM, B (eds). **Diagrammatic Reasoning: Cognitive and Computational Perspectives**. Cambridge: AAAI Press/MIT Press, p. 209-232, 1995.

BARWISE, J; SHIMOJIMA, A. "Surrogate Reasoning". **Cognitive Studies: Bulletin of Japanese Cognitive Science Society**, v.4(2), p.7-27, 1995.

BERKELEY, G. **Principles of Human Knowledge** (1710) In: ARMSTRONG, D ed). **Berkeley's Philosophical Writings**. Londres: Macmillian, 1965. [BERKELEY, G. , **Tratado sobre os Princípios do Conhecimento Humano**, in: **Obras Completas**, São Paulo: UNESP, 2010]

BLOCK, N. (ed) **Imagery**. Cambridge: MIT Press, 1981.

BOOCH, G; RUMBAUGH, J; JACOBSON, I. **The Unified Modeling Language Reference Manual**. Reading: Addison-Wesley, 1999.

COECKE, B; KISSINGER, A. **Picturing Quantum Processes. A First Course in Quantum Theory and Diagrammatic Reasoning**. Cambridge: Cambridge University Press, 2017.

CARROLL, L. **Symbolic Logic**. Nova Iorque: Dover, 1896.

CHANDRASEKARAN, B; GLASGOW, J; HARI NARAYANAN, N (eds). **Diagrammatic Reasoning: Cognitive and Computational Perspectives**. Cambridge: AAAI Press/MIT Press, 1995.

CUMMINS, R. **Representations, Targets and Attitudes**. Cambridge: MIT Press, 1996.

DE TOFFOLI, S. "Chasing The Diagram - The Use of Visualizations in Algebraic Reasoning". **Review of Symbolic Logic**, v.10(1), p. 158-186, 2017.

DE TOFFOLI, S; GIARDINO, V. "Forms and Roles of Diagrams in Knot Theory", **Erkenntnis**, v. 79(4), p. 829-842, 2014.

DENNETT, D. "The nature of images and introspective trap" em BLOCK, N. **Imagery**. Cambridge: MIT Press, 1981, p.87-107.

ENGLEBRETSEN, G. "Linear Diagrams for Syllogisms (with Relationals)". **Notre Dame Journal of Formal Logic**. v.33(1), p. 37-69, 1992.

EUCLIDES **The Thirteen Books of the Elements**. Tradução: Sir Thomas L. Heath. 2° ed, v. 1-111, Nova Iorque: Dover Publications, 1956. [EUCLIDES, **Os Elementos**, São Paulo: UNESP, 2009]

EULER, L. **Lettres à une Princesse d' Allemagne**. São Petersburgo: l'Academie Imeriale des Sciences.

FODOR, J. "Imagistic Representation" em BLOCK, N. **Imagery**. Cambridge: MIT Press, 1981, p.63-85.

FREGE, G. **Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reine Denkers**. Halle: Louis Nebert, 1879. [FREGE, G. **Conceitografia**, São Paulo: Nau, 2019]

FRIEDMAN, M. "Kant on geometry and spatial intuition". **Synthese**. v.186, p. 231-255, 2012.

GARDNER, M. **Logic Machines and Diagrams**. Sussex: Harverster Press, 1958.

GOODMAN, N. **Languages of Art: an approach to a theory of symbols**. Londres: Oxford University Press, 1968. [GOODMAN, N. **Linguagens da Arte**, Lisboa: Gradiva, 2006]

GREAVES, M. **The Philosophical Status of Diagrams**. Stanford: CSLI Publications, 2002.

GRIGNI, M; PAPADIAS, D; PAPADIMITRIOU, C. "Topological Inference" em **International Joint Conference on Artificial Intelligence (IJCAI '95)**, Cambridge: AAAI Press, p. 901-907, 1995.

GURR, C; LEE, J; STENNING, K. "Theories of diagrammatic Reasoning: Distinguishing component problems". **Minds and Machines**, v. 8, p. 533-557, 1998.

HALIMI, B. "Diagrams as Sketches". **Synthese**, v. 186(1), p. 387-409, 2012.

HAMAMI, Y; MUMMA, J. "A Prolegomena to a Cognitive Investigation of Euclidian Diagrammatic Reasoning". **Journal of Language, Logic and Information**, v. 22(4), p. 421-448, 2013.

HAMMER, E. "Reasoning with Sentences and Diagrams". **Notre Dame Journal of Formal Logic**, v. 35(1), p. 73-87, 1995a.

HAMMER, E; SHIN, S. "Euler's Visual Logic". **History and Philosophy of Logic**, v.19, p.1-29, 1998.

HAREL, D. "On Visual Formalisms". **Communications of the ACM**, v.31(5), p.514-530, 1988.

HOWELL, R. "Ordinary Pictures, Mental Representations and logical forms". **Synthese**, v. 33, p. 149-174, 1976.

JAMNIK, M. **Mathematical Reasoning with Diagrams**. Stanford: CSLI Publications, 2001.

JAMNIK, M; BUNDY, A; GREEN, 1. "On Automating Diagrammatic Proofs of Arithmetic Arguments". **Journal of Logic, Language and Information**, v.8(3), p.297-321, 1999.

KANT, I. **Critique of Pure Reason**. Tradução: P. Guyer e A. Wood. Cambridge: Cambridge University Press, 1998(1781); [KANT, I. **Crítica da Razão Pura**, Rio de Janeiro: Vozes, 2015]

KAUFFMAN, L. **Knots and Physics**. Singapura: World Scientific, 1991.

KOSSLYN, S. **Image and Mind**. Cambridge: Harvard University Press, 1980.

KOSSLYN, S. **Image and Brain: the resolutions of the imagery debate**. Cambridge: MIT Press, 1994.

LAMBERT, J. H. **Neues Organon**. Berlim: Akademie Verlag, 1990(1764).

LARKIN, J; SIMON, H. "Why a Diagram is (Sometimes) Worth 10000 Words". **Cognitive Science**, v.11, p.65-99, 1987.

LEIBNIZ, G. **New Essays Concerning Human Understanding**. LaSalle(...): Open Court Publishing, 1949(1704).

LEMON, O. "Comparing the Efficacy of Visual Languages" em BARKER-PLUMMER, D; BEAVER, D; VAN BENTHEM, J; SCOTTO DI LUZIO, P. **Words, Proofs and Diagrams**. Stanford: CSLI Publications, 2002, p.47-69.

LEMON, O; DE RIKJE, M; SHIMOJIMA, A. "Efficacy of Diagrammatic Reasoning". **Journal of Logic, Language and Information**. v. 8(3), p.265-271, 1999.

LEMON, O; PRATT, I. "Spatial Logic and the Complexity of Diagrammatic Reasoning". **Machine Graphics and Vision**, v.6(1), p.89-108, 1997.

LEMON, O; PRATT, I. "On the insufficiency of linear diagrams for syllogisms". **Notre Dame Journal of Formal Logic**, v. 39(4), p.573-580, 1998.

MALINAS, G. "A Semantics for Pictures". **Canadian Journal of Philosophy**, v. 21(3), p.275-298, 1991.

MILLER, N. **Euclid and His Twentieth Century Rivals: Diagrams in the Logic of Euclidian Geometry**. Stanford: CLSI Publications, 2007.

MILLER, N. "Computational complexity of diagram satisfacton in Euclidian Geometry". **Journal of Complexity**, v.22, p.250-274, 2006.

MORROW, G. **Proclus: A commentary on the first book of Euclid's Elements**. Princeton: Princeton University Press, 1970.

NARAYANAN, N. "Take issue/forum: The imagery debate revisited". **Computational Intelligence**, v.9(4), p.303-435, 1993.

PASCH, M. **Vorlesunger über neuere Geometrie**. Leipzig: Teubner, 1882.

PEACOCKE, C. "Depiction". **The Philosophical Review**, v;96, p.383-410, 1987.

PEIRCE, C.S. **Collected Papers**. Cambridge: Harvard University Press, 1933.

PYLYSHYN, Z. "Imagery and Artificial Intelligence" em BLOCK, N (ed). **Readings in Philosophy of Psychology**. Cambridge: Harvard University Press, v.2,

p.170-196, 1981.

ROBERTS, D. **The Existential Graphs of Charles S. Peirce**. Haia: Mouton, 1973.

RUSSELL, B. "Vagueness" em SLATER, J (ed). **Essays on Language, Mind and Matter: 1919-26 (The Collected Papers of Bertrand Russell**. Londres: Unwin Hyman, p.145-154, 1923.

SCHLIMM, D. "Pasch's Philosophy of Mathematics". **Review of Symbolic Logic**, v.3(1), p.93-118, 2010.

SHABEL, L. **Mathematics in Kant's Critical Philosophy: Reflections on Mathematical Practice**. Nova Iorque: Routledge, 2003.

SHEPARD, R; METZLER, J. "Mental rotation of three-dimensional objects". Science, v.171, p.701-703, 1971.

SHIMOJIMA, A. **On the Efficacy of Representation**. Phd Thesis(...) - Indiana: Indiana University, 1996a

SHIMOJIMA, A. "Constraint-Preserving Representations" em MOSS, L; GINZBURG, J; DE RIJKE, M (eds). **Logic, Language and Computation: Volume 2**. Stanford: CSLI Publications, p.296-317, 1999.

SHIN, S. **The Logical Status of Diagrams**. Cambridge: Cambridge University Press, 1994.

SHIN, S. **The Iconic Logic of Peirce's Graphs**. Cambridge: MIT Press, 2003.

SHIN, S. "The Mistery of Deduction and Diagrammatic Aspects of Representation". **Review of Philosophy and Psychology: Pictorial and Spatial Representation**, v.6, p.49-67, 2015.

SLOMAN, A. "Interaction between philosophy and AI: The role of intuition and non-logical reasoning in intelligence" em **Proceedings Second International Joint Conference on Artificial Intelligence**. Los Altos: Morgan Kaufmann, 1971.

SLOMAN, A. "Why we need many knowledge representation formalisms" em BRAMER, M (ed).**Research and Development in Expert Systems**.Cambridge: Cambridge University Press, p.163-183, 1985.

SLOMAN, A. "Musings on the roles of logical and nonlogical representations in intelligence" em CHANDRASEKARAN, B; GLASGOW, J; HARI NARAYANAN, N (eds). **Diagrammatic Reasoning: Cognitive and**

Computational Perspectives. Cambridge: AAAI Press/MIT Press, 1995, p.7-32. (...)

SOBER, E. "Mental Representations". **Synthese**. v. 33, p.101-148, 1976.

SOWA, J. **Conceptual Structures: Information Processing in Minf and Machine**. Londres: Addison Wesley, 1984.

STENNING, K. "Review of **Das Spiel der Logik**, by Lewis Carrol". **Journal of Symbolic Logic**, v. 64, p.1368-1370, 1999.

STENNING, K; LEMON, O. "Aligning Logical and Psychological Perspectives on Diagrammatic Reasoning". **Artificial Intelligence Review**, v. 15(1-2), p.29-62, 2001. (reimpresso em **Thinking with Diagrams**, Kluwer, 2001)

TYE, M. **The Imagery Debate**. Cambridge: MIT Press, 1991.

VENN, J. **Symbolic Logic**. Londres: Macmillan, 1881.

WANG, D; LEE, J. "Visual Reasoning: its Formal Semantics and Applications". **Journal of Visual Languages and Computing**, v.4, p.327-356, 1993.

WITTGENSTEIN, L. **Tractatus Logico-Philosophicus**. Tradução de B. PEARS e B, MCGUINESS. Londres: Routledge, 1961(1921) [WITTGENSTEIN, L. Tractatus Logico-philosophicus, São Paulo: EDUSP, 2017]

ZEMAN, J. **The Graphical Logic of C S. Peirce**. PHD Thesis (...). Chicago: University of Chicago, 1964.

Obras Relevantes

BARWISE, J; HAMMER, E. "Diagrams and the Concept of a Logical System" em GABBAY, D (ed). **What is a Logical System?**. Nova Iorque: Oxford University Press, 1994.

HAMMER, E. **Logic and Visual Information**, Studies in Logic, Language and computation. Stanford: CSLI Publications e FoLLI, 1995b.

HAMMER, E. "Semantics for Existential Graphs". **Journal of Philosophical Logic**. v. 27, p.489-503, 1998.

HAMMER, E; SHIN, S. "Euler and the Role of Visualization in Logic" em SELIGMAN, J; WESTERSTAHL, D (...)(eds). **Logic, Language and Computation: Volume 1**. Stanford CSLI Publications, p.271-286, 1996.

KNEALE, W; KNEALE, M. **The Development of Logic**. Oxford: Clarendon Press, 1962. [KNEALE, W; KNEALE, M. **O Desenvolvimento da Lógica**, Lisboa: Fundação Calouste Gulbenkian, 1962]

LEMON, O. "Review of **Logic and Visual Information** by E. M. Hammmer". **Journal of Logic, Language and Information**, v.6(2), p.213-216, 1997.

ROBERTS, D. "The Existential Graphs of Charles S. Peirce". **Computer and Math. Applic.**, (23) (...), p.639-663, 1992.

SHIMOJIMA, A. "Operational constraints in diagrammatic reasoning" em BARWISE, J; ALLWEIN, G (eds). **Logical Reasoning with Diagrams**. Nova Iorque: Orxord University Press, p.27-48, 1996b.

SHIMOJIMA, A. "Reasoning with Diagrams and Geometrical Constraints" em SELIGMAN, J; WESTERSTAHL, D (eds)(...). **Logic, Language and Computation: Volume 1**. Stanford: CSLI Publications, p.527-540, 1996c.

SHIN, S. "A Situation-Theoretic Account of Valid Reasoning with Venn Diagrams" em BARWISE, J; GAWRON, J; PLOTKIN, G; TUTIYA, S (eds). **Situation Theory and its Applications: Volume 2**. Stanford: CSLI Publications, p.581-605, 1991.

SHIN, S. "Reconstituting Beta Graphs into a Efficacious System". **Journal of Logic, Language, and Information**, v.8, p.273-295, 1999.

SHIN, S. "Reviving the Iconicity of Beta Graphs" em ANDERSON, M; CHENG, P; HAARSLEV, V (eds). **Theory and Application of Diagrams**. Berlim: Springer-Verlag, p.58-73, 2000.

SHIN, S. **The Iconic Logic of Peirce's Graphs**. Cambridge: MIT Press, 2002a.

SHIN, S. "Multiple Readings of Peirce's Alpha Graphs" em ANDERSON, M; MEYER, B; OLIVIER, P (eds). **Diagrammatic Representation and Reasoning**. Londres: Springer-Verlag, p. 297-314, 2002b.

SOWA, J. **Knowledge Representation: Logical, Philosophical, Computational Foundations**. Belmont: Brooks/Cole, 2000.

STENNING, K. **Seeing Reason: image and language in learning to think**. Oxford: Oxford University Press, 2002.

STENNING, K; OBERLANDER, J. "A Cognitive Theory of Graphical and Linguistic Reasoning". **Cognitive science**, v.19(1), p.97-140, 1995.

TUFTE, E. **The Visual Display of Quantitative Information**. Connecticut:

Graphic Press, 1983.
TUFTE, E. **Envisioning Information**. Connecticut: Graphics Press, 1990

Teorema de Bayes*

Autoria: James Joyce
Tradução: Débora de Oliveira Silva & Sérgio R. N. Miranda
Revisão: Guilherme A Cardoso

O Teorema de Bayes é uma fórmula matemática simples utilizada para o cálculo de probabilidades condicionais. Esse teorema tem grande destaque nas abordagens **subjetivistas** ou **Bayesianas** da epistemologia, estatística e lógica indutiva. Os subjetivistas, que sustentam que a crença racional é regida pelas leis da probabilidade, apoiam-se fortemente nas probabilidades condicionais em suas teorias da evidência e modelos de aprendizagem empírica. O Teorema de Bayes é central nessas abordagens por simplificar o cálculo de probabilidades condicionais e esclarecer significativamente as características da posição subjetivista. De fato, o *insight* central do Teorema — de que uma hipótese é confirmada por qualquer conjunto de dados que a torne provável — é a base de

*JOYCE, J. "Bayes' Theorem", In: ZALTA, E. N. (ed.) **Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/bayes-theorem/. Acesso em: 05 jan. 2022.

The following is the translation of the entry on Bayes' Theorem by James Joyce in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/bayes-theorem/. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the **Stanford Encyclopedia of Philosophy**, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

toda metodologia subjetivista.

1. Probabilidades Condicionais e Teorema de Bayes

A probabilidade de uma hipótese H condicionada a um conjunto de dados E é a razão entre a probabilidade incondicional da conjunção da hipótese com os dados e a probabilidade incondicional dos dados.

(1.1) Definição:
A probabilidade de H condicionada a E é definida como $\mathbf{P}_E(H) = \mathbf{P}(H\&E)/\mathbf{P}(E)$, contanto que ambos os termos dessa razão existam e $\mathbf{P}(E) > 0$.[54]

Para ilustrar, suponha que João Ninguém é um cidadão norte-americano (escolhido aleatoriamente) vivo em 1 de janeiro de 2000. Segundo o Centro de Controle de Doenças dos Estados Unidos, dentre os 275 milhões de norte-americanos vivos naquela data, aproximadamente $2,4$ milhões morreram no decorrer do próprio ano de 2000. Já dentre os aproximadamente $16,6$ milhões de idosos (indivíduos com 75 anos ou mais), cerca de $1,36$ milhões foram a óbito. A probabilidade incondicional da hipótese de que João Ninguém morreu durante o ano 2000, H, é dada pela taxa de mortalidade da população total $\mathbf{P}(H) = 2,4\text{M}/275\text{M} = 0,00873$. Para descobrir a probabilidade da morte de João Ninguém condicionada à informação E de que ele era idoso, dividimos a probabilidade de que ele era um idoso que morreu, ou seja, $\mathbf{P}(H\&E) = 1,36\text{M}/275\text{M} = 0,00495$, pela probabilidade de que ele era idoso, ou seja, $\mathbf{P}(E) = 16,6\text{M}/275\text{M} = 0,06036$. Assim, a probabilidade da morte de João Ninguém dado que ele era uma pessoa idosa é $\mathbf{P}_E(H) = \mathbf{P}(H\&E)/\mathbf{P}(E) = 0,00495/0,06036 = 0,082$. Repare como o tamanho da população **total** ficou de fora dessa última equação, de modo que $\mathbf{P}_E(H)$ é simplesmente a proporção dos idosos que morreram. Alguém poderia

[54] Mesmo que se possa tratar as probabilidades condicionais como básicas e até dar sentido a elas quando o evento condicionante tem probabilidade zero, manteremos aqui a definição padrão. Sempre que $\mathbf{P}_E$ aparece, assume-se que a probabilidade de E é positiva. Para uma discussão e referências úteis sobre as probabilidades condicionais generalizadas, *vide*RENYI, 1955; HARPER, 1976; SPHON, 1986; HAMMOND, 1994; MCGEE, 1994; JOYCE 1999, p. 201-213.

contrastar essa quantidade, que dá a taxa de mortalidade entre os idosos, com a probabilidade "inversa" de E condicionada a H, ou seja, $\mathbf{P}_H(E) = \mathbf{P}(H\&E)/\mathbf{P}(H) = 0,00495/0,00873 = 0,57$, que consiste na proporção de mortes **na população total** que ocorreram entre os idosos.

Estas são algumas consequências diretas de (1.1):

- **Probabilidade:** $\mathbf{P}_E$ é uma função de probabilidade;[55]
- **Consequência Lógica:** Se E acarreta H, então $\mathbf{P}_E(H) = 1$;
- **Preservação de Certeza:** Se $\mathbf{P}(H) = 1$, então $\mathbf{P}_E(H) = 1$;
- **Combinação :** $\mathbf{P}(H) = \mathbf{P}(E)\mathbf{P}_E(H) + \mathbf{P}(\sim E)\mathbf{P}_{\sim E}(H)$.[56]

O fato mais importante sobre as probabilidades condicionais é sem dúvida o **Teorema de Bayes**, cuja importância foi primeiramente compreendida pelo clérigo britânico Thomas Bayes em sua obra-prima publicada postumamente intitulada **An Essay Toward Solving Problem in the Doctrine of Chances** (BAYES, 1764). O Teorema de Bayes relaciona a probabilidade "direta" de uma hipótese condicionada a um conjunto de dados, $\mathbf{P}_E(H)$, com a probabilidade "inversa" dos dados condicionada à hipótese, $\mathbf{P}_H(E)$.

(1.2) Teorema de Bayes:
$\mathbf{P}_E(H) = [\mathbf{P}(H)/\mathbf{P}(E)]\mathbf{P}_H(E)$.

Por conta de uma infeliz escolha de terminologia (agora inevitável), os estatísticos se referem à probabilidade inversa $\mathbf{P}_H(E)$ como a *likelihood* de H

[55] $\mathbf{P}_E$ deve ser uma função real, limitada entre 0 e 1, que satisfaz:
Normalização: $\mathbf{P}_E(\mathbf{T}) = 1$, em que **T** é uma verdade lógica qualquer;
Aditividade Contável: $\mathbf{P}_E(X) = \sum_i \mathbf{P}_E(X_i)$ quando $\{X_1, X_2, X_3...\}$ é qualquer conjunto de pares de proposições incompatíveis cujas disjunções formam X.

[56] De modo geral, se $\{E_1, E_2, E_3...\}$ é uma **partição** contável de proposições de evidência, a combinação acarreta que $\mathbf{P}(H) = \sum_i \mathbf{P}(E_i)\mathbf{P}_{E^i}(H)$. Pode-se considerar E_i como um conjunto de resultados de algum experimento que são mutuamente exclusivos e coletivamente exaustivos. A combinação então diz que a probabilidade incondicionada de H é a **expectativa** da probabilidade de H condicionada aos resultados do experimento.

dado E. [57] Ela exprime o grau com que a hipótese **prevê** os dados a partir da informação de fundo codificada na probabilidade **P**.

No exemplo previamente discutido, a condição de que João Ninguém morreu durante o ano de 2000 é um forte previsor da sua condição de idoso. De fato, a equação $\mathbf{P}_H(E) = 0,57$ nos informa que 57% do total de mortes daquele ano ocorreram entre os idosos. O teorema de Bayes permite-nos usar essa informação para calcular a probabilidade "direta" da morte de João Ninguém dado que ele era idoso. O cálculo é feito multiplicando-se o "termo previsor" $\mathbf{P}_H(E)$ pela divisão entre o número total de mortes na população e o número de idosos ($\mathbf{P}(H)/\mathbf{P}(E) = 2,4\text{M}/16,6\text{M} = 0,144$). O resultado é o esperado: $\mathbf{P}_E(H) = 0,57 \times 0,144 = 0,082$.

Embora seja uma trivialidade matemática, o Teorema de Bayes é de grande valia no cálculo de probabilidades condicionais, pois as probabilidades inversas são tipicamente mais fáceis de serem determinadas e menos subjetivas do que as probabilidades diretas. Aqueles que têm opiniões diferentes sobre as probabilidades incondicionais de E e H frequentemente discordam acerca do valor de E como um indicador de H. Todavia, eles podem concordar sobre o grau com que a hipótese prevê os dados caso conheçam algum dos seguintes fatos intersubjetivamente disponíveis: (a) a probabilidade **objetiva** de E dado H, (b) a frequência com que eventos como E ocorrerão se H for verdadeira ou (c) o fato de que H acarreta logicamente E. Os cientistas comumente elaboram os experimentos para que as *likelihoods* possam ser conhecidas em alguma dessas maneiras "objetivas". Assim, o Teorema de Bayes garante que qualquer disputa sobre o significado dos resultados experimentais possa remontar aos desacordos "subjetivos" acerca das probabilidades incondicionais de H e E.

Quando tanto $\mathbf{P}_H(E)$ como $\mathbf{P}_{\sim H}(E)$ são conhecidas, quem realiza um experimento sequer precisa saber da probabilidade de E para determinar o valor de $\mathbf{P}_E(H)$ ao usar o Teorema de Bayes.

[57] N.T.: A tradução de dicionário de "*likelihood*" é "probabilidade", mas, obviamente, a frase "a *likelihood* de H dado E" não diz respeito à probabilidade de H dado E, mas sim à inversa: E dado H. Alguns optam por traduzir "*likelihood*" por "verossimilhança", mas decidimos manter a expressão do original em inglês. Ambas as opções são comuns.

(1.3) Teorema de Bayes (2ª forma):[58]

$$\mathbf{P}_E(H) = \mathbf{P}(H)\mathbf{P}_H(E)/[\mathbf{P}(H)\mathbf{P}_H(E) + \mathbf{P}(\sim H)\mathbf{P}_{\sim H}(E)]$$

Sob essa forma, o Teorema de Bayes é particularmente útil para inferirmos as causas a partir dos seus efeitos, uma vez que é razoavelmente fácil distinguir a probabilidade de um efeito dada a presença ou a ausência de uma suposta causa. Por exemplo, os médicos costumam rastrear doenças com incidência conhecida usando testes de diagnóstico com **sensitividade** e **especificidade** reconhecidas. A sensitividade do teste, i.e., sua taxa de "verdadeiro positivo", é a fração das vezes em que os pacientes com a doença testam positivo. A especificidade do teste, i.e., sua taxa de "verdadeiro negativo", é a proporção com que os pacientes saudáveis testam negativo. Se H for o evento em que um dado paciente tem a doença e E for o evento em que o paciente testa positivo para a doença, então a sensitividade e a especificidade do teste são dadas, respectivamente, pelas *likelihoods* $\mathbf{P}_H(E)$ e $\mathbf{P}_{\sim H}(\sim E)$, e a incidência inicial da doença na população é $\mathbf{P}(H)$. Dados esses *inputs* sobre os efeitos da doença no resultado do teste, pode-se usar (1.3) para determinar a probabilidade da doença dado um teste positivo. Para uma ilustração mais detalhada desse processo, *vide* **Exemplo 1** do complemento deste capítulo.

2. Formas Especiais do Teorema de Bayes

O Teorema de Bayes pode ser expresso nas mais variadas formas que são úteis para diferentes propósitos. Uma dessas versões emprega o que Rudolf Carnap chamou de **quociente de relevância** ou **razão de probabilidade** [*Probability Ratio* (PR)] (CARNAP, 1962, p. 466). Trata-se do fator $\mathbf{PR}(H, E) = \mathbf{P}_E(H)/\mathbf{P}(H)$ pelo qual a probabilidade incondicional de *H* deve ser multiplicada para se obter a probabilidade de H condicionada a E. O Teorema de Bayes equivale então a um princípio simples de simetria para razões de probabilidade.

---

[58] Se $H_1, H_2, H_3, ..., H_n$ é uma partição na qual cada uma das probabilidades inversas $\mathbf{P}_{H_i}(E)$ é conhecida, então pode-se expressar a probabilidade direta como $\mathbf{P}_E(H_i) = \mathbf{P}(H_i)\mathbf{P}_{H_i}(E)/[\sum_j \mathbf{P}(H_j)\mathbf{P}_{H_j}(E)]$.

(1.4) Regra da Razão de Probabilidade [*Probability Ratio*(PR)]:
PR$(H, E) =$ **PR**(E, H)

O termo da direita fornece uma medida do grau com que H **prevê** E**.** Se considerarmos que **P**(E) expressa a previsibilidade inicial de E dada a informação de fundo codificada em **P** e que $P_H(E)$ expressa a previsibilidade de E quando H é adicionada a essa informação de fundo, então **PR**(E, H) registra o grau com que o conhecimento de H torna E mais ou menos previsível com relação à base inicial: **PR**$(E, H) = 0$ significa que H prevê categoricamente que $\sim E$; **PR**$(E, H) = 1$ significa que adicionar H não altera em nada a previsão da base inicial; **PR**$(E, H) = 1/$**P**(E) significa que H prevê categoricamente que E. Uma vez que **P**$(E) =$ **P**$_{\mathbf{T}}(E)$, no qual **T** é uma verdade lógica qualquer, podemos considerar que (1.4) nos diz o seguinte:

> *A probabilidade de uma hipótese condicionada a um conjunto de dados é igual à probabilidade incondicional da hipótese multiplicada pelo grau com que a hipótese supera uma tautologia como um previsor dos dados.*

No exemplo do João Ninguém, **PR**(H, E) é obtida pela comparação entre a previsibilidade da condição de idoso dado que João Ninguém morreu em 2000 e a previsibilidade dessa mesma condição quando não é fornecida qualquer informação sobre a mortalidade do indivíduo. Ao dividirmos o primeiro "termo previsor" pelo segundo obtemos **PR**$(H, E) =$ **P**$_H(E)/$**P**$(E) = 0,57/0,06036 = 9,44$. Assim, como um previsor da condição de idoso em 2000, saber que João Ninguém morreu é mais de nove vezes melhor do que não saber se ele viveu ou morreu.

Outra forma útil do Teorema de Bayes é a **Regra de Chances** [*Odds*]. No jargão dos apostadores, a "chance" de uma hipótese é a sua probabilidade dividida pela probabilidade da sua negação: **O**$(H) =$ **P**$(H)/$**P**$(\sim H)$. Por exemplo, uma corrida de cavalo cujas chances de ganhar uma corrida em particular sejam de 7 para 5 tem $7/12$ de chances de ganho e $5/12$ de chances de perda. Algo que ajuda a entender a diferença entre chances e probabilidades é pensar nas probabilidades como **frações** da distância entre a probabilidade de uma contradição e a de uma tautologia, de modo que **P**$(H) = p$ significa que H

é p vezes tão provável de ser verdadeiro quanto uma tautologia. Por outro lado, escrever $\mathbf{O}(H) = [\mathbf{P}(H) - \mathbf{P}(\mathbf{F})]/[\mathbf{P}(\mathbf{T}) - \mathbf{P}(H)]$ (em que ***F*** é alguma contradição lógica e ***T*** uma tautologia) deixa claro que $\mathbf{O}(H)$ expressa essa mesma quantidade como a razão entre o montante com que a probabilidade de H excede a de uma contradição e o montante com que a probabilidade de H é excedida pela probabilidade de uma tautologia. Desse modo, a diferença entre "falar de probabilidades" e a "falar de chances" corresponde à diferença entre dizer que "estamos a dois terços de distância do destino" e "já percorrremos o dobro do que ainda resta para chegarmos ao destino".

O análogo da razão de probabilidade é a **razão de chances** [*Odds Ratio* (OR)] $\mathbf{OR}(H, E) = \mathbf{O}_E(H)/\mathbf{O}(H)$, o fator pelo qual a chance incondicional de H deve ser multiplicada para se obter a chance de H condicionada a E. O Teorema de Bayes então equivale ao seguinte fato acerca da razão de chances:

(1.5) Regra da razão de chances [*Odds Ratio* (OR)]:
$\mathbf{OR}(H, E) = \mathbf{P}_H(E)/\mathbf{P}_{\sim H}(E)$

Repare a similaridade entre as regras (1.4) e (1.5). Mesmo que cada uma empregue um modo diferente para **expressar** probabilidades, cada uma mostra como a **sua** expressão para a probabilidade de H condicionada a E pode ser obtida pela multiplicação da **sua** expressão para a probabilidade incondicional de H por um fator que envolve probabilidades inversas.

A quantidade $\mathbf{LR}(H, E) = \mathbf{P}_H(E)/\mathbf{P}_{\sim H}(E)$ que aparece em (1.5) é a **razão de *likelihood*** [*Likelihood Ratio* (LR)] de H dado E. Em cenários de teste como o descrito no Exemplo 1, a razão de *likelihood* é a taxa de verdadeiro positivo dividida pela taxa de falso positivo: $\mathbf{LR} = \text{sensitividade}/(1 - \text{especificidade})$. Assim como no caso da razão de probabilidade, podemos construir a razão de *likelihood* como uma medida do grau com que *H* prevê *E*. Ao invés de compararmos a probabilidade de E dado H com a probabilidade incondicional de E, iremos compará-la com a probabilidade de E condicionada a $\sim H$. $\mathbf{LR}(H, E)$ é, portanto, o grau com que a hipótese supera a sua negação como uma previsora dos dados. Mais uma vez, o Teorema de Bayes nos diz como fatorar as probabilidades condicionais em probabilidades incondicionais e medidas de poder de previsão.

A chance de uma hipótese condicionada a um conjunto de dados é igual à chance incondicional da hipótese multiplicada pelo grau com que essa hipótese supera a sua negação como uma previsora dos dados.

No exemplo de João Ninguém, **LR**(H, E) é obtida pela comparação entre a previsibilidade da condição de idoso dado que João Ninguém morreu em 2000 e a previsibilidade dessa mesma condição dado que viveu durante todo aquele ano. Ao dividirmos o primeiro "termo previsor" pelo segundo obtemos **LR**(H, E) $=$ **P**$_H(E)/$**P**$_{\sim H}(E) = 0,57/0,056 = 10,12$. Assim, como um previsor da condição de idoso no ano de 2000, saber que João Ninguém morreu é mais de dez vezes melhor do que saber que viveu.

As similaridades entre as versões da "razão de probabilidade" e da "razão de chances" do Teorema de Bayes podem ser ainda mais desenvolvidas se expressarmos a probabilidade de H como um múltiplo da probabilidade de alguma outra hipótese H^* usando a **função da probabilidade relativa** **B**$(H, H^*) =$ **P**$(H)/$**P**(H^*). Que fique claro que **B** generaliza tanto **P** quanto **O**, pois **P**$(H) =$ **B**$(H,$ **T**$)$ e **O**$(H) =$ **B**$(H, \sim H)$. Pela comparação dos valores condicionais e incondicionais de **B** obtemos o **Fator de Bayes:**

$$\begin{aligned}\textbf{BR}(H, H^*; E) &= \textbf{B}_E(H, H^*)/\textbf{B}(H, H^*) \\ &= [\textbf{P}_E(H)/\textbf{P}_E(H^*)]/[\textbf{P}(H)/\textbf{P}(H^*)]\end{aligned}$$

Também podemos generalizar a razão de *likelihood* ao definirmos **LR**$(H, H^*; E) =$ **P**$_H(E)/$**P**$_{H^*}(E)$. Essa fórmula compara a previsibilidade de E com base em H com a previsibilidade de E com base em H^*. Podemos usar essas duas quantidades para formular uma forma ainda mais geral do Teorema de Bayes.

(1.6) Teorema de Bayes (Forma Geral):
BR$(H, H^*; E) =$ **LR**$(H, H^*; E)$.

A mensagem contida em (1.6) é a seguinte:

> *As razões de probabilidade para duas hipóteses condicionadas a um conjunto de dados é igual à razão das probabilidades incondicionais dessas hipóteses multiplicadas pelo grau com que a primeira hipótese supera a segunda como uma previsora dos dados.*

As diversas versões do Teorema de Bayes diferem apenas no que diz respeito às funções usadas para expressar as probabilidades incondicionais ($\mathbf{P}(H)$, $\mathbf{O}(H)$, $\mathbf{B}(H)$) e ao termo de *likelihood* usado para representar o poder de previsão ($\mathbf{PR}(E, H)$, $\mathbf{LR}(H, E)$, $\mathbf{LR}(H, H^*, E)$). Contudo, em cada caso a mensagem subjacente é a mesma:

> *probabilidade condicional = probabilidade incondicional x poder de previsão*

(1.2)-(1.6) são formas multiplicativas do Teorema de Bayes que fazem uso da divisão para comparar as disparidades entre as probabilidades incondicionais e condicionais. Às vezes essas comparações são melhor expressas aditivamente ao substituirmos as razões por **diferenças**. A tabela a seguir contém o análogo aditivo de cada medida de razão.

Tabela 1

Razão	Diferença
Razão de Probabilidade: $\mathbf{PR}(H, E) = \mathbf{P}_E(H)/\mathbf{P}(H)$	**Diferença de Probabilidade:** $\mathbf{PD}(H, E) = \mathbf{P}_E(H) - \mathbf{P}(H)$
Razão de Chances (*Odds*): $\mathbf{OR}(H, E) = \mathbf{O}_E(H)/\mathbf{O}(H)$	**Diferença de Chances(*Odds*):** $\mathbf{OD}(H, E) = \mathbf{O}_E(H) - \mathbf{O}(H)$
Fator de Bayes: $\mathbf{BR}(H, H^*; E) = \mathbf{B}_E(H, H^*)/\mathbf{B}(H, H^*)$	**Diferença de Bayes:** $\mathbf{BD}(H, H^*; E) = \mathbf{B}_E(H, H^*) - \mathbf{B}(H, H^*)$

Podemos usar o teorema de Bayes para obter os análogos aditivos de (1.4)-(1.6), que são exibidos na tabela a seguir juntamente com suas contrapartes multiplicativas.

Tabela 2

	Razão	Diferença
(1.4)	$\mathbf{PR}(H, E) = \mathbf{PR}(E, H) = \mathbf{P}_H(E)/\mathbf{P}(E)$	$\mathbf{PD}(H, E) = \mathbf{P}(H)[\mathbf{PR}(E, H) - 1]$
(1.5)	$\mathbf{OR}(H, E) = \mathbf{LR}(H, E) = \mathbf{P}_H(E)/\mathbf{P}_{\sim H}(E)$	$\mathbf{OD}(H, E) = \mathbf{O}(H)[\mathbf{OR}(H, E) - 1]$
(1.6)	$\mathbf{BR}(H, H^*; E) = \mathbf{LR}(H, H^*; E) = \mathbf{P}_H(E)/\mathbf{P}_{H^*}(E)$	$\mathbf{BD}(H, H^*; E) = \mathbf{B}(H, H^*)[\mathbf{BR}(H, H^*; E) - 1]$

Repare como cada medida aditiva é obtida multiplicando-se a probabilidade incondicional de H, expressa na escala relevante, **P**, **O** ou **B**, pela medida multiplicativa associada subtraída por 1.

Os resultados dessa seção são úteis para quem utiliza o cálculo de probabilidades, mas eles têm uma importância especial para as abordagens **subjetivistas** ou "Bayesianas" da estatítica, epistemologia e inferência indutiva.[59]. Os subjetivistas apoiam-se fortemente nas probabilidades condicionais em sua teoria do apoio evidencial e na sua explicação da aprendizagem empírica. Dado que o Teorema de Bayes é o fato mais importante sobre as probabilidades condicionais, não é de modo algum surpreendente que ele receba grande destaque na metodologia subjetivista.

3. O Papel do Teorema de Bayes nas Explicações Subjetivistas da Evidência

Os subjetivistas mantêm que as crenças têm graus variados de força e que as crenças graduadas de um sujeito idealmente racional podem ser representadas por uma **função de probabilidade subjetiva P**. Para cada hipótese H sobre a qual o sujeito forma uma opinião, $\mathbf{P}(H)$ mede o seu nível de confiança (ou "grau de crença") na verdade de H.[60] As crenças condicionais são representadas por probabilidades condicionais, de modo que $\mathbf{P}_E(H)$ mede a confiança do sujeito em

[59] Para uma excelente discussão geral do subjetivismo, *vide* o verbete **Interpretations of Probability** [https://plato.stanford.edu/archives/win2021/entries/probability-interpret/] na SEP

[60] Quando as opiniões do sujeito sobre *H* não são suficientemente definidas para serem mensuradas por um único número, seu estado de crença é representado por uma família de funções de probabilidade. Para discussões úteis sobre estados indeterminados de crença, *vide* LEVI, 1985; JEFFREY, 1987; KAPLAN, 1996, p. 23-31.

H **dada a suposição de que** E **é um fato**.[61]

Uma das características mais influentes do programa subjetivista é a sua explicação do **apoio evidencial**. As ideias que norteiam a **Teoria Bayesiana da Confirmação** são as seguintes:

- **Relatividade Confirmacional:** As relações de evidência devem ser relativizadas aos indivíduos e seus graus de crença;
- **Proporcionalismo de Evidência:**[62] Um crente racional regulará a sua confiança em uma hipótese H por sua **evidência total** para H, de modo que sua probabilidade subjetiva para H reflita o saldo geral das suas razões a favor ou contra a verdade de H;
- **Confirmação incremental:**[63] Um conjunto de dados fornece evidência **incremental** para H na medida em que o condicionamento a esses dados aumenta a probabilidade de H.

O primeiro princípio diz que os enunciados sobre relações de evidência sempre farão referência implícita aos sujeitos e seus graus de crença, de modo que, e.g., o enunciado "E é evidência para H" deveria ser realmente lido como "E é evidência para H relativamente à informação codificada na probabilidade subjetiva **P**".

Segundo o proporcionalismo de evidência, o nível de confiança de um sujeito em H deveria variar diretamente com a força da sua evidência a favor da

[61] Pode-se ter uma probabilidade subjetiva determinada para H condicionada a E mesmo quando não se tem probabilidades determinadas para H & E e E. A evidência estatística frequentemente justifica atribuições de probabilidade condicional sem fornecer qualquer informação sobre as probabilidades incondicionais subjacentes. Por exemplo, estudos cuidadosos de tabelas atuariais podem convencer-me de que as minhas chances de viver além dos oitenta anos dado que eu tenha um ataque cardíaco severo aos cinquenta estão entre $0,04$ e $0,02$, mas a mesma tabela não me dá qualquer informação sobre as chances de que eu vá sofrer um ataque cardíaco severo aos cinquenta e, portanto, sobre as minhas chances incondicionais de viver até os oitenta.

[62] Tomei esse termo emprestado de Alvin Goldman, que argumenta **contra** tal perspectiva (GOLDMAN, 1986, p. 89-93). Embora nem todos os Bayesianos aceitem o proporcionalismo de evidência, a explicação da evidência incremental como uma mudança na probabilidade subjetiva só faz realmente sentido se supomos que o nível de confiança do sujeito em uma proposição varia diretamente com a força da sua evidência para a verdade dessa proposição.

[63] A distinção entre evidência total e incremental é essencialmente a mesma que a distinção de Carnap entre "confirmação como firmeza" e "confirmação como aumento de firmeza" (CARNAP, 1962, novo prefácio. Compare MAHER, 1996, p. 162).

verdade de H. Da mesma maneira, seu nível de confiança em H condicionada a E deveria variar diretamente com a força da sua evidência a favor da verdade de H quando essa evidência é aumentada pela suposição de E. É difícil precisar o que constitui a evidência de um sujeito[64] e explicar como suas crenças deveriam ser "proporcionais" a essa evidência. Contudo, a ideia de que a evidência incremental é refletida em disparidades entre as probabilidades condicionais e incondicionais só faz sentido se as diferenças na probabilidade subjetiva espelharem as diferenças na evidência **total**.

Uma unidade de dados fornece ao sujeito evidência **incremental** a favor ou contra uma hipótese na medida em que o recebimento dos dados aumenta ou diminui sua evidência total para a verdade da hipótese. Quando as probabilidades mensuram a evidência total, o incremento da evidência que E fornece a H é uma questão de disparidade entre $\mathbf{P}_E(H)$ e $\mathbf{P}(H)$. Já quando as chances são usadas é uma questão de disparidade entre $\mathbf{O}_E(H)$ e $\mathbf{O}(H)$. Confira o **Exemplo 2** do complemento deste capítulo, que ilustra a diferença entre a evidência incremental e a total e explica a "falácia de taxa base", que pode surgir quando elas não são distinguidas adequadamente.

[64] Em uma concepção puramente subjetivista, a evidência total de um sujeito para uma hipótese é derivada das suas próprias perspectivas "subjetivas" sobre as plausibilidades intrínsecas das proposições e da informação que ele adquire via aprendizagem. O sujeito inicia com uma probabilidade subjetiva $\mathbf{P}_0$ que abrange seus julgamentos "prévios" sobre as plausibilidades das proposições (ou, mais acuradamente, seus preconceitos epistêmicos iniciais). Subsequentemente, ele revisa essas probabilidades à luz da experiência, incorporando assim novas informações em seu sistema doxástico. Suas probabilidades subjetivas a qualquer momento são então o resultado do aumento das suas opiniões "prévias" sobre as plausibilidades intuitivas das proposições com a informação adquirida via aprendizagem. Embora alguns subjetivistas falem como se cada pessoa iniciasse sua vida epistêmica com um tipo de prévias primitivas que registrariam o estado das suas opiniões antes que qualquer informação empírica viesse à tona, esse é um modo enganoso de colocar as coisas. Falar de probabilidades "prévias" e "posteriores" só faz sentido com relação a uma sequência futura específica de experiências de aprendizagem. A "prévia" não é nada mais do que a função de probabilidade que reflete as crenças do sujeito, independentemente de como foram alcançadas, antes do início da aprendizagem. Como Elliot Sober coloca, "a probabilidade prévia é propriamente chamada assim não porque é *a priori* (e não é), mas porque ela vigora antes que novas evidências sejam levadas em consideração". (SOBER, 2002, p. 24). Alguns probabilistas, menos inclinados a abordagens subjetivas, adotam concepções menos permissivas de evidência. Por exemplo, *vide* WILLIAMSON, 2000, p. 184-208 e MAHER, 1996.

Será útil distinguir dois conceitos relacionados com a evidência total:

- A **evidência líquida a favor de** H é o grau com que a evidência total do sujeito a favor de H excede a sua evidência total a favor de $\sim H$.
- O **saldo da evidência total de** H **sobre** H^* é o grau com que a evidência total do sujeito a favor de H excede a sua evidência total a favor de H^*.

O conteúdo preciso dessas noções dependerá de como a evidência total é entendida e mensurada, além de como as disparidades na evidência total são caracterizadas. Por exemplo, se a evidência total é dada em termos de probabilidades e as disparidades são tratadas como razões, então a evidência líquida para H é $\mathbf{P}(H)/\mathbf{P}(\sim H)$. Já se a evidência total é expressada em termos de chances e as diferenças são usadas para expressar disparidades, então a evidência líquida para H será $\mathbf{O}(H) - \mathbf{O}(\sim H)$. Para uma lista completa das possibilidades, *vide* **Tabela 3** (no complemento deste capítulo).

Como essas observações deixam claro, pode-se interpretar $\mathbf{O}(H)$ como uma medida da evidência líquida ou uma medida da evidência total. Para ver a diferença, imagine que 750 bolas vermelhas e 250 bolas pretas foram retiradas ao acaso (e depois repostas por outras) de uma urna que continha 10.000 bolas vermelhas e pretas. Assumindo que essa é a nossa única evidência sobre o conteúdo da urna, torna-se razoável estabelecer que $\mathbf{P}(\textbf{Vermelha}) = 0,75$ e $\mathbf{P}(\sim \textbf{Vermelha}) = 0,25$. Em uma leitura da probabilidade como evidência total, essas atribuições de valor refletem o fato de que temos muitas evidências a favor da cor **Vermelha** (nomeadamente, que 750 das 1000 bolas retiradas eram vermelhas) e o fato de que também temos algumas evidências contra isso (nomeadamente, que 250 das bolas retiradas eram pretas). A evidência **líquida** para a cor **Vermelha** é então a disparidade entre a nossa evidência total para a cor **Vermelha** e a nossa evidência total contra a cor **Vermelha**. Isso pode ser expresso de maneira multiplicativa se dissermos que vimos a cor vermelha sendo retirada três vezes mais do que a cor preta, o que é simplesmente dizer que $\mathbf{O}(\textbf{Vermelha}) = 3$. Alternativamente, podemos usar $\mathbf{O}(\textbf{Vermelha})$ como uma medida da evidência total se tomarmos nossa evidência total para a cor **Vermelha** como a razão entre as bolas retiradas das cores vermelha e preta ao invés do número total de retiradas da cor vermelha, além da nossa evidência para a cor

~Vermelha como sendo a razão entre as bolas pretas e vermelhas ao invés do número total de retiradas da cor preta. Embora a decisão acerca de se usar **O** como uma medida da evidência total ou líquida faça pouca diferença para as questões sobre o montante **absoluto** da evidência total para uma hipótese (uma vez que $\mathbf{O}(H)$ é uma função crescente de $\mathbf{P}(H)$), essa decisão pode fazer uma grande diferença quando se está considerando as **mudanças** incrementais na evidência total que são provocadas pelo condicionamento a novas informações.

Os filósofos interessados em caracterizar padrões corretos de raciocínio indutivo e em fornecer "reconstruções racionais" da metodologia científica tenderam a se concentrar na evidência incremental como sendo crucial para os seus empreendimentos. Quando os cientistas (ou pessoas comuns) dizem que E suporta ou confirma H, o que eles geralmente estão querendo dizer é que saber da verdade de E aumentará o montante da evidência total para a verdade de H. Uma vez que os subjetivistas caracterizam a evidência total em termos de probabilidades subjetivas ou chances, eles analisam a evidência incremental em termos das mudanças nessas quantidades. Nessas perspectivas, a maneira mais simples de caracterizar a força da evidência incremental é fazer comparações ordinais das probabilidades condicionais e incondicionais ou das chances.

(2.1) Um Tratamento Comparativo da Evidência incremental:
Relativamente a uma função de probabilidade subjetiva **P**,

- E confirma incrementalmente (infirma, é irrelevante para) H se e somente se $\mathbf{P}_E(H)$ é maior que (menor que, igual a) $\mathbf{P}(H)$;
- H recebe um incremento maior (ou decréscimo menor) de suporte evidencial de E do que de E^* se e somente se $\mathbf{P}_E(H)$ excede $\mathbf{P}_{E^*}(H)$.

Essas equivalências ainda valem mesmo quando as probabilidades são substituídas por chances. Assim, essa parte da teoria subjetivista da evidência não depende de como a evidência total é mensurada.

O Teorema de Bayes ajuda a esclarecer o conteúdo de (2.1) ao tornar claro que o *status* de E como evidência incremental para H aumenta na medida em que H prevê E. Essa observação constitui a base para as seguintes conclusões sobre a confirmação incremental (que valem enquanto $1 > \mathbf{P}(H), \mathbf{P}(E) > 0$):

(2.1a) Se E confirma incrementalmente H, então H confirma incrementalmente E.

(2.1b) Se E confirma incrementalmente H, então E infirma incrementalmente $\sim H$.

(2.1c) Se H implica E, então E confirma incrementalmente H.

(2.1d) Se $\mathbf{P}_H(E) = \mathbf{P}_H(E^*)$, então H recebe mais apoio incremental de E do que de E^* se e somente se E é incondicionalmente menos provável do que E^*.

(2.1e) **Princípio Fraco de *likelihood*:** Efornece evidência incremental para H se e somente se $\mathbf{P}_H(E) > \mathbf{P}_{\sim H}(E)$. De modo geral, se $\mathbf{P}_H(E) > \mathbf{P}_{H^*}(E)$ e $\mathbf{P}_{\sim H}(\sim E) \geq \mathbf{P}_{\sim H^*}(\sim E)$, então E fornece mais evidência incremental para H do que a H^*.

(2.1a) nos diz que a confirmação incremental é uma questão de **reforço mútuo**: um sujeito que vê E como evidência para H investe mais confiança na possibilidade de que ambas as proposições sejam verdadeiras do que em qualquer outra possibilidade em que apenas uma delas é verdadeira.

(2.1b) diz que a evidência relevante tem de ser capaz de discriminar entre a verdade e a falsidade da hipótese sob teste.

(2.1c) fornece uma razão subjetivista para o **modelo hipótetico dedutivo de confirmação**. Segundo esse modelo, as hipóteses são confirmadas de modo incremental por qualquer evidência que elas acarretem. Embora os subjetivistas rejeitem a ideia de que as relações de evidência possam ser caracterizadas de uma maneira independente da crença — a confirmação Bayesiana é **sempre** relativizada ao sujeito e às suas probabilidades subjetivas —, eles tentam preservar a ideia básica do modelo H-D ressaltando que as hipóteses são apoiadas incrementalmente pelas evidências que elas acarretam **para qualquer pessoa que já não tenha se decidido sobre a hipótese ou a evidência**. Mais precisamente, se H acarreta E, então $\mathbf{P}_E(H) = \mathbf{P}(H)/\mathbf{P}(E)$, que excede $P(H)$ sempre que $1 > \mathbf{P}(E)$, $\mathbf{P}(H) > 0$. Isso explica por que os cientistas frequentemente elaboram experimentos que se encaixam no paradigma H-D. Mesmo quando as relações de evidência são relativizadas às probabilidades subjetivas, os experimentos em que a hipótese sob teste implica os dados serão considerados como evidencialmente

relevantes por **qualquer pessoa** que ainda não tenha se decidido sobre a hipótese ou os dados. O **grau** de confirmação incremental vai variar entre os sujeitos a depender dos níveis anteriores de confiança que depositam em H e E, mas todos concordarão que os dados apoiam incrementalmente a hipótese em algum grau, pelo menos.

Os subjetivistas recorrem a (2.1d) para explicar por que os cientistas tão frequentemente consideram que as evidências improváveis ou surpreendentes têm mais potencial confirmatório do que as evidências previamente conhecidas. Embora não seja **geralmente** verdadeiro que as evidências improváveis tenham mais potencial de confirmação, é verdade que o poder de confirmação incremental de E relativamente a H varia inversamente à probabilidade incondicionada de E **quando o valor da probabilidade inversa** $\mathbf{P}_H(E)$ **permanece fixo**. Se H acarreta tanto E quanto E^*, então o Teorema de Bayes acarreta que o menos provável dentre os dois apoiará H mais fortemente. Por exemplo, mesmo se ataques cardíacos forem invariavelmente acompanhados por fortes dores no peito e falta de ar, o primeiro sintoma é uma evidência muito melhor para um ataque cardíaco do que o segundo, pois fortes dores no peito são muito menos comuns do que falta de ar.

(2.1e) registra uma mensagem central do Teorema de Bayes para as teorias da confirmação. Diremos que H é **uniformamente melhor** do que H^* como um previsor do valor de verdade de E quando (a) H prevê E mais fortemente do que H^* e (b) $\sim H$ prevê $\sim E$ mais fortemente do que $\sim H^*$. Segundo o princípio fraco de *likelihood*, as hipóteses que são, uniformemente, melhores previsoras dos dados são mais bem apoiadas por esses dados. Por exemplo, o fato de que Joãozinho é cristão é uma evidência melhor para pensarmos que seus pais são cristãos ao invés de que são hindus, pois (a) uma proporção muito maior de pais cristãos têm filhos cristãos em comparação com os pais hindus e (b) uma proporção muito maior de pais que não são cristãos têm filhos não cristãos em comparação com os pais que não são hindus.

O Teorema de Bayes também pode ser usado como base para o desenvolvimento e avaliação de medidas **quantitativas** de apoio evidencial. Os resultados listados na Tabela 2 acarretam que todas as quatro funções **PR**, **OR**, **PD** e **OD** concordam entre si no que diz respeito à questão mais simples da

confirmação: E fornece evidência incremental para H?

(2.2) Corolário.
Cada uma das formulações que se seguem são equivalentes à asserção de que E fornece evidência incremental a favor de H: **PR**$(H, E) > 1$, **OR**$(H, E) > 1$, **PD**$(H, E) > 0$ e **OD**$(H, E) > 0$.

Assim, todas as quadro medidas concordam com a explicação comparativa de evidência incremental presente em (2.1).

Por conta de todo esse acordo não deveria ser surpreendente que **PR**(H, E), **OR**(H, E) e **PD**(H, E) tenham sido propostas como medidas do **grau** de apoio incremental que E fornece a H.[65] Apesar de **OD**(H, E) não ter sido sugerida para esse propósito, iremos considerá-la por razões de simetria. Alguns autores sustentam que alguma dessas funções em particular é a única medida correta da evidência incremental; já outros pensam que é melhor utilizar uma variedade de medidas que registram diferentes relações de evidência. Embora aqui não seja o lugar para discutir tais problemas, podemos recorrer ao Teorema de Bayes para entender o que essas diversas funções mensuram e também para caracterizar as relações formais entre elas.

Todas as quatro medidas concordam nas suas conclusões sobre o montante **comparativo** de evidência incremental que diferentes unidades de dados fornecem a uma hipótese **fixa**. Em particular, as funções concordam ordinariamente acerca dos seguintes conceitos derivados da evidência incremental:

[65] Para um levantamento completo da literatura nessa área com comentários penetrantes, *vide* FITELSON, 2001, capítulos 1 e 2. Outras funções têm sido propostas como medidas de apoio da evidência, tais como $\mathbf{P}(H\&E) - \mathbf{P}(H)\mathbf{P}(E)$ (*vide* CARNAP, 1962, p. 360), $\mathbf{P}_H(E) - \mathbf{P}_{\sim H}(E)$ (*vide* NOZICK, 1981, p. 252) e $\mathbf{P}_H(E) - \mathbf{P}(E)$ (*vide* HALINA, 1988), além de $P_E(H) - P_{\sim E}(H)$ (*vide* JOYCE, 1999; CHRISTENSEN, 1999). Como Fitelson mostra, nenhuma dessas medidas satisfaz a segunda cláusula de (2.1), o que significa que nenhuma delas capta a noção de evidência **incremental** que buscamos.

- O **incremento efetivo de evidência**[66] que E fornece a H é o montante com que a evidência incremental que E fornece a H excede a evidência incremental que $\sim E$ fornece a H.
- O **diferencial** na evidência incremental que E e E^* fornecem a H é o montante com que a evidência incremental que E fornece a H excede a evidência incremental que E^* fornece a H.

A evidência efetiva é uma questão do grau que a evidência total de um sujeito para H depende da sua opinião sobre E. Quando $\mathbf{P}_E(H)$ e $\mathbf{P}_{\sim E}(H)$ (ou $\mathbf{O}_E(H)$ e $\mathbf{O}_{\sim E}(H)$) diferem muito, a crença do sujeito em E afeta demasiadamente sua crença em H: do seu ponto de vista, muita coisa depende do valor de verdade de E quando consideramos as questões sobre o valor de verdade de H. Um grande diferencial na evidência incremental entre E e E^* nos diz que o conhecimento de E aumenta a evidência total do sujeito para H em uma quantidade muito maior do que o conhecimento de E^* o faz. Os leitores podem consultar a **Tabela 4** (no complemento deste capítulo) para verem as medidas quantitativas de evidência efetiva e diferencial.

A segunda claúsula de (2.1) afirma que E fornece mais evidência incremental para H do que E^* somente quando a probabilidade de H condicionada a E excede a probabilidade de H condicionada a E^*. Agora trata-se de um passo simples mostrar que todas as quatro medidas de apoio incremental concordam ordinariamente quanto a questões de evidência efetiva e diferenciais na evidência incremental.

(2.3) Corolário:

Os aspectos seguintes são equivalentes para todo H, E^* e E com probabilidade positiva:

- E fornece mais evidência incremental para H do que E^*
- $\mathbf{PR}(H, E) > \mathbf{PR}(H, E^*)$
- $\mathbf{OR}(H, E) > \mathbf{OR}(H, E^*)$

[66] Embora o termo "evidência efetiva" não seja padrão, a ideia de que a disparidade entre $\mathbf{P}_E(H)$ e $\mathbf{P}_{\sim E}(H)$ registra uma relação de evidência importante é defendida em (JOYCE, 1999, p. 203-213) e (CHRISTENSEN, 1999). Ambos os autores argumentam que essa medida auxilia os Bayesianos a contornarem o chamado problema da "velha evidência" descrito em (GLYMOUR, 1980).

– **PD**$(H, E) >$ **PD**(H, E^*)
– **OD**$(H, E) >$ **OD**(H, E^*)

As quatro medidas de apoio incremental podem discordar quanto ao grau **comparativo** com que cada unidade particular dos dados confirma duas hipóteses distintas. Os **Exemplos 3**, **4** e **5** (no complemento deste capítulo) mostram as várias maneiras pelas quais isso pode acontecer.

Em última análise, todas as diferenças entre as medidas têm a ver com (a) se a evidência **total** a favor de uma hipótese deveria ser mensurada em termos de probabilidades ou em termos de chances e (b) se as **disparidades** na evidência total são melhor registradas como razões ou diferenças. As linhas na tabela a seguir correspondem às diferentes medidas da evidência total. As colunas correspondem aos diferentes modos de tratamento das disparidades.

Tabela 5: Quatro medidas de evidência incremental

	Razão	**Diferença**
P $=$ Total	**PR**$(H, E) =$ **P**$_E(H)/$**P**(H)	**PD**$(H, E) =$ **P**$_E(H) -$ **P**(H)
O $=$ Total	**OR**$(H, E) =$ **O**$_E(H)/$**O**(H)	**OD**$(H, E) =$ **O**$_E(H) -$ **O**(H)

Tabelas similares podem ser elaboradas para as medidas de evidência líquida e as medidas de saldo da evidência total (*vide* **Tabela 5A** no complemento deste capítulo).

Podemos usar várias formas do Teorema de Bayes para esclarecer as similaridades e diferenças entre essas medidas reescrevendo-as em termos de razões de *likelihood*.

Tabela 6: As quatro medidas expressas em termos de razões de *likelihood*

	Razão	Diferença
$\mathbf{P} = \text{Total}$	$\mathbf{PR}(H, E) = \mathbf{LR}(H, \mathbf{T}; E)$	$\mathbf{PD}(H, E) = \mathbf{P}(H)[\mathbf{LR}(H, \mathbf{T}; E) - 1]$
$\mathbf{O} = \text{Total}$	$\mathbf{OR}(H, E) = \mathbf{LR}(H, \sim H; E)$	$\mathbf{OD}(H, E) = \mathbf{O}(H)[\mathbf{LR}(H, \sim H; E) - 1]$

Essa tabela mostra que há duas diferenças entre cada medida multiplicativa e sua contraparte aditiva. Em primeiro lugar, o termo da *likelihood* que aparece em uma dada medida multiplicativa é subtraído por 1 na medida aditiva associada a ela. Em segundo, em cada medida aditiva o termo da *likelihood* subtraído é multiplicado por uma expressão para a probabilidade de H: $\mathbf{P}(H)$ ou $\mathbf{O}(H)$, conforme seja o caso. A primeira distinção não faz diferença; ela deve-se unicamente ao fato de que as medidas multiplicativas e aditivas empregam um ponto zero diferente para medirem a evidência. Se estabelecermos o ponto de independência probabilística $\mathbf{P}_E(H) = \mathbf{P}(H)$ como um zero natural comum e subtraírmos 1 de cada medida multiplicativa,[67] termos de *likelihood* equivalentes aparecerão em ambas as colunas.

A diferença real entre as medidas de uma dada linha dizem respeito ao efeito das probabilidades incondicionais nas relações de confirmação incremental. Na coluna da direita, o grau que E fornece evidência incremental para H é diretamente proporcional à probabilidade de H expressada em unidades de $\mathbf{P}(\mathbf{T})$ ou $\mathbf{P}(\sim H)$. Na coluna da esquerda, a probabilidade de H não faz diferença para o montante de evidência incremental que E fornece a H, uma vez que $\mathbf{P}_H(E)$, $\mathbf{P}(E)$ ou $\mathbf{P}_{\sim H}(E)$ são fixados.[68]

À luz do Teorema de Bayes, a diferença entre as medidas de razão e

[67] Uma alternativa seria considerar o **logaritmo** de cada medida. No entanto, as escalas logarítmicas são bem diferentes das escalas aditivas, pois elas expressam as quantidades em termos de múltiplos de uma base comum ao invés de distâncias de um zero comum. Assim, em uma escala logarítmica, distâncias iguais representam **razões** iguais de aumento de evidência em vez de **incrementos** iguais de evidência. Para comparar coisas similares, devemos expressar as medidas multiplicativas em uma escala aditiva (ou as medidas aditivas em uma escala multiplicativa).

[68] Considere, por exemplo, duas probabilidades, **P** e **Q**, relacionadas pela seguinte transformação:

$$\mathbf{Q}(X) = h\mathbf{P}_H(X) + [\mathbf{P}(E) - h\mathbf{P}_H(E)]\mathbf{P}_{\sim H\&E}(X) + [\mathbf{P}(\sim E) - h\mathbf{P}_H(\sim E)]\mathbf{P}_{\sim H\&\sim E}(X)$$

medidas de diferença se resume a uma questão:

> *Um conjunto de dados fornece um maior incremento de apoio evidencial para uma hipótese mais provável do que para uma hipótese menos provável quando ambas as hipóteses predizem os dados igualmente bem?*

A medida de diferença dá uma resposta positiva para essa questão, a medida de razão dá uma resposta negativa.

O Teorema de Bayes também pode nos ajudar a entender a diferença entre as linhas. As medidas presentes em uma dada linha concordam sobre o papel da **previsibilidade** na confirmação incremental. Na linha de cima a evidência incremental que E fornece a H aumenta linearmente com $\mathbf{P}_H(E)/\mathbf{P}(E)$, enquanto na linha inferior ela aumenta linearmente com $\mathbf{P}_H(E)/\mathbf{P}_{\sim H}(E)$. Desse modo, o que importa quando as probabilidades mensuram a evidência total é o grau com que H excede **T** como um previsor de E, mas o que importa quando as chances mensuram a evidência total é o grau com que H excede $\sim H$ como um previsor de E.

Aqui, o problema central consiste no status da razão de *likelihood*. Enquanto todos concordam que a razão de *likelihood* deveria ter um papel central em qualquer teoria quantitativa da evidência, há visões conflitantes sobre qual é exatamente a relação de evidência que ela registra. Há três interpretações possíveis:

---

na qual $1/\mathbf{PR}(H, E) > \mathbf{Q}(H) = h \geq 0$. O leitor pode verificar que as razões de probabilidade de **Q** e **P** são as mesmas, mas que a diferença da probabilidade-**Q** é $\mathbf{Q}(H)/\mathbf{P}(H)$ vezes a diferença da probabilidade-**P**. Similarmente, se **P** e **Q** são relacionados por

$$\mathbf{Q}(X) = h\mathbf{P}_H(X) + (1-h)\mathbf{P}_{\sim H}(X)$$

em que $1 > h > 0$, então as razões de *likelihood* são as mesmas, embora suas diferenças de chance sejam um fator de $[\mathbf{Q}(H)/\mathbf{Q}(\sim H)]/[\mathbf{P}(H)/\mathbf{P}(\sim H)]$ em separado.

Tabela 7: Três interpretações da razão de *likelihood*

Leitura da Probabilidade como evidência total	• **PR**(H, E) mede mudança incremental na evidência total. • **LR**(H, E) mede mudança incremental na evidência líquida. • **LR**(H, H^*, E) mede a mudança incremental no saldo de evidência que E fornece a H sobre H^*.
Leitura das Chances como evidência total	• **LR**(H, E) mede mudanças incrementais na evidência total. • **LR**$(H, E)^2$ mede mudança incremental na evidência líquida. • **LR**(H, H^*, E)/**LR**$(\sim H, \sim H^*, E)$ mede a mudança incremental no saldo de evidência que E fornece a H sobre H^*.
Leitura Likelihoodista	• Nem **P** nem **O** mensuram evidência total porque as relações de evidência são essencialmente **comparativas**; elas sempre envolvem o saldo de evidência. • **LR**(H, E) mede o saldo de evidência que E fornece a H sobre H^*. • **LR**(H, H^*, E) mede o saldo de evidência que E fornece a H sobre H^*.

Na primeira leitura não há qualquer conflito entre usar as razões de probabilidade ou as razões de *likelihood* para medir evidências. Uma vez que tenhamos clareza sobre as distinções entre evidência total, evidência líquida e o saldo de evidência veremos que cada **PR**(H, E), **LR**(H, E) e **LR**(H, H^*, E) mede uma relação de evidência importante, embora as relações que elas mensurem sejam importantemente diferentes.

Quando as chances mensuram a evidência total, nem **PR**(H, E) ou **LR**(H, H^*, E) cumprem um papel fundamental na teoria da evidência. As

mudanças na razão de probabilidade para H dado E apenas indicam mudanças na evidência incremental dada a presença de informação sobre as mudanças na razão de probabilidade para $\sim H$ dado E. Da mesma forma, mudanças na razão de *likelihood* para H e H^* dado E apenas indicam mudanças no balanço da evidência à luz de informação sobre mudanças na razão de *likelihood* para $\sim H$ e $\sim H$ dado E. Assim, enquanto cada uma das duas funções pode figurar como componente em uma medida significativa da confirmação, nenhuma delas nos diz algo sobre a evidência incremental quando tomadas isoladamente.

A terceira explicação, o "likelihoodismo", é popular entre os estatísticos não Bayesianos. Seus proponentes negam o proporcionalismo da evidência. Eles sustentam que a probabilidade subjetiva que um sujeito atribui a uma hipótese reflete meramente o seu grau de incerteza sobre a verdade dessa hipótese; tal probabilidade não precisa estar vinculada de modo algum ao montante de evidência que ele tem a seu favor.[69] São as razões de *likelihood*, ao invés das probabilidades subjetivas, que registram o "significado científico" das relações de evidência. Vejamos dois enunciados clássicos dessa posição:

> Toda a informação que os dados fornecem quanto aos méritos relativos de duas hipóteses está contido na razão de *likelihood* da hipótese sobre os dados. (EDWARDS, 1972, p. 30).
>
> O "significado evidencial" dos resultados experimentais é totalmente caracterizado pela função de *likelihood* (…) De modo geral, os relatórios de resultados experimentais em jornais científicos devem ser descrições de funções de *likelihood*. (BRINBAUM, 1962, p. 272).

De acordo com essa explicação, tudo o que pode ser dito sobre a importância evidencial de E para H está incorporado na seguinte generalização do princípio fraco de *likelihood*:

> **"Lei de *likelihood*":** Se H implica que a probabilidade de E é **x** e H^* implica que a

[69] Para um enunciado claro dessa posição, *vide* ROYALL, 1997, p. 8-11).

> probabilidade de E é **x***, então E é uma evidência que suporta H em relação a H^* se e somente se **x** excede **x*** e a razão de *likelihood,* **x/x***, mensura a força desse apoio. (HACKING, 1965, p. 106-109), (ROYALL, 1997, p. 3).

O bioestatístico Richard Royall é um defensor particulamente lúcido do likelihoodismo (ROYALL, 1997). Ele sustenta que qualquer conceito de evidência cientificamente respeitável deve analisar o impacto evidencial de E sobre H somente em termos de *likelihoods*. Ela não deve dizer respeito às probabilidades incondicionais de E ou H porque as *likelihoods* são mais conhecidas e objetivas do que as probabilidades incondicionais. Royall argumenta intensamente contra a ideia de que a evidência incremental possa ser medida em termos de disparidade entre as probabilidades condicionais e incondicionais. O ponto central da sua queixa é o seguinte:

> Enquanto $[\mathbf{LR}(H, H^*; E)]$ mede o apoio para uma hipótese H relativamente a uma alternativa específica H^* sem considerar as probabilidades prévias das duas hipóteses ou quais outras hipóteses também poderiam ser examinadas, a lei da mudança de probabilidade [como mensurada por $\mathbf{PR}(H, E)$] mede o apoio a H relativamente a uma distribição prévia específica entre H e as hipóteses a ela alternativas (...) A lei da mudança de probabilidade tem utilidade limitada nos discursos científicos pela sua dependência da distribuição da probabilidade prévia, a qual é geralmente desconhecida e/ou pessoal. Embora concordemos (com base na lei da *likelihood*) que a evidência dada apoia H em relação a H^*, e H^{**} em relação a embos H e H^*, podemos discordar sobre se essa é uma evidência que apoia H (com base na lei da mudança de probabilidade) nos baseando apenas em nossos diferentes julgamentos das probabilidades prévias de H, H^*

> e H^{**}. (ROYALL, 1977, p. 10-11, com variações na notação).

O ponto em questão é que nem a razão de probabilidade ou a diferença de probabilidade registrarão o tipo de evidência objetiva requerida pela ciência porque os seus valores dependem de termos "subjetivos" $\mathbf{P}(E)$ e $\mathbf{P}(H)$ e não apenas das *likelihoods* "objetivas" $\mathbf{P}_H(E)$ e $\mathbf{P}_{\sim H}(E)$.

Concordar com essa avaliação será uma questão de temperamento filosófico, particularmente da disposição para tolerar as probabilidades subjetivas em uma explicação das relações de evidência. Isso também dependerá crucialmente de quão se está convencido de que as *likelihoods* são melhor conhecidas e mais objetivas do que as probabilidades subjetivas ordinárias. Casos como o que é previsto na lei de *likelihood*, nos quais as hipóteses **acarretam dedutivamente** uma probabilidade definida para os dados, são relativamente raros. Portanto, a menos que se esteja disposta a adotar uma teoria da evidência com um campo bastante restrito de aplicação, muito dependerá de quão fácil é determinar as *likelihoods* objetivas em situações em que a conexão de previsão entre a hipótese e os dados é ela mesma o resultado de inferências **indutivas**. Contudo, seja como for que se decidam essas questões, não há como negar que as razões de *likelihood* terão um papel central em qualquer explicação probabilista da evidência.

De fato, o princípio fraco da probabilidade (2.1e) contém uma forma mínima de Bayesianismo com o qual todas as partes podem concordar. Isso é mais claro ainda quando recolocado em termos de *likelihoods*:

(2.1e) O Princípio Fraco de *likelihood* (expresso em termos de razões de *likelihood*):
Se $\mathbf{LR}(H, H^*; E) \geq 1$ e $\mathbf{LR}(\sim H, \sim H^*; \sim E) \geq 1$, com uma desigualdade estrita, então E fornece mais evidência incremental para H do que a H^* e $\sim E$ fornece mais evidência incremental para $\sim H$ do que a $\sim H^*$.

Adeptos do "likelihoodismo" vão endossar (2.1e) porque as relações descritas no seu antecedente dependem apenas de probabilidades inversas. Os

proponentes das interpretações de "probabilidade" e de "chance" da evidência total aceitarão (2.1e) porque a satisfação do seu antecedente garante que o condicionamento a E aumenta a probabilidade de H e as suas chances estritamente mais do que as de H^*. Certamente, o princípio fraco de *likelihood* deve ser uma parte integrante de qualquer explicação de relevância da evidência que merece o título de "Bayesianismo". Negar isso é não ter compreendido ainda a principal mensagem do Teorema de Bayes para as questões de evidência: nomedamente, as hipóteses são confirmadas pelos dados que elas prevêem. Como veremos na próxima seção, a forma "mínima" do Bayesianismo aparece de modo importante nos modelos subjetivistas de aprendizagem a partir da experiência.

4. O Papel do Teorema de Bayes nos Modelos Subjetivistas de Aprendizagem

Os subjetivistas entendem a aprendizagem como um processo de **revisão de crenças** no qual a probabilidade subjetiva "prévia" **P** é substituída por uma probabilidade posterior **Q** que incorpora informações recém-adquiridas. Esse processo procede em dois estágios. No primeiro, algumas das probabilidades do sujeito são **diretamente alteradas** pela experiência, intuição, memória ou algum outro processo de aprendizado **não inferencial**. No segundo, o sujeito "atualiza" o restante das suas opiniões para conciliá-las com o seu conhecimento recém-adquirido.

Muitos subjetivistas consideram as crenças iniciais como *sui generis* e independentes do estado de opinião prévio do crente. No entanto, na medida em que a primeira fase do processo de aprendizagem é entendida como não inferencial, o subjetivismo pode ser compatibilizado com uma epistemologia "externalista" que permite a crítica às mudanças de crença em termos da confiabilidade dos processos causais que as geram. Além disso, também é capaz de acomodar o pensamento de que o efeito direto da experiência pode depender causalmente da probabilidade prévia do crente.

Os subjetivistas estudaram detalhadamente a segunda fase do processo de aprendizagem. Aqui, as mudanças imediatas de crença são vistas como impondo restrições da forma "a probabilidade posterior **Q** tem tais e tais

propriedades". O objetivo é descobrir quais tipos de restrições a experiência tende a impor e explicar como as opiniões **prévias** do sujeito podem ser usadas para justificar a escolha de uma probabilidade posterior dentre muitas outras que poderiam satisfazer uma dada restrição. Os subjetivistas abordam o último problema assumindo que o agente está justificado em adotar qualquer probabilidade posterior elegível que se **afaste minimamente** das suas opiniões prévias. Esse é um requisito do tipo "sem conclusões precipitadas". Isso é aqui explicado como um resultado natural da ideia de que aprendizes racionais deveriam regular as suas crenças de acordo com a força da evidência que eles coletam.

As experiências de aprendizagem mais simples são aquelas nas quais o aprendiz se torna certo da verdade de alguma proposição E sobre a qual previamente estava incerto. A restrição aqui é que todas as hipóteses inconsistentes com E devem receber probabilidade zero. Os subjetivistas modelam esse tipo de aprendizagem como **condicionalização simples**, o processo no qual a probabilidade prévia de cada proposição H é substituída pela posterior que coincide com a probabilidade prévia de H condicionada a E.

(3.1) Condicionalização Simples

Se uma pessoa com uma "prévia" tal que $0 < \mathbf{P}(E) < 1$ tem uma experiência de aprendizagem cujo único efeito imediato é o de aumentar sua probabilidade subjetiva a favor de E para 1, então, pós-aprendizagem, a sua "posterior" para qualquer proposição H deveria ser $\mathbf{Q}(H) = \mathbf{P}_E(H)$.

Resumidamente, um crente racional que aprende com certeza que E é verdadeiro deveria introduzir essa informação no seu sistema doxástico por meio da sua condicionalização.

Embora seja útil como um ideal, a simples condicionalização não tem muita aplicação porque ela requer que o aprendiz esteja absolutamente **certo** da verdade de E. Como Richard Jeffrey argumentou (JEFFREY, 1987), a evidência que recebemos é quase sempre muito vaga ou ambígua para justificar tal "dogmatismo". Em modelos mais realísticos, o efeito direto de uma experiência de aprendizagem será o de **alterar** a probabilidade subjetiva de alguma proposição

sem elevá-la a 1 ou abaixá-la a 0. As experiências desse tipo são apropriadamente modeladas pelo o que veio a ser chamado de "**condicionalização de Jeffrey**" (apesar de Jeffrey preferir o termo "cinemática de probabilidade").

(3.2) Condicionalização de Jeffrey

Se uma pessoa com uma prévia tal que $0 < \mathbf{P}(E) < 1$ tem uma experiência de aprendizagem cujo único efeito imediato é o de mudar sua probabilidade subjetiva a favor de E para q, então, pós-aprendizagem, a sua "posterior" para qualquer H deveria ser $\mathbf{Q}(H) =_q \mathbf{P}_E(H) + (1-q)\mathbf{P}_{\sim E}(H)$.

Obviamente a condicionalização de Jeffrey se reduz à condicionalização simples quando $q = 1$.

Uma variedade de argumentos a favor da condicionalização (simples ou ao modo de Jeffrey) pode ser encontrada na literatura, mas aqui não podemos explorar esses argumentos.[70] No entanto, há um tipo de justificação no qual o Teorema de Bayes tem destaque. Essa justificação explora as conexões entre a revisão de crença e a noção de evidência incremental para mostrar que a condicionalização é a **única** regra de revisão que permite aos aprendizes regularem corretamente as suas crenças posteriores à nova evidência que eles recebem.

A chave para esse argumento repousa na junção da versão "mínima" do Bayesianismo expressa em (2.1e) com um requerimento de "proporcionalização" muito modesto para as regras de revisão de crença.

(3.3) O Princípio Fraco de Evidência

Se, relativamente à prévia **P**, E fornece pelo menos tanta evidência incremental para H quanto a H^* e se H é anteriormente mais provável que H^*, então H deveria permanecer mais provável do que H^* após qualquer experiência de aprendizagem cujo efeito imediato é o de aumentar a probabilidade de E.

[70] *vide*, por exemplo, (TELLER, 1976), (ARMENDT, 1980), (SKYRMS, 1987) and (VAN FRAASSEN, 1999).

Isso requer um agente que conserve seus pontos de vista sobre a probabilidade relativa das duas hipóteses quando adquire evidência que apoia mais fortemente a hipótese mais provável. Obviamente, isso exclui revisões de crenças irracionais como esta: George está mais confiante de que o New York Yankees vencerá o campeonato do que está confiante de que o Boston Red Sox vencerá o campeonato, mas ele inverte a opinião quando vem a saber (apenas) que os Yankees derrotaram os Red Sox no jogo da noite passada.

Combinar (3.3) com o Bayesianismo mínimo produz o seguinte:

(3.4) Consequência
Se a prévia de uma pessoa é tal que $\mathbf{LR}(H, H^*; E) \geq 1$, $\mathbf{LR}(\sim H, \sim H^*; \sim E) \geq 1$ e $\mathbf{P}(H) > \mathbf{P}(H^*)$, então qualquer experiência de aprendizagem cujo efeito imediato seja aumentar sua probabilidade subjetiva a favor de *E* deveria resultar em uma posterior tal que $\mathbf{Q}(H) > \mathbf{Q}(H^*)$.

Na suposição razoável de que **Q** é definido no mesmo conjunto de proposições sobre os quais **P** é definido, essa condição é suficiente para escolher a condicionalização simples como o **único** método correto de revisão de crença para experiências de aprendizagem que tornam E indubitável. Ela seleciona a condicionalização de Jeffrey como o **único** método correto quando a aprendizagem meramente altera a probabilidade subjetiva de alguém para E. O argumento para essas conclusões utiliza os dois fatos seguintes sobre probabilidades:

(3.5) Lema
Se H e H^* implicam E quando $\mathbf{P}(H) > \mathbf{P}(H^*)$, então $\mathbf{LR}(H, H^*; E) = 1$ e $\mathbf{LR}(\sim H, \sim H^*; \sim E) > 1$.
[*vide* o **esboço da prova** no complemento deste capítulo]

(3.6) Lema
A condicionalização simples a E é a única regra para a revisão de probabilidades subjetivas que produz uma posterior com as propriedades seguintes para **qualquer** prévia tais que $\mathbf{P}(E) > 0$:

i. $\mathbf{Q}(E) = 1$;

ii. **Similaridade Ordinal:**
Se H e H^* implicam E, então $\mathbf{P}(H) \geq \mathbf{P}(H^*)$ se e somente se $\mathbf{Q}(H) \geq \mathbf{Q}(H^*)$.

[*vide* o **esboço da prova** no complemento deste capítulo]

A partir daqui o argumento para a condicionalização simples é uma questão de utilizar (3.4) e (3.5) para estabelecer a similaridade ordinal. Suponha que H e H^* implicam E e que $\mathbf{P}(H) > \mathbf{P}(H^*)$. Segue-se de (3.5) que $\mathbf{LR}(H, H*; E) = 1$ e $\mathbf{LR}(\sim H, \sim H*; \sim E) > 1$. (3.4) então implica que qualquer experiência de aprendizagem que aumente a probabilidade de E deve resultar em uma posterior com $\mathbf{Q}(H) > \mathbf{Q}(H^*)$. Assim, **Q** e **P** são ordinalmente similares quanto às hipóteses que implicam H. Se continuarmos a supor que as experiências de aprendizagem aumentam a probabilidade de E para 1, então (3.6) garante que **Q** surge de **P** pela condicionalização simples a E.

O caso da condicionalização de Jeffrey é igualmente direto. Uma vez que o argumento da similaridade ordinal não depende totalmente da suposição de que $\mathbf{Q}(E) = 1$, estabelecemos realmente que:

(3.7) Corolário

- Se H e H^* implicam E, então $\mathbf{P}(H) > \mathbf{P}(H^*)$ se e somente se $\mathbf{Q}(H) > \mathbf{Q}(H^*)$;
- Se H e H^* implicam $\sim E$, então $\mathbf{P}(H) > \mathbf{P}(H^*)$ se e somente se $\mathbf{Q}(H) > \mathbf{Q}(H^*)$.

Por conseguinte, **Q** é ordinalmente similar a **P** quando restrito a hipóteses que implicam E ou $\sim E$. Ademais, já que a divisão por números positivos não atrapalha as relações ordinais, também se segue que $\mathbf{Q}_E$ é ordinalmente similar a **P** quando restrito a hipóteses que implicam E e $\mathbf{Q}_{\sim E}$ é ordinalmente similar a **P** quando restrito a hipóteses que implicam $\sim E$. Dado que $\mathbf{Q}_E(E) = 1 = \mathbf{Q}_{\sim E}(E)$, (3.6) implica:

(3.8) Consequência
Para cada proposição H, $\mathbf{Q}_E(H) = \mathbf{P}_E(H)$ e $\mathbf{Q}_{\sim E}(H) = \mathbf{P}_{\sim E}(H)$

É fácil mostrar que (3.8) é necessário e suficiente para que **Q** surja de **P** pela condicionalização de Jeffrey a E. Sujeito à restrição $\mathbf{Q}(E) = q$, garante-se que $\mathbf{Q}(H) =_q \mathbf{P}_E(H) + (1-q)\mathbf{P}_{\sim E}(H)$.

A moral geral é clara:

> *O insight Bayesiano básico presente no princípio fraco de likelihood (2.1e) implica que a condicionalização simples e a condicionalização de Jeffrey a E são as únicas maneiras racionais de revisar as crenças em resposta à experiência de aprendizagem cujo efeito imediato é o de alterar a probabilidade de E.*

Ainda que muito mais possa ser dito sobre a condicionalização simples, a condicionalização de Jeffrey e outras formas de revisão de crença, essas observações devem dar ao leitor uma noção da importância do Teorema de Bayes nas abordagens subjetivistas da aprendizagem e do apoio evidencial. Embora seja uma trivialidade matemática, o *insight* central do Teorema — de que uma hipótese é apoiada por qualquer conjunto de dados que a torne provável — está no cerne de todas as abordagens subjetivistas da epistemologia, estatística e lógica indutiva.

Bibliografia

ARMENDT, B. Is there a Dutch book argument for probability kinematics?, **Philosophy of Science**, v. 47, n. 4, p. 583-588, 1980.

BAYES, T; PRICE, M. An Essay Toward Solving a Problem in the Doctrine of Chances, **Philosophical Transactions of the Royal Society of London**, n. 53, p. 370-418.

BIRNBAUM, A. On the foundations of statistical inference, **Journal of the American Statistical Association**, v. 57, n. 298, p. 269-306, 1962.

CARNAP, R. **Logical foundations of probability**. Chicago: University of Chicago press, 1962.

CHIHARA, C. S. Some problems for Bayesian confirmation theory, **The British Journal for the Philosophy of Science**, v. 38, n. 4, p. 551-560, 1987.

CHRISTENSEN, D. Measuring Evidence, **Journal of Philosophy**, v. 96, 437-61,

1999.

DALE, A. I. Thomas Bayes: a memorial, **The Mathematical Intelligencer**, v. 11, p. 18-19, 1989.

DALE, A. I. **A history of inverse probability: From Thomas Bayes to Karl Pearson**. Springer Science & Business Media, 1999.

EARMAN, J. **Bayes or Bust? A Criticial Examination of Bayesian Confirmation Theory**. Cambridge, ma: MIT Press, 1992.

EDWARDS, A. W. F. **Likelihood**. Cambridge University Press, 1972.

GLYMOUR, C. **Theory and Evidence**. Princeton University Press, 1980.

HACKING, I. **Logic of Statistical Inference**. Cambridge University Press, 1965.

HÁJEK, A. Interpretations of the Probability Calculus. *In*: ZALTA, E. N. (ed.). **Stanford Encyclopedia of Philosophy**. Summer Edition. URL = <https://plato.stanford.edu/archives/sum2003/entries/probability-interpret/>.

HAMMOND, P. Elementary non-Archimedean representations of probability for decision theory and games, *In*: HUMPHREYS, P., ed., **Patrick Suppes: Scientific Philosopher**, Springer, Dordrecht, p. 25-61, 1994.

HARPER, W. Rational Belief Change, Popper Functions and Counterfactuals, *In*: HARPER, W.; HOOKER, C., eds., *In*: **Foundations of probability theory, statistical inference, and statistical theories of science**, Springer, Dordrecht, p. 73-115, 1976.

HARTIGAN, J. A. **Bayes Theory**. Springer Science & Business Media, 1983.

HOWSON, C. Some recent objections to the Bayesian theory of support, **The British journal for the philosophy of science**, v. 36, n. 3, p. 305-309, 1985.

JEFFREY, R. Alias Smith and Jones: The testimony of the senses, **Erkenntnis**, p. 391-399, 1987.

JEFFREY, R. **Probability and the Art of Judgment**. Cambridge University Press, 1992.

JOYCE, J. M. **The foundations of causal decision theory**. Cambridge University Press, 1999.

KAHNEMAN, D.; TVERSKY, A. On the psychology of prediction, *Psychological Review*, v. 80, n. 4, p. 237, 1973.

KAPLAN, M. **Decision Theory as Philosophy**. Cambridge University Press,

1996.

LEVI, I. Imprecision and indeterminacy in probability judgment, **Philosophy of Science**, v. 52, n.3, p. 390-409, 1985.

MAHER, P. Subjective and Objective Confirmation, **Philosophy of Science**, v. 63, n. 2, p. 149-174, 1996.

MCGEE, V. Learning the Impossible, *In*: EELLS, E.; SKYRMS, B., eds., **Probability and conditionals: Belief revision and rational decision**, p. 179-199, 1994.

MORTIMER, H. **The logic of induction**. 1988.

NOZICK, R. **Philosophical Explanations**. Harvard University Press, 1981.

RÉNYI, A. On a new axiomatic theory of probability. **Acta Mathematica Hungarica**, v. 6, n. 3-4, p. 285-335, 1955.

ROYALL R. **Statistical Evidence: A Likelihood Paradigm**. CRC Press, 1997.

SKYRMS, B. Dynamic coherence and probability kinematics. **Philosophy of Science**, v. 54, n. 1, p. 1-20, 1987.

SOBER, E. Bayesianism—Its scope and limits, *In*: **Proceedings-British Academy**. OXFORD UNIVERSITY PRESS INC., p. 21-38, 2002.

SPHON, W. The representation of Popper measures, **Topoi**, v. 5, n. 1, p. 69-74, 1986.

STIGLER, S. M. Thomas Bayes's bayesian inference, **Journal of the Royal Statistical Society: Series A (General)**, v. 145, n. 2, p. 250-258, 1982.

SWINBURNE, R. **Bayes's Theorem**. 2002.

TALBOT, W. Bayesian Epistemology. *In*: ZALTA, E. N. (ed.). **Stanford Encyclopedia of Philosophy**. Fall Edition. 2001. URL = <https: //plato.stanford.edu/archives/fall2001/entries/epistemology-bayesian/>.

TELLER, P. Conditionalization, observation, and change of preference", *In*: HARPER, W.; HOOKER, C.A., eds., **Foundations of probability theory, statistical inference, and statistical theories of science**. Springer, Dordrecht, p. 205-259, 1976.

WILLIAMSON, T. **Knowledge and its Limits**. Oxford University Press, 2000.

VAN FRAASSEN, B. C. A New Argument for Conditionalization. **Topoi**, v. 18, n. 2, p. 93-96, 1999.

Complemento - Exemplos, Tabelas e Esboços de Provas*

Autoria: James Joyce
Tradução: Débora de Oliveira Silva & Sérgio R. N. Miranda
Revisão: Guilherme A. Cardoso

Exemplo 1: Teste Aleatório de Drogas

João Ninguém é membro (aleatoriamente escolhido) de uma grande população na qual 3% são usuários de heroína. Ele testa positivo para heroína em um teste de drogas que identifica corretamente os usuários 95% das vezes e os não usuários 90% das vezes. Para determinar a probabilidade de que João

*JOYCE, J. 'Bayes' Theorem", In: ZALTA, E. N. (ed.) **Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/bayes-theorem/supplement.html#1. Acesso em: 05 jan. 2022.

The following is the translation of the supplement entry on Bayes' Theorem by James Joyce in the **Stanford Encyclopedia of Philosophy.** The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/bayes-theorem/supplement.html#1. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. We'd like to thank the Editors of the **Stanford Encyclopedia of Philosophy**, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

Ninguém seja um usuário de heroína ($= H$) dado o resultado positivo do teste ($= E$), aplicamos o Teorema de Bayes utilizando os seguintes valores:

- Sensitividade $=$ $\mathbf{P}_H(E) = 0,95$
- Especificidade $= 1 -$ $\mathbf{P}_{\sim H}(E) = 0,90$
- Probabilidade "Prévia" Inicial $=$ $\mathbf{P}(H) = 0,03$

O cálculo nos mostra que $\mathbf{P}_E(H) = 0,03 \text{ x } 0,95/[0,03 \text{ x } 0,95 + 0,97 \text{ x } 0,1] = 0,227$. Portanto, mesmo que a probabilidade do pós teste de João Ninguém ser um usuário seja mais de sete vezes maior do que a da população em geral, ainda é bastante improvável que ele seja um usuário. (Repare como um resultado positivo de um teste bastante confiável pode deixar a probabilidade de H muito baixa quando a sua probabilidade de base inicial é baixa no começo!).

Exemplo 2: Novamente o Teste Aleatório de Drogas

Lembremos que João Ninguém, um membro (aleatório) de uma população na qual 3% são usuários de heroína, testa positivo para heroína em um teste com sensitividade de $0,95$ e especificidade de $0,90$. Uma vez que $\mathbf{P}_E(H) = 0,227$ excede $\mathbf{P}(H) = 0,03$, esse resultado fornece forte evidência **incremental** para pensarmos que ele usa heroína. Porém, a evidência **total** para essa conclusão permanece fraca. Como o uso de heroína é raro na população em geral é muito mais provável o teste estar errado nessa circunstância do que João Ninguém ser um usuário.

Repare como a evidência incremental e a total fazem usos diferentes da informação acerca da taxa base de uso de heroína na população. Ao avançar questões sobre a confirmação incremental, ignora-se inteiramente a taxa base porque ela é incorporada em $\mathbf{P}(H)$ e $\mathbf{P}_E(H)$. Mas quando se avançam questões sobre a evidência total, deve-se acompanhar atentamente a taxa base, pois ela quase sempre fornece informação evidencialmente relevante acerca da hipótese. No caso de João Ninguém, por exemplo, a taxa base baixa afeta o resultado positivo do teste. Muitas das vezes as pessoas cometem a "falácia da taxa base" (KAHNEMAN; TVERSKY, 1973, p. 237-251) ao confundirem a evidência

incremental com a evidência total. Elas tratam o resultado de um teste altamente (mas não completamente) confiável como se ele fornecesse evidência conclusiva para a verdade de alguma hipótese mesmo quando a baixa probabilidade anterior da hipótese devesse levá-los a questionar a acurácia do resultado do teste em questão.

Tabela 3: Medidas Relacionadas à Evidência Total

P(H) como evidência total para H

	Multiplicativa	**Aditiva**
LÍQUIDA	**P**$(H)/$**P**$(\sim H) =$ **O**(H)	**P**$(H) -$ **P**$(\sim H) = 2[$**P**$(H) - 1/2]$
SALDO	**P**$(H)/$**P**$(H^*) =$ **B**(H, H^*)	**P**$(H) -$ **P**$(H^*) =$ **P**$(H\& \sim H^*) -$ **P**$(\sim H \& H^*)$

O(H) como evidência total para H

	Multiplicativa	**Aditiva**
LÍQUIDA	**O**$(H)/$**O**$(\sim H) =$ **O**$(H)^2$	**O**$(H) -$ **O**$(\sim H) = [$**P**$(H) -$ **P**$(\sim H)]/$**P**$(\sim H)$**P**(H)
SALDO	**O**$(H)/$**O**$(H^*) =$ **B**$(H, H^*)/$**B**$(\sim H, \sim H^*)$	**O**$(H) -$ **O**$(H^*) = [$**P**$(H) -$ **P**$(H^*)]/$**P**$(\sim H)$**P**$(\sim H^*)$

Tabela 4: Medidas da evidência incremental para *H* fixa e *E* variável

P(H) como evidência total para H

	Razão	**Diferença**
Efetiva	**PR**(H,E)/**PR**$(H,\sim E)$ = **P**$_E(H)$/**P**$_{\sim E}(H)$ = **LR**(E,H)	**PD**(H,E) − **PD**$(H,\sim E)$ = **P**$_E(H)$ − **P**$_{\sim E}(H)$
Diferencial	**PR**(H,E)/**PR**(H,E^*) = **P**$_E(H)$/**P**$_{E^*}(H)$ = **LR**$(E,E^*;H)$	**PD**(H,E) − **PD**(H,E^*) = **P**$_E(H)$ − **P**$_{E^*}(H)$

O(H) como evidência total para H

	Razão	**Diferença**
Efetiva	**OR**(H,E)/**OR**$(H,\sim E)$ = **O**$_E(H)$/**O**$_{\sim E}(H)$ = [**P**$(E\&H)$**P**$(\sim E\& \sim H)$]/[**P**$(E\& \sim H)$**P**$(\sim E\&H)$]	**CD**(H,E) − **CD**$(H,\sim E)$ = [**P**$(E\&H)$/[**P**$(E\& \sim H)$] − [**P**$(\sim E\&H)$/**P**$(\sim E\& \sim H)$]
Diferencial	**OR**(H,E)/**OR**(H,E^*) = **O**$_E(H)$/**O**$_{E^*}(H)$ = [**P**$(E\&H)$**P**$(E^*\& \sim H)$]/[**P**$(E\& \sim H)$**P**$(E^*\&H)$]	**PD**(H,E) − **PD**(H,E^*) = [**P**$(E\&H)$/[**P**$(E\& \sim H)$] − [**P**$(E^*\&H)$/**P**$(E^*\& \sim H)$]

Exemplo 3: Uma ilustração da diferença entre PR e CR

A razão de probabilidade [*Probability Ratio* (PR)] e a razão de chances [*Odds Ratio* (OD)] geram diferentes vereditos sobre o grau relativo com que E confirma incrementalmente H e H^* quando (a) de alguma maneira H prevê E mais fortemente do que H^*, mas (b) $\sim H$ prevê E muito mais fortemente do que $\sim H^*$. A condição geral é a seguinte:

$$\textbf{PR}(H,E) > \textbf{OR}(H,E) \text{ e } \textbf{OR}(H^*,E) < \textbf{PR}(H^*,E)$$

se e somente se

$$1 < \textbf{LR}(H, H^*; E) < \textbf{LR}(\sim H, \sim H^*; E)$$

Para ilustrar, suponha que um exame que testa o conhecimento de matemática elementar seja aplicado aos estudantes no fim do segundo ano do ensino médio. Cada estudante que realiza o exame fez um curso de geometria ($= H$), um curso de álgebra ($= H^*$), ambos os cursos ou não fez qualquer curso de matemática. Estatísticas confiáveis de uma grande população mostram que os estudantes tendem a passar no exame nas seguintes proporções:

	$H\&H^*$	$H\& \sim H^*$	$\sim H\&H^*$	$\sim H\& \sim H^*$
$E =$ APROVADO	$0,12$	$0,001$	$0,378$	$0,001$
$\sim E =$ REPROVADO	$0,003$	$0,005$	$0,2$	$0,292$

Cerca de 12% dos estudantes fazem ambos os cursos e eles são aprovados $97,5\%$ das vezes. A proporção minúscula de estudantes $(0,6\%)$ que fazem geometria, mas não álgebra, são aprovados apenas uma vez em seis. Cerca de 58% dos estudantes fazem álgebra, mas não geometria, sendo então aprovados com uma taxa de 65%. Finalmente, um pouco menos de 30% dos estudantes não fazem nenhum dos cursos, sendo aprovados menos do que $0,4\%$ das vezes. Em geral, apenas a álgebra é um indicador moderadamente forte de aprovação, mas a geometria sozinha promove um pouco a aprovação (e é melhor do que nada), e adicionar a geometria à álgebra torna quase certa a aprovação. Dados esses números, saber que um estudante foi aprovado no exame confirmará incrementalmente a hipótese de que ele fez álgebra e a hipótese de que ele fez geometria. **PR** e **OR** discordam sobre quais das duas hipóteses recebem **mais** suporte incremental.

- $\textbf{PR}(H, E) = 1,88 > \textbf{PR}(H^*, E) = 1,42$
- $\textbf{OR}(H, E) = 2,16 < \textbf{OR}(H^*, E) = 106,2$

Uma comparação das razões de probabilidade nos diz que E fornece um pouco mais de evidência incremental para H do que a H^*, enquanto uma comparação das razões de chances indicam que E fornece muito mais evidência incremental para H^* do que a H. A comparação da razão de probabilidade reflete o fato de que um curso de geometria é muito melhor do que um curso de álgebra como

um indicador de que um estudante tenha feito **ambos** os cursos. Em contraste, a grande, opostamente direcionada disparidade nas razões de chance é devida ao fato de que os estudantes têm uma chance decente de serem aprovados sem geometria ((44%)) porque é mais provável que eles tenham feito álgebra, mas eles quase não têm nenhuma chance de serem aprovados sem álgebra ($(0,7\%)$) porque eles provavelmente não terão feito nenhum curso de matemática.

Exemplo 4: Uma ilustração da diferença entre PR e PD

A razão de probabilidade e a diferença de probabilidade discordam sobre o grau relativo com que E confirma incrementalmente H e H^* quando (a) de alguma maneira H prevê E mais fortemente do que H^*, mas (b) H^* é muito mais provável do que H. Aqui está a condição geral:

$$\textbf{PR}(H,E) > \textbf{PR}(H,E) \text{ e } \textbf{PD}(H,E) < \textbf{PD}(H,E)$$

se e somente se

$$1 < (\textbf{PR}(H,E)-1)/(\textbf{PR}(H^*,E)-1) < \textbf{P}(H^*)/\textbf{P}(H)$$

Para ilustrar, suponha que um paciente apareça em uma sala de emergência com dor de cabeça severa, dores musculares e fadiga. Esses sintomas são consistentes com a doença de Lyme ($= H$), a qual é rara na área, e a influenza ($= H^*$), que é mais comum. Nós estamos prestes a saber se o paciente tem febre ($= E$). As estatísticas conhecidas para as pessoas que exibem os sintomas do paciente são as seguintes:

	$H\&H^*$	$H\& \sim H^*$	$\sim H\&H^*$	$\sim H\& \sim H^*$
$E =$ febre	$0,007$	$0,020$	$0,184$	$0,004$
$\sim E =$ sem febre	$0,001$	$0,002$	$0,080$	$0,702$

Enquanto somente 3% das pessoas têm doença de Lyme, essa doença é acompanhada por febre 90% das vezes. Um grupo muito maior de pacientes (cerca de 27%) que têm gripe ficam febris em apenas 70% das vezes. Um tanto

paradoxalmente, os pacientes com gripe e doença de Lyme apresentam febre um pouco menos frequentemente do que os pacientes que têm apenas a doença de Lyme (talvez porque febres induzidas por gripe tendem a aparecer mais rápido do que aquelas causadas pela doença de Lyme). Dadas essas estatísticas, saber que um paciente tem febre confirma incrementalmente a hipótese de que ele tem doença de Lyme e a hipótese de que ele tem gripe. **PR** e **PD** discordam sobre quais das duas hipóteses recebe maior incremento de apoio.

- **PR**$(H, E) = 4,19 >$ **PR**$(H^*, E) = 3,27$
- **PD**$(H, E) = 0,09 <$ **PD**$(H^*, E) = 0,61$

Segundo a medida de razão, uma febre fornece mais evidência incremental para a doença de Lyme do que para a gripe simplesmente porque a febre acompanha mais frequentemente a primeira do que a última. No entanto, segundo a medida de razão, uma febre confirma incrementalmente mais um diagnóstico de influenza do que um diagnóstico de doença de Lyme. Já que muitos mais pacientes são acometidos por gripe do que por doença de Lyme e dado que ambas as doenças produzem febre com uma taxa alta, a probabilidade de H^* acaba sendo aumentada em uma quantidade absoluta maior do que a probabilidade de H. Em geral, quando duas hipóteses têm poder preditivo similar no que diz respeito a alguma porção de evidência, a medida de diferença de probabilidade tem um incremento maior de confirmação para a hipótese que é antecedentemente mais provável.

Exemplo 5: Uma ilustração da diferença entre OR e PD

A razão de chances e a diferença de probabilidade geram vereditos díspares sobre o grau relativo com que E confirma incrementalmente H e H^* nas condições que se seguem:

$$\textbf{PD}(H, E) > \textbf{PD}(H, E) \text{ e } \textbf{OR}(H, E) < \textbf{OR}(H, E)$$

se e somente se

$$[\textbf{P}(\sim H\&E)\textbf{P}(H)]/[\textbf{P}(\sim H^*\&E)\textbf{P}(H^*)] >$$
$$[\textbf{P}(H\&E) - \textbf{P}(H)\textbf{P}(E)]/[\textbf{P}(H^*\&E) - \textbf{P}(H^*)\textbf{P}(E)] > 1$$

Para ter uma ideia do que isso envolve, imagine uma corporação na qual os funcionários podem ou não ser muito bem pagos ($= H$) e podem ou não ter trabalhos fáceis ($= H^*$). A distribuição de trabalhos entre homens ($= E$) e mulheres na folha de pagamento é a seguinte:

	$H\&H^*$	$H\& \sim H^*$	$\sim H\&H^*$	$\sim H\& \sim H^*$
$E =$ *homem*	$0,018$	$0,102$	$0,019$	$0,162$
$\sim E =$ mulher	$0,002$	$0,098$	$0,001$	$0,598$

Em uma firma sexista, apenas 2% dos funcionários têm trabalhos fáceis que pagam bem e 90% deles são homens. Já 20% da equipe têm trabalhos que pagam bem, mas são difíceis de serem executados; esses trabalhos são divididos quase igualmente entre homens (51%) e mulheres (49%). Outros 2% dos funcionários têm trabalhos fáceis e de baixo salário; 95% deles são homens. A maioria dos trabalhadores (76%) é mal remunerada e têm trabalhos difíceis. Essas vagas são ocupadas predominantemente (79%) pelas mulheres. Dadas essas estatísticas, saber que um funcionário é um homem confirmará incrementalmente a hipótese de que ele é bem pago e a hipótese de que ele tem um trabalho fácil. **PD** e **OR** discordam acerca de qual das duas hipóteses recebe o maior incremento de apoio da evidência.

- **PD**$(H, E) = 0,18 >$ **PD**$(H^*, E) = 0,08$
- **OR**$(H, E) = 2,36 <$ **OR**$(H^*, E) = 3,37$

A medida de diferença de probabilidade tem H ganhando um incremento (ligeiramente) maior de confirmação do que H^*. Isso ocorre em grande parte porque: (i) o número de homens em trabalhos bem pagos e difíceis é muito maior do que o número de homens em trabalhos mal pagos e fáceis; e (ii) o número de homens em trabalhos de baixo salário e difíceis é grande relativamente ao número total de homens. A medida da razão de chance tem H^* ganhando um incremento (ligeiramente) maior de confirmação do que H porque: (i*) trabalhos com baixo salário ou difíceis são igualmente bons como contraindicadores de E; mas (ii) trabalhos fáceis são muito melhores do que trabalhos bem pagos como indicadores positivos de E.

Tabela 5A: Medidas de evidência incremental para *E* fixo e *H* variável

Quatro medidas de mudança incremental na evidência LÍQUIDA

	Razão	**Diferença**
P = Total	**PR**(H, E)/**PR**$(\sim H, E)$ = **LR**(H, E)	**PD**(H, E) − **PD**$(\sim H, E)$ = 2**PD**(H, E)
O = Total	**OR**(H, E)/**OR**$(\sim H, E)$ = **LR**$(H, E)^2$	**CD**(H, E) − **CD**$(\sim H, E)$ = [**P**(H)−**P**$(\sim H)$]/[**P**(H)**P**$(\sim H)$**O**(E)]

Quatro medidas de mudança incremental no SALDO da evidência TOTAL

	Razão	**Diferença**
P = Total	**PR**(H, E)/**PR**(H^*, E) =	**PD**(H, E) − **PD**(H^*, E) = [**P**$_E(H)$−**P**$_E(H^*)$]−[**P**(H)− **P**(H^*)]
O = Total	**OR**(H, E)/**OR**(H^*, E) = **LR**$(H, H^*; E)$/**LR**$(\sim H, \sim H^*; E)$	**CD**(H, E) − **CD**$(\sim H, E)$ = [**P**(H) − **P**(H^*)]/[**P**$(\sim H)$**P**$(\sim H^*)$**O**(E)]

Esboço de Prova: Lema 3.5

(3.5) Lema: Se H e H^* implicam E e se **P**$(H) >$ **P**(H^*), então **LR**$(H, H^*, E) = 1$ e **LR**$(\sim H, \sim H^*; \sim E) > 1$.

Esboço da Prova: Se H e H^* implicam E, então

a. **P**$_H(E) =$ **P**$_{H^*}(E) = 1$
b. **P**$_{\sim H}(\sim E) =$ **P**$(\sim E)$/**P**$(\sim H)$
c. **P**$_{\sim H^*}(\sim E) =$ **P**$(\sim E)$/**P**$(\sim H^*)$

(a) implica que **LR**$(H, H^*, E) = 1$. (b) e (c) implicam que **LR**$(\sim H, \sim H^*; \sim E) > 1$ se e somente se **P**$(\sim H^*)$/**P**$(\sim H) > 1$, que sempre será assim quando **P**$(H) >$ **P**(H^*).

Esboço de Prova: Lema 3.6

(3.6) Lema: A condicionalização simples a E é a única regra para a revisão de probabilidades subjetivas que produz uma posterior **Q** com as seguintes propriedades para **quaisquer** prévia tais que $\mathbf{P}(E) > 0$:

i. $\mathbf{Q}(E) = 1$.
ii. **Similaridade Ordinal:** Se H e H^* implicam E, então $\mathbf{P}(H) \geq \mathbf{P}(H^*)$ se e somente se $\mathbf{Q}(H) \geq \mathbf{Q}(H^*)$.

Esboço da Prova: A condicionalização a E obviamente satisfaz (i)-(ii) para qualquer **P**. Para ver por que ela é a **única** regra de revisão que possui essas propriedades para **todas** as probabilidades, repare que **P** pode ser **não atômica** no sentido de que qualquer hipótese à qual ela atribui uma probabilidade positiva pode ser subdividida em hipóteses disjuntas que também têm probabilidades positivas. Para uma não atômica **P** com $\mathbf{P}(E) > 0$, $\mathbf{P}_E$ também será não atômica. Uma vez que **Q** é definida sobre o mesmo conjunto de proposições que **P**, as claúsulas (i)-(ii) garantem que **Q** é não atômica e ordinariamente similar a $\mathbf{P}_E$. Acontece que funções de probabilidade não atômicas definidas sobre o mesmo conjunto de proposições só podem ser ordinariamente similares se elas são idênticas (JOYCE, 1999, p. 134-135). Assim, $\mathbf{Q} = \mathbf{P}_E$, o que significa que (i)-(ii) só pode valer em plena generalidade se a regra de revisão em questão é condicionante.

Teoria dos Jogos*

Autoria: Don Ross
Tradução: Arthur de Castro Machado & Sérgio R. N. Miranda
Revisão: Guilherme A. Cardoso

A teoria dos jogos é o estudo dos modos em que **decisões interativas** de **agentes econômicos** produzem **resultados** a respeito das **preferências** (**utilidades**) desses agentes, em que os resultados em questão poderiam não ter sido pretendidos por nenhum dos agentes. O significado desse enunciado não será claro para o não especialista até que cada uma das expressões em negrito tenha sido explicada e destacada em alguns exemplos. A tarefa principal deste artigo será fazer isso. Contudo, oferecemos, primeiramente, algum contexto histórico e filosófico a fim de motivar o leitor para o trabalho técnico à frente.

---

*ROSS, D. Game Theory, In: ZALTA, E. N. (ed.) **Stanford Encyclopedia of Philosophy**, Winter Edition, Stanford, CA: The Metaphysics Research Lab, 2021. Disponível em: https://plato.stanford.edu/archives/win2021/entries/game-theory/. Acesso em: 18 jan. 2022.

The following is the translation of the entry on Game Theory by Don Ross, in the **Stanford Encyclopedia of Philosophy**. The translation follows the version of the entry in the SEP's archives at https://plato.stanford.edu/archives/win2021/entries/game-theory/. This translated version may differ from the current version of the entry, which may have been updated since the time of this translation. The current version is located at a https://plato.stanford.edu/entries/game-theory/. We'd like to thank the Editors of the **Stanford Encyclopedia of Philosophy**, mainly Prof. Dr. Edward Zalta, for granting permission to translate and to publish this entry. Finally, we would like to thank to John Templeton Foundation for financially supporting this project.

1. Motivação Filosófica e Histórica

A teoria dos jogos, no formato conhecido por economistas, cientistas sociais e biólogos, teve sua primeira formulação matemática geral fornecida por John von Neuman e Oskar Morgenstern (1944). Por razões a serem discutidas depois, limitações em sua estrutura formal tornaram a teoria aplicável inicialmente apenas sob condições especiais e limitadas. Como veremos, essa situação mudou dramaticamente nas últimas sete décadas, à medida que essa estrutura foi aprofundada e generalizada. Refinamentos ainda estão sendo feitos, e no final do artigo revisaremos alguns problemas pendentes de ponta. Contudo, desde pelo menos o final da década de 1970, tem sido possível dizer com confiança que a teoria dos jogos é a ferramenta mais importante e útil no kit do analista sempre que ele confronta situações nas quais o que conta (para o agente) como a melhor ação depende de expectativas acerca do que um ou mais outros agentes farão, e o que conta para esses agentes como as melhores ações depende similarmente de expectativas acerca do primeiro agente.

Apesar do fato de a teoria dos jogos ter sido tratada de maneira matemática e logicamente sistemática apenas desde 1944, *insights* teóricos podem ser encontradas entre comentadores desde os tempos antigos. Por exemplo, em dois textos de Platão, o **Laques** e o **Banquete**, Sócrates recorda um episódio da Batalha de Délio que alguns comentadores têm interpretado (provavelmente de modo anacrônico) como envolvendo a seguinte situação. Considere um soldado no fronte, esperando com seus companheiros para repelir um ataque inimigo. Pode ocorrer a ele que, se a defesa é provável de ser bem sucedida, então não é muito provável que sua própria contribuição pessoal será essencial. Mas caso permaneça no fronte, ele corre o risco de ser morto ou ferido - por nenhuma razão, aparentemente. Por outro lado, se o inimigo ganhará a batalha, então suas chances de morrer ou de se ferir ainda são altas, e, agora, muito claramente, por nenhuma razão, uma vez que a linha será completamente destruída de qualquer modo. Com base nesse raciocínio, parece que seria melhor para o soldado fugir da batalha, independentemente de quem seja o vencedor. Obviamente, se todos os soldados pensarem assim - como todos eles aparentemente **deveriam** pensar, uma vez que estão todos em situações

idênticas -, então a batalha será perdida. Obviamente, esse ponto, visto que ocorreu a nós analistas, pode ocorrer aos soldados também. Isso dá à eles uma razão para permanecerem em seus postos? Justamente o contrário: quanto maior for o medo dos soldados de que a batalha será perdida, maior serão seus incentivos para saírem do caminho do perigo. E quanto maior for a crença dos soldados de que a batalha será ganha, sem serem necessárias contribuições de qualquer indivíduo em particular, menor razão eles têm para permanecer e lutar. Se cada soldado **antecipar** esse tipo de raciocínio por parte dos demais, todos irão rapidamente levar a si mesmos ao pânico e desespero, e seu comandante, horrorizado, terá uma derrota em suas mãos antes mesmo do inimigo atacar.

Muito antes de ter surgido a teoria dos jogos para mostrar aos analistas como pensar sistematicamente sobre esse tipo de problema, ela ocorreu a alguns líderes militares reais e influenciou suas estratégias. Assim, o conquistador espanhol, Cortez, quando atracou no México com uma pequena armada que ele tinha boas razões para temer sua capacidade para repelir os ataques dos muito mais numerosos Astecas, removeu o risco de que suas tropas pudessem pensar em uma retirada ao queimar os navios atracados. Com a retirada tendo assim se tornado fisicamente impossível, os soldados espanhóis não possuíam um melhor curso de ação do que permanecer e lutar - e, além disso, lutar com tanta determinação quanto conseguissem reunir. Ainda melhor do que isso, do ponto de vista de Cortez, a sua ação teve um efeito desencorajador na motivação dos Astecas. Ele tomou o cuidado de queimar os navios de modo bastante visível para que os Astecas pudessem ver claramente o que ele havia feito. Os Astecas então teriam reciocinado assim: qualquer comandante que pudesse ser tão confiante a ponto de destruir deliberadamente sua própria opção de ser prudente caso a batalha acabasse mal para ele deve ter boas razões para tão extremo otimismo. Não pode ser sensato atacar um oponente que tenha uma boa razão (qualquer que ela seja, exatamente) para ter a certeza de que não perderá. Os Astecas, então, recuaram para as montanhas em torno, e Cortez teve a vitória mais fácil possível.

Essas duas situações, em Délio e aquela manipulada por Cortez, possuem por trás uma lógica interessante e comum. Note que os soldados não são motivados a recuar **apenas**, ou mesmo principalmente, por sua avaliação

racional dos perigos da batalha e por seu interesse próprio. Eles descobrem uma razão sólida para fugir ao se darem conta de que o que faz sentido para eles fazerem depende do que fará sentido para os outros fazerem, e que todos os outros também podem perceber isso. Mesmo um soldado bastante corajoso pode preferir correr ao invés de morrer heroicamente, mas sem sentido, tentando conter sozinho a maré que se aproxima. Por isso, nós poderíamos imaginar, sem contradição, uma circunstância em que um exército, cujos membros são todos corajosos, foge em alta velocidade antes que o inimigo faça qualquer movimento. Se os soldados **são** realmente corajosos, então esse com certeza não é o resultado que qualquer um deles gostaria; cada um teria preferido que todos permanecessem e lutassem. Assim, o que temos aqui é um caso em que a **interação** de muitos processos individualmente racionais de tomada de decisão - um processo para cada soldado - produz um resultado que não é pretendido por ninguém. (A maioria dos exércitos tenta evitar esse problema assim como Cortez o fez. Como eles geralmente não podem tornar a retirada **fisicamente** impossível, eles a tornam **economicamente** impossível: eles atiram nos desertores. Assim, permanecer e lutar é o curso de ação individualmente racional de cada soldado, pois o custo de correr é certamente tão alto quanto o de permanecer.)

Outra fonte clássica que convida a essa sequência de raciocínio é encontrada em **Henrique V** de Shakespeare. Durante a Batalha de Azincourt, Henrique decidiu abater seus prisioneiros franceses à vista do inimigo e para a surpresa de seus subordinados, que descrevem a ação como ausente de caráter moral. As razões que Henrique oferece aludem a considerações não-estratégicas: ele está com medo de que os prisioneiros se libertem e ameacem sua posição. Contudo, um teórico dos jogos poderia ter fornecido a ele uma justificação estratégica suplementar (e similarmente prudencial, embora talvez não de caráter moral). Suas próprias tropas observam que os prisioneiros foram mortos, e observam que o inimigo observou isso. Portanto, eles sabem qual destino os aguardará nas mãos do inimigo se não vencerem. Metaforicamente, mas de modo muito efetivo, seus barcos foram queimados. Plausivelmente, o massacre dos prisioneiros enviou um sinal para os soldados de ambos os lados, mudando, assim, seus incentivos de modo a favorecer os prospectos inglesas de vitória.

Esses exemplos podem parecer relevantes apenas àqueles que se

encontram em situações sórdidas de competição de cortar gargantas. Alguém poderia pensar que talvez sejam importantes para generais, políticos, mafiosos, treinadores esportivos e demais cujos trabalhos envolvem a manipulação estratégica dos outros, mas que o filósofo deveria apenas deplorá-los como imorais. Mas essa conclusão seria altamente prematura. O estudo da **lógica** que governa as interrelações entre incentivos, interações estratégicas e resultados tem sido fundamental na filosofia política moderna, desde séculos antes que alguém tivesse um nome explícito para esse tipo de lógica. Filósofos compartilham com cientistas sociais a necessidade de serem capazes de representar e sistematicamente modelar não só o que eles pensam que as pessoas normativamente **devem** fazer, mas o que elas frequentemente fazem **de fato** em situações interativas.

O **Leviatã** de Hobbes é frequentemente considerado a obra fundadora da filosofia política moderna, o texto que começou a rodada contínua de análises da função e justificação do estado e das restrições das liberdades individuais. O núcleo do raciocínio de Hobbes pode ser apresentado diretamente como se segue. A melhor situação para todas as pessoas é aquela em que cada uma é livre para fazer o que lhe apetece. (Pode-se ou não concordar com isso por uma questão de psicologia ou ideologia, mas essa é a suposição de Hobbes.) Frequentemente, as pessoas livres desejarão cooperar umas com as outras com o objetivo de realizar projetos que seriam impossíveis para um indivíduo agindo sozinho. Mas se houver qualquer agente imoral ou amoral por perto, elas notarão que seus interesses podem, pelo menos por vezes, ser melhor atendidos quando recebem os benefícios da cooperação e não dão retorno. Por exemplo, suponha que você concorde em me ajudar a construir a minha casa em troca da promessa de eu ajudá-lo a construir a sua casa. Após a construção da minha casa terminar, posso fazer com que o seu trabalho saia de graça simplesmente negando a minha promessa. Contudo, depois me dou conta de que se isso o deixar sem casa, você terá um incentivo para tomar a minha. Isso me colocará em um estado constante de medo de você. Eu posso minimizar melhor esse custo ao lhe atacar e matar na primeira oportunidade. Obviamente, você pode antecipar todo esse raciocínio feito por mim, e então ter boas razões para me atacar e matar. Como posso antecipar **esse** raciocínio feito por **você**, meu medo original não foi paranoico;

nem o foi o seu medo de mim. Na verdade, nenhum de nós precisa realmente ser imoral para manter essa linha de raciocínio mútuo em funcionamento; só precisamos pensar que há alguma **possibilidade** de que o outro possa tentar trapacear na negociação. Assim que uma pequena parcela de dúvida entra na mente de alguém, o incentivo induzido pelo medo das consequências de ser **antecipado** - atacado antes de atacar primeiro - rapidamente se torna opressivo em ambos os lados. Se qualquer um de nós tiver quaisquer recursos próprios que o outro possa querer, essa lógica assassina pode tomar conta muito antes de sermos tão tolos a ponto de imaginar que poderíamos realmente ir longe fazendo acordos para ajudar uns aos outros a construir casas em primeiro lugar. Deixados aos seus próprios recursos, os agentes que possuem, ao menos por vezes, um ligeiro interesse próprio podem falhar repetidamente em derivar os benefícios da cooperação, e, em vez disso, ficar presos num estado de "guerra de todos contra todos", nas palavras de Hobbes. Nessas circunstâncias, a vida humana, como ele vívida e famosamente colocou, será "solitária, pobre, desagradável, bruta e curta".

A solução proposta por Hobbes para esse problema é a tirania. As pessoas podem contratar um agente - um governo - cujo trabalho é punir qualquer pessoa que quebrar alguma promessa. Enquanto a punição ameaçadora for suficientemente terrível, o custo de descumprir promessas excederá o custo de mantê-las. A lógica aqui é idêntica àquela usada por um exército quando ele ameaça atirar em desertores. Se todas as pessoas souberem que esses incentivos valem para a maioria dos outros, então a cooperação será não apenas possível, mas poderá ser a norma esperada, tal que a guerra de todos contra todos se torne uma paz geral.

Hobbes leva a lógica desse argumento a uma conclusão muito forte, argumentando que ele implica não só um governo com o direito e o poder para impor a cooperação, mas um governo "não-dividido", no qual a vontade arbitrária de um único governante deve impor obrigação absoluta a todos. Poucos teóricos políticos contemporâneos pensam que os passos seguidos por Hobbes em seu raciocínio até essa conclusão são sólidos e válidos. Contudo, trabalhar com esses problemas aqui nos levaria longe do nosso tópico para detalhes da filosofia política contratualista. O que é importante no presente contexto é que esses detalhes, como eles são de fato investigados nos debates contemporâneos,

envolvem uma interpretação sofisticada dos problemas que usa recursos da teoria dos jogos contemporânea. Além disso, o ponto mais básico de Hobbes de que a justificação fundamental para a autoridade e as práticas coercitivas do governo são as próprias necessidades das pessoas de se protegerem do que os teóricos dos jogos chamam de "dilemas sociais" é aceito por muitos, se não pela maioria, dos teóricos políticos. Note ainda que Hobbes **não** argumentou que a tirania é uma coisa desejável por si mesma. A estrutura de seu argumento é que a lógica da interação estratégica deixa apenas dois resultados políticos gerais possíveis: tirania ou anarquia. Agentes sensatos escolhem então tirania como o menor de dois males.

Os raciocínios dos soldados atenienses, de Cortez, e dos agentes políticos de Hobbes possuem uma lógica em comum, uma lógica derivada de suas situações. Em cada caso, o mais importante aspecto do ambiente para a conquista dos resultados preferidos de um agente é o conjunto de expectativas de outros agentes e as reações possíveis desses agentes às estratégias que ele adota. A distinção entre agir **parametricamente** num mundo totalmente passivo e agir **não-parametricamente** em um mundo que tenta agir por antecipação de ações é fundamental. Se você deseja chutar uma pedra ladeira abaixo, você precisa se preocupar apenas com a massa da pedra relativa à força do seu golpe, a extensão com que ela está ligada à sua superfície de suporte, a inclinação do chão do outro lado da pedra, e o impacto esperado da colisão com o seu pé. Os valores de todas essas variáveis são independentes de seus planos e intenções, uma vez que a pedra não tem interesses próprios e não realiza ações para tentar ajudá-lo ou frustrá-lo. Por outro lado, se você deseja chutar uma pessoa morro abaixo, então, a não ser que essa pessoa esteja inconsciente, amarrada ou, de outra maneira, incapacitada, você provavelmente não obterá sucesso a menos que possa disfarçar seus planos até que seja tarde demais para essa pessoa realizar uma ação evasiva ou preventiva. Além disso, espera-se que as respostas prováveis dessa pessoa tenham custos para você, o que seria prudente levar em consideração. Por sua vez, as probabilidades relativas dessas respostas dependerão das expectativas dessa pessoa sobre como você responderá à sua resposta à agressão. (Considere a diferença que isso fará para ambos os seus raciocínios caso um de vocês ou ambos estejam armados, ou um de vocês seja

maior que o outro, ou um de vocês seja o chefe do outro.) Os problemas lógicos associados com o segundo tipo de situação (chutar a pessoa em vez da pedra) são comumente muito mais complicados, como um simples exemplo hipotético o ilustrará.

Primeiramente, suponha que você deseja atravessar um rio cujas margens são ligadas por três pontes, e que nadar, caminhar ou atravessar de barco seja impossível. A primeira ponte é conhecida por ser segura e livre de obstáculos; se você tentar atravessar por lá, certamente terá sucesso. A segunda ponte está sob um penhasco do qual às vezes rolam grandes pedras. A terceira é habitada por cobras mortais. Agora suponha que você deseje classificar as três pontes em ordem no que diz respeito às suas preferências como pontos de passagem. A não ser que você tenha uma satisfação prazerosa em arriscar sua vida - o que, como um ser humano, você poderia ter, uma complicação que nós abordaremos mais adiante neste artigo -, o seu problema de decisão aqui é simples. A primeira ponte é obviamente a melhor, uma vez que é segura. Para classificar ordenadamente as outras duas pontes, você precisa de informações sobre seus níveis relativos de perigo. Se você puder estudar a frequência dos deslizamentos de pedras e dos movimentos das cobras por um tempo, poderá ser capaz de calcular que a probabilidade de ser esmagado por uma pedra na segunda ponte é de 10% e de ser atacado por uma cobra na terceira ponte é de 20%. Seu raciocínio aqui é estritamente paramétrico, pois nem as pedras nem as cobras estão tentando influenciar suas ações (por exemplo, escondendo seus padrões usuais de comportamento porque sabem que você os estuda). É óbvio o que você deve fazer aqui: atravessar a ponte segura. Vamos agora complicar um pouco a situação. Suponha que a ponte com as pedras estivesse imediatamente diante de você e a ponte segura estivesse a um dia de uma difícil caminhada rio acima. A sua tomada de decisão aqui é ligeiramente mais complicada, mas ainda é estritamente paramétrica. Você teria que decidir se valeria a pena trocar o custo de uma longa caminhada pela penalidade de 10% de chance de ser atingido por uma pedra. Contudo, isso é tudo o que você tem que decidir, e sua probabilidade de ter uma travessia bem-sucedida depende inteiramente de você; o ambiente não está interessado em seus planos.

No entanto, se complicarmos a situação adicionando um elemento não-

paramétrico, ela se torna mais desafiadora. Suponha que você seja um fugitivo, e que, esperando na outra margem do rio com uma arma, esteja o seu perseguidor. Vamos supor que ele irá pegá-lo e atirar em você só se ele esperar na ponte em que você tentar atravessar; de outro modo, você escapará. Enquanto você raciocina sobre qual ponte escolher, ocorre a você que ele está logo ali tentando antecipar seu raciocínio. Parecerá, certamente, que escolher a ponte segura logo de cara seria um erro, visto que é justamente onde ele irá esperá-lo, e suas chances de morrer beirarão a certeza. Então, talvez você devesse arriscar as pedras, visto que lá suas chances são muito melhores. Mas espere... se você pode chegar a essa conclusão, seu perseguidor, que é justamente tão racional e bem informado quanto você, pode antecipar que você irá chegar a ela, e estará lhe esperando caso você fuja na ponte das pedras. Então, talvez você deva arriscar suas chances com as cobras; isso é o que ele deve menos esperar. Mas, então, não... se ele espera que você irá esperar que ele irá menos esperar por isso, então ele esperará por isso ainda mais. Com pavor, você percebe que esse dilema é geral: você deve fazer o que o seu perseguidor menos espera; mas o que quer que seja que você mais espere que ele menos espere é automaticamente o que ele irá mais esperar. Parece que você está preso em um impasse. Tudo o que deve consolá-lo um pouco aqui é que, do outro lado do rio, seu perseguidor esteja preso exatamente na mesma perplexidade; incapaz de decidir em qual ponte esperar porque, assim que se imagina se comprometendo com alguma delas, notará que se ele pode encontrar uma melhor razão para escolher uma ponte, você pode antecipar essa mesma razão e assim tentar evitá-lo.

Sabemos por experiência que, em situações como essa, as pessoas não costumam ficar paradas e hesitantes para sempre. Como veremos posteriormente, **há** uma única melhor solução disponível para cada jogador. Contudo, até os anos 1940, nem filósofos nem economistas sabiam como encontrá-la matematicamente. Por essa razão, os economistas eram forçados a tratar influências não-paramétricas como se fossem complicações das paramétricas. É provável que isso pareça estranho ao leitor, uma vez que, como nosso exemplo do problema da travessia da ponte pretendia mostrar, características não-paramétricas são frequentemente características fundamentais dos problemas de tomada de decisão. Parte da explicação para a

entrada em campo relativamente tardia da teoria dos jogos reside nos problemas com os quais economistas têm historicamente se preocupado. Economistas clássicos, tais como Adam Smith e David Ricardo, estavam principalmente interessados na questão de como agentes em mercados muito grandes - nações inteiras - poderiam interagir de modo a adquirir o máximo de riqueza monetária para si próprios. O *insight* básico de Smith, de que a eficiência é melhor maximizada pelos agentes diferenciando primeiramente as suas contribuições potenciais e então buscando de maneira livre negociações mutuamente vantajosas, foi verificado matematicamente no século XX. Todavia, a demonstração desse fato se aplica apenas em condições de "competição perfeita", isto é, quando indivíduos ou empresas se deparam com nenhum custo de entrada ou saída nos mercados, quando não há economias de escala, e quando nenhuma das ações dos agente possui efeitos colaterais não-intencionados no bem-estar de outros agentes. Os economistas sempre reconheceram que esse conjunto de pressuposições é puramente uma idealização para fins de análise, não um possível estado de coisas que alguém poderia tentar (ou deveria querer tentar) estabelecer institucionalmente. Mas até a matemática da teoria dos jogos amadurecer próximo do final dos anos 1970, os economistas tinham que esperar que quanto mais próximo um mercado **se aproximasse** da competição perfeita, mais eficiente ele seria. Mas nenhuma expectativa dessas pode ser justificada matemática ou logicamente em geral; de fato, estritamente como uma generalização, a pressuposição se mostrou falsa já na década de 1950.

Esse artigo não é sobre os fundamentos da economia, mas saber que mercados perfeitamente competitivos têm embutido neles uma característica que os tornam suscetíveis à análise paramétrica é importante para entender as origens e o alcance da teoria dos jogos. Porque os agentes não se deparam com custos de entrada em mercados, eles abrirão firmas em qualquer mercado dado até que a competição leve todos os lucros a zero. Isso implica que, se os custos de produção são fixos e a demanda é exógena, os agentes não possuem opção sobre o quanto produzir caso estejam tentando maximizar as diferenças entre seus custos e suas receitas. Esses níveis de produção podem ser determinados de maneira separada para cada agente, assim ninguém precisa prestar atenção

no que os outros estão fazendo; cada agente trata suas contrapartes como características passivas do ambiente. O outro tipo de situação à qual a análise econômica clássica pode ser aplicada sem recorrer à teoria dos jogos é a de um monopólio frente a muitos clientes. Aqui, desde que nenhum cliente possua uma parcela de demanda grande o suficiente para exercer influência estratégica, as considerações não-paramétricas desaparecem e a tarefa da empresa é apenas identificar a combinação de preço e a quantidade de produção em que ela maximiza o lucro. Todavia, competições tanto perfeitas quanto monopolísticas são arranjos de mercado muito especiais e incomuns. Portanto, anteriormente ao advento da teoria dos jogos, os economistas estavam bastante limitados quanto às classes de circunstâncias às quais poderiam aplicar diretamente seus modelos.

Filósofos compartilham com economistas um interesse profissional quanto às condições e técnicas para a maximização do bem-estar. Além disso, os filósofos possuem uma preocupação especial com a justificação lógica de ações, e as ações devem ser frequentemente justificadas por referência aos seus resultados esperados. (Uma tradição em filosofia moral, o utilitarismo, é baseada na ideia de que todas as ações justificáveis devem ser justificadas dessa maneira.) Sem a teoria dos jogos, esses problemas resistem à análise sempre que aspectos não-paramétricos forem relevantes. Demonstraremos isso brevemente em relação ao jogo mais famoso (embora não seja o mais comum), o assim chamado **Dilema do Prisioneiro**, e a outros jogos mais comuns. Ao fazer isso, precisaremos introduzir, definir e ilustrar os elementos básicos e as técnicas da teoria dos jogos.

2. Elementos Básicos e Pressuposições da Teoria dos Jogos

2.1. A Utilidade

Por definição, um agente econômico é uma entidade com **preferências**. Os teóricos dos jogos, assim como economistas e filósofos que estudam a tomada de decisão racional, descrevem as preferências por meio de um conceito abstrato chamado de **utilidade**. Ele se refere a uma classificação, em alguma escala especificada, do bem-estar subjetivo ou mudança no bem-estar subjetivo

que um agente deriva de um objeto ou de um evento. Por "bem-estar", nós nos referimos a algum índice normativo de alinhamento relativo entre os estados do mundo e a avaliação dos agentes dos estados em questão, justificado por referência a algum *framework* de pano de fundo. Por exemplo, podemos avaliar o bem-estar relativo de países (que para certos propósitos podemos modelar como agentes) por referência à sua renda per capita. E podemos avaliar o bem-estar relativo de um animal, no contexto de prever e explicar suas disposições comportamentais, por referência à sua aptidão evolutiva esperada. No caso de pessoas, é mais comum em economia e em aplicações da teoria dos jogos avaliar seu bem-estar relativo por referência aos seus próprios julgamentos implícitos ou explícitos a esse respeito. É por isso que nos referimos acima ao bem-estar **subjetivo**. Considere uma pessoa que adora o sabor de picles mas não gosta de cebolas. Pode-se dizer que ela associa maior utilidade a estados do mundo em que, todo o resto sendo igual, ela consome mais picles e menos cebolas do que estados em que ela consome mais cebolas e menos picles. Exemplos desse tipo sugerem que "utilidade" denota uma medida de realização **psicológica** subjetiva, e isso é de fato como o conceito foi originalmente interpretado por economistas e filósofos influenciados pelo utilitarismo de Jeremy Bentham. Contudo, no início do século XX, os economistas reconheceram cada vez mais claramente que seu principal interesse era a propriedade do mercado de demanda marginal decrescente, independentemente de ela ter sido produzida por consumidores individuais saciados ou por outros fatores. Nos anos 1930, essa motivação dos economistas se encaixou confortavelmente com o predomínio do behaviourismo e do empirismo radical em psicologia e na filosofia da ciência. Behaviouristas e empiristas radicais se opuseram ao uso teórico de entidades inobserváveis como "quocientes de realização psicológica". O clima intelectual era então receptivo aos esforços do economista Paul Samuelson (1938) de redefinir a utilidade de tal modo que se tornasse um conceito puramente técnico ao invés de um conceito desenvolvido com raízes na psicologia especulativa. Como a redefinição de Samuelson se tornou *standard* nos anos 1950, quando dizemos que um agente age de modo a maximizar sua utilidade, queremos dizer com "utilidade" seja o que for que o comportamento do agente sugere a ele agir consistentemente de modo a tornar mais provável de obter. Se isso parece circular a você, assim deveria: os

teóricos seguidores de Samuelson **entendem** o enunciado “os agentes agem de modo a maximizar sua utilidade” como uma tautologia, em que um “agente (econômico)” é qualquer entidade que pode ser precisamente descrita como agindo para maximizar uma função de utilidade, uma “ação” é qualquer seleção maximizadora de utilidade de um conjunto de alternativas possíveis, e uma “função de utilidade” é o que um agente econômico maximiza. Como outras tautologias que ocorrem nos fundamentos de teorias científicas, esse sistema emaranhado (recursivo) de definiçoes é útil não em si mesmo, mas porque ele ajuda a fixar os contextos de investigação.

Embora o behaviourismo dos anos 1930 tenha sido desde então substituído por um interesse generalizado por processos cognitivos, muitos teóricos continuam seguindo a maneira de Samuelson de compreender a utilidade porque pensam ser importante que a teoria dos jogos se aplique a **qualquer** tipo de agente - uma pessoa, um urso, uma abelha, uma empresa ou um país - e não só a agentes com mentes humanas. Quando tais teóricos dizem que os agentes agem de modo a maximizar sua utilidade, eles querem que isso seja parte da **definição** do que é ser um agente, não uma reivindicação empírica sobre possíveis estados internos ou motivações. A concepção de utilidade de Samuelson, definida por meio da **Teoria da Preferência Revelada** (TPR) introduzida em seu artigo clássico (SAMUELSON, 1938), satisfaz essa demanda.

Economistas e outros que interpretam a teoria dos jogos em termos de TPR não devem pensar na teoria dos jogos como uma explicação empírica das motivações de alguns atores de carne e osso (tal como pessoas reais). Em vez disso, eles devem considerar a teoria dos jogos como parte do corpo da matemática que é usada para modelar aquelas entidades (que podem ou não literalmente existir) que selecionam consistentemente elementos de conjuntos de ações mutuamente exclusivos, o que resulta em padrões de escolhas que podem ser modelados estatisticamente como maximização de funções de utilidade ao se permitir alguma aleatoriedade e interferência. Nessa interpretação, a teoria dos jogos não poderia ser refutada por nenhuma observação empírica, visto que ela não seria uma teoria empírica. Obviamente, a observação e a experiência poderiam levar alguém a favor dessa interpretação a concluir que a teoria dos jogos é de pouca **ajuda** para descrever o comportamento dos seres humanos

reais.

Alguns teóricos compreendem o ponto da teoria dos jogos de maneira diferente. Eles veem a teoria dos jogos como provedora de um critério explicativo para processos de raciocínio estratégico de seres humanos reais. Para que essa ideia seja aplicável, devemos supor que os agentes ao menos às vezes fazem o que fazem em cenários não-paramétricos **porque** a lógica da teoria dos jogos recomenda determinadas ações como "racionais". Tal compreensão da teoria dos jogos incorpora um aspecto **normativo**, uma vez que a "racionalidade" é tomada como denotando uma propriedade que um agente deveria, pelo menos geralmente, querer ter. Essas duas maneiras muito gerais de pensar acerca dos possíveis usos da teoria dos jogos são compatíveis com a interpretação tautológica de maximização de utilidade. Contudo, a diferença filosófica não é insignificante para o teórico dos jogos. Como veremos em uma seção posterior, aqueles que esperam usar a teoria dos jogos para explicar o **raciocínio** estratégico, em oposição ao **comportamento** meramente estratégico, enfrentam alguns problemas filosóficos e práticos especiais.

Como a teoria dos jogos é uma tecnologia para modelagem formal, devemos ter um dispositivo para tratar a maximização de utilidade em termos matemáticos. Tal dispositivo é chamado de **função de utilidade**. Introduziremos a ideia geral de uma função de utilidade por meio do caso especial de uma função de utilidade **ordinal**. (Posteriormente, encontraremos funções de utilidade que incorporam mais informações.) O mapa de utilidade para um agente é chamado de "função" porque ele mapeia as **preferências ordenadas** nos números reais. Suponha que um agente *x* prefere o pacote *a* ao pacote *b* e o pacote *b* ao pacote *c*. Em seguida, nós os mapeamos em uma lista de números, em que a função mapeia o pacote mais bem classificado no maior número na lista, o segundo pacote mais bem classificado no segundo maior número na lista, e assim por diante, então:

pacote $a \gg 3$
pacote $b \gg 2$
pacote $c \gg 1$

A única propriedade mapeada por essa função é a **ordem**. A magnitude dos números é irrelevante; isto é, não deve ser inferido que *x* recebe 3 vezes mais

utilidade do pacote *a* do que obtém do pacote *c*. Assim, poderíamos representar **exatamente a mesma** função de utilidade que a função acima por

pacote $a \gg 7.326$
pacote $b \gg 12,6$
pacote $c \gg -1.000.000$

Portanto, os números que aparecem em uma função de utilidade ordinal não estão medindo **quantidade**. Uma função de utilidade em que a magnitude **de fato** importa é chamada de "cardinal". Sempre que alguém se refere a uma função de utilidade sem especificar qual tipo é referido, deve-se assumir que ela é ordinal. Esses são os tipos de que precisamos para o primeiro conjunto de jogos que examinaremos. Posteriormente, quando chegarmos a ver como resolver jogos que envolvem (*ex ante*) incerteza - nosso jogo de atravessar o rio na Parte 1 acima, por exemplo -, precisaremos construir funções de utilidade cardinais. A técnica para fazer isso nos foi dada por von Neumann e Morgenstern (1944), e era um aspecto essencial de sua invenção da teoria dos jogos. Contudo, por agora, precisaremos apenas de funções ordinais.

2.2. Jogos e Racionalidade

Todas as situações em que ao menos um agente pode agir para maximizar sua utilidade somente por meio da antecipação (consciente ou apenas implicitamente em seu comportamento) das respostas às suas ações por um ou mais agentes diferentes são chamadas de "**jogos**". Os agentes envolvidos em jogos são chamados de "**jogadores**". Se todos os agentes possuem ações ótimizadas independentemente do que os outros façam, como em situações puramente paramétricas ou em condições de monopólio ou de competição perfeita (*vide* **Seção 1**), podemos modelá-las sem apelar à teoria dos jogos; de outro modo, precisamos da teoria.

Os teóricos dos jogos assumem que os jogadores possuem conjuntos de capacidades comumente referidas na literatura da economia como envolvendo "racionalidade". Usualmente, isso é formulado por enunciados simples tais como "é assumido que os jogadores são racionais". Na literatura crítica da economia em

geral, ou crítica da importação da teoria dos jogos nas ciências humanas, esse tipo de retórica se tornou cada vez mais um ímã para ataques. Há uma densa e intrincada teia de conexões associadas à "racionalidade" na tradição cultural ocidental, e a palavra tem sido frequentemente usada para marginalizar normativamente características tão normais e importantes quanto emoções, feminilidade e empatia. O uso do conceito por parte dos teóricos dos jogos não precisa implicar, e geralmente não implica, tal ideologia. Para os presentes propósitos, usaremos "racionalidade econômica" como um termo estritamente técnico, não normativo, que se refere a um conjunto estreito e específico de restrições sobre preferências compartilhadas pela versão original da teoria dos jogos de von Neumann e Morgenstern e pela TPR. Os economistas usam um segundo, e igualmente importante (para eles), conceito de racionalidade quando estão modelando mercados, a que chamam de "expectativas racionais". Nessa frase, "racionalidade" se refere não a restrições em preferências, mas sim às **ausências** de restrições no processamento de informações: as expectativas racionais são crenças idealizadas que refletem estatistica e precisamente o uso ponderado de toda informação disponível a um agente. O leitor deve notar que esses dois usos de uma única palavra dentro da mesma disciplina estão tecnicamente desconectados. Além disso, a TPR original tem sido especificada ao longo dos anos por vários conjuntos diferentes de axiomas para diferentes propósitos de modelagem. Uma vez que decidimos tratar a racionalidade como um conceito técnico, modificaremos efetivamente o conceito sempre que ajustarmos os axiomas. Portanto, em qualquer discussão que envolva juntamente economistas e filósofos, podemos nos encontrar em uma situação na qual diferentes participantes usam a mesma palavra para se referirem a coisas diferentes. Para leitores novatos em economia, teoria dos jogos, teoria da decisão e filosofia da ação, essa situação naturalmente apresenta um desafio.

Neste texto, "racionalidade econômica" será usado no sentido técnico compartilhado pela teoria dos jogos, microeconomia e teoria formal da decisão, como se segue. Um jogador economicamente racional é aquele que pode (i) avaliar resultados, no sentido de classificá-los e ordená-los com respeito às suas contribuições para seu bem-estar; (ii) calcular vias para resultados, no sentido de reconhecer quais sequências de ações são probabilisticamente associadas com

quais resultados; e (iii) selecionar ações de conjuntos de alternativas (o que descreveremos como "escolhendo" ações) que geram resultados de maior preferência, dadas as ações dos outros jogadores. Podemos resumir a intuição por trás disso tudo como se segue: uma entidade é modelada de forma apropriada como um agente economicamente racional na medida em que ela tem alternativas, e escolhe entre essas de uma maneira motivada, pelo menos na maioria das vezes, pelo que lhe parece melhor para seus propósitos. (Para leitores previamente familiarizados com as obras do filósofo Daniel Dennett, poderíamos equiparar a ideia de um agente economicamente racional com o tipo de entidade que Dennett caracteriza como **intencional**, e então dizer que nós podemos prever de forma apropriada o comportamento de um agente economicamente racional a partir "da postura intencional".)

A racionalidade econômica pode, em alguns casos, ser satisfeita por cálculos internos executados por um agente, e ele pode ou não estar ciente de calcular ou de ter calculado suas condições e implicações. Em outros casos, a racionalidade econômica pode simplesmente ser incorporada em disposições comportamentais construídas por seleção natural, cultural ou de mercado. Em particular, ao chamar uma ação de "escolhida", não implicamos nenhuma deliberação necessária, consciente ou de outra forma. Queremos dizer, meramente, que a ação foi tomada quando uma ação alternativa estava disponível, em algum sentido de "disponível" normalmente estabelecido pelo contexto da análise particular. ("Disponível", como usado por teóricos dos jogos e economistas, nunca deve ser lido como se quisesse dizer meramente "metafísica" ou "logicamente"disponível; a disponibilidade é quase sempre pragmática, contextual e infinitamente revisável por modelagens mais refinadas.)

Cada jogador em um jogo se depara com uma escolha dentre duas ou mais **estratégias** possíveis. Uma estratégia é um "programa de jogo" predeterminado que diz a um jogador quais ações executar em resposta a **toda estratégia possível que outros jogadores possam usar**. A importância da frase em negrito se tornará clara quando na sequência tomarmos alguns jogos como amostra.

Um aspecto crucial da especificação de um jogo envolve as informações que os jogadores possuem quando escolhem estratégias. Os jogos mais simples

(considerando a estrutura lógica) são aqueles nos quais agentes possuem **informações perfeitas**, o que significa que, em todo ponto em que cada estratégia de um agente lhe diz para executar uma ação, ele sabe tudo o que aconteceu no jogo até aquele ponto. Um jogo de tabuleiro de movimentos sequenciais no qual ambos os jogadores assistem toda a ação (e sabem as regras em comum), tal como o xadrez, é uma instância de tal jogo. Por outro lado, o exemplo do jogo de atravessar a ponte da **Seção 1** acima ilustra um jogo de **informações imperfeitas**, uma vez que o fugitivo deve escolher uma ponte para atravessar sem saber a ponte em que o perseguidor escolheu esperar, e o perseguidor, similarmente, toma sua decisão em ignorância quanto às escolhas de sua presa. Uma vez que a teoria dos jogos é acerca da ação economicamente racional dadas as ações estrategicamente importantes de outros, não gera surpresa dizer que o que os agentes nos jogos acreditam, ou falham em acreditar, sobre as ações uns dos outros faça uma diferença considerável para a lógica de nossa análise, como veremos.

2.3. Árvores e Matrizes

A diferença entre jogos de informações perfeitas e jogos de informações imperfeitas está relacionada (embora certamente não seja idêntica!) com uma distinção entre **maneiras de representar** jogos que é baseada na **ordem de jogo**. Comecemos por distinguir entre jogos de movimento sequencial e jogos de movimento simultâneo em termos de informações. É natural, como uma primeira aproximação, pensar em jogos de movimento sequencial como aqueles em que os jogadores escolhem suas estratégias um após o outro, e em jogos de movimento simultâneo como aqueles em que os jogadores escolhem suas estratégias ao mesmo tempo. Contudo, isso não seria estritamente correto porque o que é de importância estratégica não é a **ordem** temporal em si, mas se e quando os jogadores **sabem sobre** as ações dos demais no momento de escolher as suas próprias ações. Por exemplo, se duas firmas concorrentes estão planejando campanhas de marketing, uma delas poderia se comprometer com sua estratégia meses antes da outra; contudo, se nenhuma delas sabe com o que a outra se comprometeu ou com o que se comprometerá quando tomarem suas

decisões, esse é um jogo de movimento simultâneo. O xadrez, por outro lado, é normalmente jogado como um jogo de movimento sequencial: você vê o que o seu oponente fêz antes de escolher a sua próxima ação. (O xadrez **pode** ser transformado em um jogo de movimento simultâneo se cada um dos jogadores fizer movimentos em um tabuleiro comum estando isolado do outro; mas esse é um jogo muito diferente do xadrez convencional.)

Foi dito acima que a distinção entre jogos de movimento sequencial e jogos de movimento simultâneo não é idêntica à distinção entre jogos de informações perfeitas e jogos de informações imperfeitas. Explicar por que isso é assim é uma boa maneira de estabelecer uma compreensão completa de ambos os conjuntos de conceitos. Como jogos de movimento simultâneo foram caracterizados no parágrafo anterior, tem de ser verdadeiro que todos os jogos de movimento simultâneo são jogos de informações imperfeitas. Contudo, alguns jogos podem conter combinações de movimentos sequenciais e movimentos simultâneos. Por exemplo, duas firmas poderiam se comprometer com suas estratégias de marketing de maneira independente e em segredo uma da outra, mas, depois disso, se envolverem numa competição de preços escancarada. Se as estratégias de marketing ideais fossem parcial ou inteiramente dependentes do que se esperava que acontecesse no subsequente jogo de preços, então os dois estágios precisariam ser analisados como um único jogo, no qual um estágio de jogo sequencial se seguiu a um estágio de jogo simultaneo. Jogos inteiros que envolvem estágios mistos desse tipo são jogos de informações imperfeitas, não importa o quão temporalmente distantes esses estágios possam estar uns dos outros. Jogos de informações perfeitas (como o próprio nome já diz) denotam casos em que **nenhum** movimento é simultâneo (e nenhum jogador se esquece do que aconteceu anteriormente).

Como notado previamente, os jogos de informações perfeitas são os tipos de jogos (logicamente) mais simples. Isso é assim porque em tais jogos (na medida em que os jogos são finitos, isto é, terminam após um número conhecido de ações) jogadores e analistas podem usar um procedimento simples para prever resultados. Um jogador em tal jogo escolhe sua primeira ação considerando cada série de respostas e contrarrespostas que resultarão de cada ação aberta a ele. Ele então se pergunta qual dos resultados finais disponíveis lhe

traz a maior utilidade, e escolhe a ação que leva a esse resultado ao iniciar a cadeia. Esse processo é chamado de **indução reversa** (porque o raciocínio funciona de trás para frente a partir resultados eventuais para problemas de escolha atuais).

Haverá muito mais a ser dito sobre a indução reversa e suas propriedades em uma seção posterior (quando chegarmos a discutir o equilíbrio e a seleção de equilíbrio). Por enquanto, ela é descrita apenas para que possamos utilizá-la para introduzir um dos dois tipos de objetos matemáticos usados para representar jogos: as **árvores de jogos**. Uma árvore de jogo é um exemplo do que os matemáticos chamam de **grafo direto**. Isto é, ele é um conjunto de nós conectados em que o grafo como um todo possui uma direção. Podemos desenhar árvores do topo da página para o final, ou da esquerda para a direita. No primeiro caso, nós do topo da página são interpretados como vindo antes na sequência de ações. No caso de uma árvore desenhada da esquerda para a direita, os nós da esquerda são anteriores na sequência aos da direita. Uma árvore não rotulada possui uma estrutura do seguinte tipo:

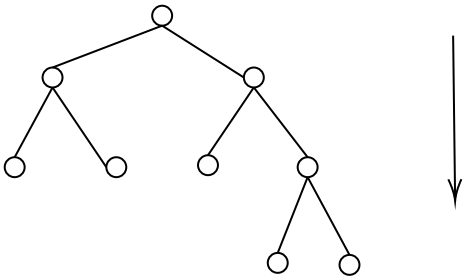


Figura 1

O ponto de representar jogos usando árvores pode ser melhor compreendido visualizando o seu uso para apoiar o raciocínio de indução reversa. Apenas imagine o jogador (ou o analista) começando pelo final da árvore, em que os resultados são exibidos, e, em seguida, trabalhando de trás para frente a partir desses resultados, procurando por conjuntos de estratégias que descrevam caminhos que levam a eles. Uma vez que a função de utilidade do jogador indica

quais resultados ele prefere a outros, nós também sabemos quais caminhos ele preferirá. Obviamente, nem todos os caminhos serão possíveis porque o outro jogador também tem um papel na seleção de caminhos, e não tomará ações que levam a resultados menos preferidos por ele. Apresentaremos alguns exemplos dessa seleção de caminhos interativa e de técnicas detalhadas para raciocinar através desses exemplos depois de termos descrito uma situação que pode ser modelada por uma árvore.

Árvores são usadas para representar jogos **sequenciais**, porque mostram a ordem em que as ações são tomadas pelos jogadores. Contudo, os jogos são por vezes representados em **matrizes** em vez de árvores. Esse é o segundo tipo de objeto matemático usado para representar jogos. Diferentemente de árvores, as matrizes mostram simplesmente os resultados, representados em termos das funções de utilidade dos jogadores, para toda combinação possível de estratégias que os jogadores poderiam usar. Por exemplo, faz sentido exibir o jogo de atravessar o rio da **Seção 1** em uma matriz, uma vez que, nesse jogo, o fugitivo e o perseguidor possuem ambos só um movimento cada, e cada um deles escolhe seu movimento ignorando o que o outro decide fazer. Esta é **uma parte** da matriz:

		Perseguidor		
		Ponte Segura	Ponte das Pedras	Ponte das Cobras
Fugitivo	Ponte Segura	0,1	1,0	1,0
	Ponte das Pedras	?	0,1	?
	Ponte das Cobras	?	?	0,1

Figura 2

As três estratégias possíveis do fugitivo - atravessar pela ponte segura, arriscar as pedras, ou arriscar as cobras - formam as linhas da matriz. Similarmente, as três estratégias possíveis do perseguidor - esperar na ponte segura, esperar na ponte com pedras e esperar na ponte com cobras - formam as colunas da matriz. Cada célula da matriz mostra - ou, melhor, **mostraria** caso a nossa matriz estivesse completa - um **resultado** definido em termos das **recompensas** [*payoffs*] dos jogadores. A recompensa de um jogador é simplesmente o número

designado por sua função de utilidade ordinal ao estado de coisas correspondente ao resultado em questão. Para cada resultado, as recompensas das linhas são sempre listadas primeiro, seguidas pelas recompensas das colunas. Assim, por exemplo, o canto superior esquerdo da matriz acima mostra que quando o fugitivo atravessa a ponte segura e o perseguidor está esperando por ele nessa ponte, o fugitivo obtém uma recompensa de 0 e o perseguidor obtém uma recompensa de 1. Interpretamos isso por referência às funções de utilidade dos dois jogadores, que neste jogo são muito simples. Se o fugitivo conseguir atravessar o rio com segurança, ele recebe a recompensa de 1; se ele não conseguir, obtém 0. Se o fugitivo não conseguir atravessar o rio, seja porque foi baleado pelo perseguidor ou atingido por uma pedra ou mordido por uma cobra, então o perseguidor obtém uma recompensa de 1 e o fugitivo obtém uma recompensa de 0.

Explicaremos brevemente as partes da matriz que foram preenchidas e por que ainda não conseguimos completá-la. Sempre que o perseguidor esperar na ponte escolhida pelo fugitivo, o fugitivo é baleado. Todos esses resultados produzem o vetor de recompensa (0,1). Você pode encontrá-los descendo diagonalmente através da matriz a partir do canto superior esquerdo. Sempre que o fugitivo escolher a ponte segura mas o perseguidor esperar em outra, o fugitivo consegue atravessar em segurança, o que produz o vetor de recompensa (1,0). Esses dois resultados são mostrados nas duas segundas células da linha superior. **Por enquanto**, todas as outras células estão marcadas com pontos de interrogação. Por quê? O problema aqui é este: se o fugitivo atravessar a ponte com pedras ou a ponte com cobras, ele introduz fatores paramétricos no jogo. Nesses casos, ele assume algum risco de ser morto, e assim produz o vetor de recompensa (0,1), que é independente de qualquer coisa que o perseguidor faça. Ainda não introduzimos conceitos que são suficientes para sermos capazes representar esses resultados em termos de funções de utilidade - mas nós o faremos até que tenhamos terminado, e isso fornecerá a chave para solucionar o nosso enigma da **Seção 1**.

Jogos de matrizes são referidos como jogos de “formato normal” ou de “formato estratégico”, e jogos como árvores são referidos como jogos de “formato extensivo”. Os dois tipos de jogos não são equivalentes, pois os jogos de formato extensivo contêm informações - sobre sequências de jogo e níveis de

informações dos jogadores sobre a estrutura do jogo - que os jogos de formato estratégico não contêm. Em geral, um jogo de formato estratégico poderia representar qualquer um dos vários jogos de formato extensivo, portanto um jogo de formato estratégico é melhor concebido como sendo um **conjunto** de jogos de formato extensivo. Quando a ordem de jogo é irrelevante para o seu resultado, então você deve estudar seu formato estratégico, uma vez que é sobre todo o conjunto que você quer saber. Quando a ordem de jogo **é** relevante, o formato extensivo **deve** ser especificado ou suas conclusões não serão confiáveis.

2.4. O Dilema do Prisionieiro como um exemplo de representação de formato estratégico vs representação de formato extensivo

As distinções descritas acima são difíceis de se compreender inteiramente caso tudo o que se tenha à disposição sejam descrições abstratas. Elas são melhor ilustradas por meio de um exemplo. Para esse propósito, usaremos o mais famoso de todos os jogos: o Dilema do Prisioneiro. De fato, ele fornece a lógica do problema deparado pelos soldados de Cortez e Henrique V (*vide* **Seção 1**), e pelos agentes de Hobbes antes de eles darem poderes ao tirano. Contudo, por razões que se tornarão claras um pouco mais tarde, você não deve tomar o DP como um jogo **comum**; ele não o é. Nós o usamos aqui como um exemplo apenas porque ele é particularmente útil para ilustrar a **relação** entre jogos de formato estratégico e jogos de formato extensivo (e, posteriormente, para ilustrar as relações entre jogos de lance único [*one-shot games*] e jogos repetidos; *vide* **Seção 4**).

O nome do jogo do Dilema do Prisioneiro é derivado da seguinte situação comumente utilizada para exemplificá-lo. Suponha que a polícia tenha prendido duas pessoas que ela sabe terem cometido um assalto à mão armada juntos. Infelizmente, faltam evidências admissíveis suficientes para obter a condenação pelo júri. No entanto, a polícia **de fato** possui evidência suficiente para mandar para a prisão cada prisioneiro por dois anos pelo furto do carro utilizado na fuga. O delegado faz agora a seguinte oferta para cada prisioneiro: se você confessar o assalto, comprometendo seu parceiro, e ele não confessar, então você ficará livre e ele pegará dez anos de prisão. Se vocês dois

confessarem, cada um de vocês pegará 5 anos. Se nenhum de vocês confessar, então cada um de vocês pegará dois anos pelo roubo do carro.

Nosso primeiro passo para modelar a situação dos dois prisioneiros como um jogo é representá-la em termos de funções de utilidade. Seguindo-se a convenção usual, vamos nomear os prisioneiros "Jogador I" e "Jogador II". Ambas as funções de utilidade ordinal são idênticas:

Sair impune $\gg 4$
2 anos $\gg 3$
5 anos $\gg 2$
10 anos $\gg 0$

Os números na função acima são agora utilizados para expressar as **recompensas** de cada jogador nos vários resultados possíveis na situação. Nós podemos representar o problema enfrentado por ambos em uma única matriz que mostra a maneira como suas escolhas separadas interagem; este é o formato estratégico do jogo:

		Jogador II	
		Confessa	Recusa
Jogador I	Confessa	2,2	4,0
	Recusa	0,4	3,3

Figura 3

Cada célula da matriz fornece as recompensas a ambos os jogadores para cada combinação de ações. A recompensa do Jogador I aparece como o primeiro número de cada par, e a do Jogador II como o segundo. Assim, se ambos os jogadores confessarem, cada um deles receberá uma recompensa de 2 (5 anos de prisão cada). Isso aparece na célula superior esquerda. Se nenhum deles confessar, cada um deles receberá uma recompensa de 3 (2 anos de prisão cada). Isso aparece como a célula inferior direita. Se o Jogador I confessa e o Jogador II recusa-se a confessar, então o Jogador I recebe uma recompensa de 4 (saindo impune) e o Jogador II recebe uma recompensa de 0 (dez anos de prisão). Isso aparece na célula superior direita. A situação inversa, em que o Jogador II confessa e o Jogador I se recusa a confessar, aparece na célula

superior esquerda.

Cada jogador avalia suas duas ações possíveis comparando suas recompensas pessoais em cada coluna, uma vez que isso mostra qual das ações é preferível, só para cada um, em cada lance possível do parceiro. Assim, observe: se o Jogador II confessar, então o Jogador I obtém uma recompensa de 2 por confessar e uma recompensa de 0 por se recusar. Se o Jogador II se recusa, então o Jogador I obtém uma recompensa de 4 por confessar e uma recompensa de 3 por se recusar. Portanto, o Jogador I se sai melhor confessando independentemente do que o Jogador II faça. O jogador II, enquanto isso, avalia suas ações comparando suas recompensas em cada linha, e ele chega exatamente à mesma conclusão que o Jogador I. Sempre que uma ação para um jogador for superior às suas outras ações para cada ação possível do oponente, dizemos que a primeira ação **domina estritamente** a segunda. Para ambos os jogadores no Dilema do Prisioneiro, a confissão domina estritamente a recusa. Cada jogador tem conhecimento disso a respeito do outro, eliminando-se, assim, qualquer tentação de se afastar do caminho estritamente dominado. Desse modo, ambos os jogadores confessarão, e ambos irão para a prisão por 5 anos.

Os jogadores, e analistas, podem prever esse resultado usando um procedimento mecânico, conhecido como eliminação iterativa de estratégias estritamente dominadas. O Jogador I pode ver, ao examinar a matriz, que suas recompensas em cada célula da linha superior são maiores que suas recompensas em cada célula correspondente da linha inferior. Portanto, nunca pode ser um maximizador de utilidade para ele jogar a estratégia da linha inferior, ou seja, se recusar a confessar, **independentemente** do que faça o Jogador II. Como a estratégia da linha inferior do Jogador I nunca será jogada, podemos simplesmente **deletar** a linha inferior da matriz. Agora, é óbvio que o Jogador II não se recusará a confessar, uma vez que sua recompensa por confessar nas duas células restantes é maior que a sua recompensa por se recusar. Assim, novamente, podemos deletar a coluna de uma célula do jogo. Temos agora apenas uma célula remanescente, aquela que corresponde ao resultado oriundo da confissão mútua. Como o raciocínio que nos levou a deletar todos os outros resultados possíveis dependeu em cada passo apenas da premissa de que ambos os jogadores são economicamente racionais - isto é, eles escolherão

estratégias que levam a maiores recompensas em vez de estratégias que levam a menores recompensas -, há fortes razões para ver a confissão mútua como a **solução** para o jogo, o resultado para o qual suas jogadas **devem** convergir na medida em que a racionalidade econômica modela corretamente o comportamento dos jogadores. Você deve notar que a ordem em que linhas e colunas estritamente dominadas são deletadas não importa. Caso tivéssemos começado por deletar a coluna da direita e então deletado a linha inferior, teríamos chegado à mesma solução.

Tem-se dito algumas vezes que o DP não é um jogo comum sob muitos aspectos. Um desses aspectos é que todas as suas linhas e colunas são ou estritamente dominadas ou estritamente dominantes. Em qualquer jogo de formato estratégico em que isso é verdadeiro, a eliminação iterativa de estratégias estritamente dominadas garante a produção de uma única solução. No entanto, posteriormente, veremos que para muitos jogos essa condição não se aplica, e então nossa tarefa analítica será menos direta.

O leitor provavelmente terá notado algo perturbador acerca do resultado do Dilema do Prisioneiro. Caso os dois jogadores tivessem se recusado a confessar, eles teriam chegado ao resultado inferior direito em que cada um dos dois vai para a prisão por apenas 2 anos, portanto **ambos** ganhariam maior utilidade do que qualquer um receberia quando ambos confessam. Esse é o fato mais importante sobre o DP e sua importância para a teoria dos jogos é bastante geral. Retornaremos a ele quando discutirmos os conceitos de equilíbrio na teoria dos jogos. Todavia, por enquanto, vamos permanecer com nosso uso desse jogo em particular para ilustrar a diferença entre formatos estratégico e extensivo.

Quando o DP é introduzido em discussões populares, frequentemente se ouve dizer que o delegado de polícia deve trancar seus prisioneiros em salas separadas de tal modo que eles não possam comunicar entre si. O raciocínio por trás dessa ideia parece óbvio: se os jogadores pudessem comunicar entre si, eles veriam que cada um deles se sairia melhor caso ambos se recusassem a confessar, e poderiam entrar em acordo para fazer exatamente isso, não? Presume-se que isso removeria a convicção de cada jogador de que ele ou ela deve confessar porque, caso contrário, cada um deles seria traído pelo outro. Na verdade, contudo, essa intuição é enganadora e sua conclusão é falsa.

Quando representamos o DP como um jogo de formato estratégico, assumimos implicitamente que os prisioneiros não podem tentar um acordo de conluio, visto que eles escolhem suas ações simultaneamente. Nesse caso, um acordo antes do fato não pode ajudar. Se o Jogador I está convencido de que seu parceiro cumprirá o acordo, então ele pode aproveitar a oportunidade para sair ileso ao se confessar. Obviamente, ele percebe que a mesma tentação ocorrerá ao Jogador II; porém, nesse caso, ele novamente terá certeza de que confessará, pois esse é o seu único meio de evitar seu pior resultado. O acordo dos prisioneiros é mal sucedido porque eles não têm maneiras de fazê-lo ser cumprido; as promessas um ao outro constituem o que os teóricos dos jogos chamam de "conversa fiada".

Todavia, suponha que os prisioneiros **não** se movimentam simultaneamente. Isto é, suponha que o Jogador II pode escolher **depois** de observar a ação do Jogador I. Esse é o tipo de situação que as pessoas que pensam que a não comunicação é importante devem ter em mente. Agora, o Jogador II será capaz de ver que o Jogador I permaneceu firme em sua escolha e não precisa se preocupar em ser enganado. Contudo, isso não muda nada, um ponto que é melhor colocado quando representamos novamente o jogo sob o formato extensivo. Isso nos dá a oportunidade de introduzir árvores de jogos e o método de análise apropriado.

Primeiramente, contudo, aqui estão as definições de alguns conceitos que serão úteis para analisar as árvores de jogos:

- **Nó:** um ponto em que um jogador escolhe uma ação.
- **Nó inicial**: o ponto em que a primeira ação no jogo ocorre.
- **Nó final:** qualquer nó que, se alcançado, finaliza o jogo. Cada nó final corresponde a um **resultado**.
- **Subjogo:** qualquer conjunto conectado de nós e ramos que descendem unicamente de um nó.
- **Recompensa** [*payoff*]**:** um número de utilidade ordinal atribuído a um jogador em um resultado.
- **Resultado:** uma atribuição de um conjunto de recompensas, uma para cada jogador no jogo.
- **Estratégia:** um programa que instrui um jogador sobre qual ação tomar em

cada nó na árvore no qual ele poderia ser chamado para fazer uma escolha.

Essas rápidas definições podem não significar muito para você até que as veja colocadas em uso na nossa análise das árvores abaixo. Provavelmente, será melhor repetidamente ir aos exemplos e voltar às definições à medida em que avançamos. No momento em que você entender cada exemplo, você verá que os conceitos e suas definições são naturais e intuitivos.

Para tornar esse exercício o mais instrutivo possível, vamos supor que os Jogadores I e II tenham estudado a matriz acima e tenham feito um acordo para cooperar ao verem que ambos se saem melhor no resultado representado pela célula inferior direita. O Jogador I se comprometerá com uma recusa primeiro, após a qual o Jogador II lhe retribuirá a recusa quando a polícia pedir sua escolha. Vamos chamar a estratégia de manter o acordo de "cooperação", e a denotaremos na árvore abaixo com "C". A estratégia de quebrar o acordo será chamada de "defecção", e a denotaremos na árvore abaixo com "D". Cada nó é enumerado 1, 2, 3, ..., de cima para baixo, para facilitar a referência na discussão. Eis então a árvore:

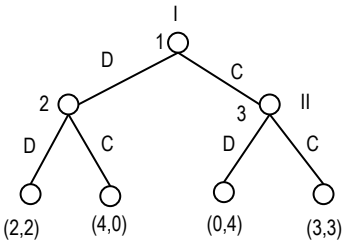


Figura 4

Olhe primeiro para cada um dos nós finais (aqueles ao longo da parte inferior). Eles representam resultados possíveis. Cada um dos nós é identificado com uma atribuição de recompensas, assim como no jogo de formato estratégico, com a recompensa do Jogador I aparecendo primeiro em cada conjunto e a do Jogador II aparecendo em segundo. Cada uma das estruturas descendentes dos nós 1, 2 e 3, respectivamente, são um subjogo. Começamos a nossa análise de indução reversa - usando a técnica chamada de **algoritmo de Zermelo** - com os subjogos

que surgem por último na sequência de jogo. Se o subjogo descendente do nó 3 é jogado, então o Jogador II enfrentá uma escolha entre uma recompensa de 4 e uma recompensa de 3. (veja o segundo número, representando sua recompensa, em cada conjunto em um nó final descendente do nó 3.) II recebe sua recompensa mais alta jogando D. Portanto, podemos substituir todo o subjogo por uma atribuição da recompensa (0,4) diretamente ao nó 3, uma vez que esse é o resultado que será realizado se o jogo atingir esse nó. Agora, considere o subjogo descendente do nó 2. Aqui, II se depara com uma escolha entre uma recompensa de 2 e uma de 0. Ele obtém sua recompensa mais alta, 2, ao jogar D. Portanto, podemos atribuir a recompensa (2,2) diretamente ao nó 2. Movemo-nos agora para o subjogo descendente do nó 1. (Esse subjogo é, obviamente, idêntico a todo o jogo; todos os jogos são subjogos de si mesmos.) O Jogador I se depara com uma escolha entre os resultados (2,2) e (0,4). Consultando os primeiros números em cada um desses conjuntos, ele vê que consegue sua recompensa mais alta - 2 - ao jogar D. Obviamente, D é a opção de confessar. Assim, o Jogador I confessa, e então o Jogador II também confessa, produzindo o mesmo resultado que na representação de formato estratégico.

Intuitivamente, o que aconteceu aqui é que o Jogador I percebe que se ele jogar C (se recusar a confessar) no nó 1, então o Jogador II será capaz de maximizar sua utilidade enganando-o e jogando D. (Na árvore, isso acontece no nó 3.) Isso deixa o Jogador I com uma recompensa de 0 (dez anos de prisão), o que, para começo de conversa, ele pode evitar apenas ao jogar D. Desse modo, ele abre mão do acordo.

Vimos assim que as versões simultânea e sequencial produzem o mesmo resultado no caso do Dilema do Prisioneiro . Mas isso muitas vezes não será verdadeiro para outros jogos. Além disso, apenas jogos (sequenciais) finitos de formato extensivo de informações perfeitas podem ser resolvidos usando o algoritmo de Zermelo.

Como notado anteriormente nessa seção, às vezes devemos representar movimentos simultâneos **dentro** de jogos que, contrariamente, são sequenciais. (Em todos esses casos, o jogo como um todo será de informações imperfeitas e não seremos capazes de resolvê-lo usando o algoritmo de Zermelo.) Representamos tais jogos usando o dispositivo de **conjuntos informacionais**.

Considere a seguinte árvore:

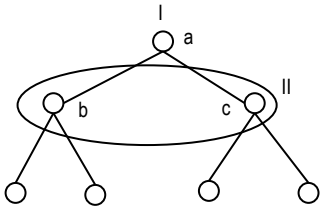


Figura 5

A elipse desenhada ao redor dos nós *b* e *c* indica que elas residem dentro de um conjunto informacional comum. Isso significa que, nesses nós, os jogadores não podem inferir de volta o caminho pelo qual vieram; ao escolher sua estratégia, o Jogador II não sabe se ele está em *b* ou em *c*. (Por essa razão, o que propriamente contém números em jogos de formato extensivo são conjuntos informacionais, concebidos como "pontos de ação", em vez de nós em si mesmos; é por isso que os nós dentro da elipse são designados com letras em vez de números.) Colocado de outro modo, ao escolher, o Jogador II não sabe o que o Jogador I fêz no nó *a*. Você deve se se lembrar que isso é justamente o que define dois movimentos como simultâneos. Desse modo, podemos ver que o método de representar jogos como árvores é inteiramente geral. Se nenhum nó após o nó inicial está sozinho em um conjunto informacional na sua árvore, de tal modo que o jogo tenha apenas um subjogo (ele mesmo), então o jogo como um todo é um jogo simultâneo. Se pelo menos um nó compartilha seu conjunto informacional com outro, enquanto os outros estão sozinhos, o jogo envolve ambos os jogos simultâneo e sequencial, e portanto ainda é um jogo de informações imperfeitas. Teremos um jogo de informações perfeitas somente se todos os conjuntos informacionais forem habitados por apenas um nó.

2.5. Conceitos de Solução e Equilíbrio

No Dilema do Prisioneiro, o resultado de defecção mútua, que representamos como (2,2), foi dito ser a "solução" para o jogo. Seguindo a prática geral em economia, os teóricos dos jogos se referem às soluções de jogos como **equilíbrio**. Os leitores com mentalidade filosófica poderão colocar esta questão conceitual: Em alguns resultados de jogos, o que está "em equilíbrio" de modo que somos motivados a chamá-los de "soluções"? Quando dizemos que um sistema físico está em equilíbrio, queremos dizer que ele está em um estado **estável**, aquele em que todas as forças causais internas ao sistema se equilibram mutuamente e então o deixam "em repouso" até que, e a não ser que, ele seja perturbado pela intervenção de alguma força exógena (isto é, "externa"). Isso é o que os economistas tradicionalmente querem dizer ao falar em "equilíbrio"; eles leem sistemas econômicos como redes de relações mutuamente restritivas (frequentemente causais), assim como sistemas físicos, e o equilíbrio de tais sistemas são, desse modo, seus estados estáveis de modo endógeno. (Note que, tanto em sistemas físicos quanto em sistemas econômicos, estados estáveis de modo endógeno podem nunca ser diretamente observados porque os sistemas em questão nunca estão isolados de influências exógenas que os movem e os desestabilizam. Tanto na mecânica clássica quanto na economia, os conceitos de equilíbrio são **ferramentas para análise**, não previsões do que esperamos observar.) Como veremos em seções posteriores, é possível manter esse entendimento de equilíbrio no caso da teoria dos jogos. Contudo, como notamos na **Seção 2.1**, algumas pessoas interpretam a teoria dos jogos como sendo uma teoria explicativa do raciocínio estratégico. Para elas, uma solução para um jogo deve ser um resultado que um agente racional preveria **ao usar somente os mecanismos de cálculo racional**. Tais teóricos se deparam com alguns enigmas acerca dos conceitos de solução que são menos importantes para o teórico que não tenta usar a teoria dos jogos para subscrever uma análise geral da racionalidade. Frequentemente, o interesse de filósofos em teoria dos jogos é motivado mais por essa ambição do que o do economista ou outro cientista.

É útil começar a discussão a partir do caso do Dilema do Prisioneiro porque ele é extraordinariamente simples da perspectiva dos enigmas sobre

conceitos de solução. O que nos referimos como a sua "solução" é o único **equilíbrio de Nash** do jogo. (o "Nash" aqui se refere a John Nash, o matemático ganhador do prêmio Nobel que em (NASH, 1950) fez o máximo para estender e generalizar o trabalho pioneiro de Neumann e Morgenstern.) O equilíbrio de Nash (daqui em diante, "EN") se aplica (ou falha em se aplicar, conforme o caso) a **conjuntos** inteiros de estratégias, um para cada jogador em um jogo. Um conjunto de estratégias é um EN apenas no caso de nenhum jogador poder melhorar sua recompensa pela alteração de sua estratégia, dadas as estratégias de todos os outros jogadores no jogo. Note o quão intimamente essa ideia está relacionada à ideia de dominância estrita: nenhuma estratégia poderia ser uma estratégia EN se ela for estritamente dominada. Portanto, se a eliminação iterativa de estratégias estritamente dominadas nos levar a um único resultado, sabemos que o vetor de estratégias que leva a ele é o único EN do jogo. Agora, quase todos os teóricos concordam que evitar estratégias estritamente dominadas é um requerimento **mínimo** da racionalidade econômica. Um jogador que conscientemente escolhe uma estratégia estritamente dominada viola diretamente a cláusula (iii) da definição de agência econômica dada na **Seção 2.2**. Isso implica que **se** um jogo possui um resultado que é um único EN, como no caso da confissão mútua no DP, essa deve ser sua única solução. Esse é um dos aspectos mais importantes pelos quais o DP é um jogo "fácil"(e incomum).

Podemos especificar uma classe de jogos em que EN é sempre, não só necessário, mas também **suficiente** como um conceito de solução. Há jogos de informações perfeitas finitos que também são **soma zero**. Um jogo de soma zero (no caso de um jogo que envolve apenas dois jogadores) é aquele em que um jogador pode se sair melhor ao fazer o outro jogador ficar pior. (O jogo da velha é um exemplo simples de tal jogo: qualquer movimento que aproxima um jogador da vitória aproxima o seu oponente da derrota, e vice-versa.) Podemos determinar se um jogo é de soma zero examinando as funções de utilidade dos jogadores: em jogos de soma zero, elas serão imagens espelhadas uma da outra, com os resultados altamente classificados de um jogador sendo de baixa classificação para o outro, e vice-versa. Em um jogo assim, se estou jogando uma estratégia tal que, dada a sua estratégia, eu não possa fazer nada melhor, e se você **também** está jogando tal estratégia, então, uma vez que qualquer mudança de estratégia

feita por mim teria que fazê-lo ficar pior, e vice-versa, segue-se que o nosso jogo não pode ter solução compatível com a nossa racionalidade econômica mútua além de seu único EN. Podemos colocar isso de outra maneira: em um jogo de soma zero, eu jogar uma estratégia que maximiza minha recompensa mínima se você jogar o melhor que puder, e você simultaneamente fazer a mesma coisa, é **equivalente** a nós dois jogarmos nossas melhores estratégias, e, desse modo, garante-se que esse par de assim chamados procedimentos "máximo-mínimos" encontrarão a única solução para o jogo, que é o seu único EN. (No jogo da velha, esse é um empate. Você não pode fazer nada melhor do que empatar, e nem eu, se estamos ambos tentando ganhar e tentando não perder.)

Contudo, a maioria dos jogos não possui essa propriedade. Nesse artigo, não será possível enumerar **todas** as maneiras que jogos podem ser problemáticos da perspectiva de suas soluções possíveis. (Seja pelo único motivo que é altamente improvável que os teóricos já tenham descoberto todos os problemas possíveis.) No entanto, podemos tentar generalizar um pouco as questões.

Primeiramente, há o problema de que na maioria dos jogos que não são de soma zero há mais de um EN, mas nem todos os EN parecem igualmente plausíveis como soluções que jogadores estrategicamente alertas encontrariam. Considere o jogo de formato estratégico abaixo (retirado de KREPS, 1990, p. 403):

		II	
		t1	t2
I	s1	10,10	0,0
	s2	0,0	1,1

Figura 6

Esse jogo tem dois EN: s1-t1 e s2-t2. (Note que nenhuma linha ou coluna é estritamente dominante. Mas se o Jogador I está jogando s1, então o Jogador II não pode fazer melhor do que t1, e vice-versa; e similarmente para o par s2-t2.) Se EN for nosso único conceito de solução, então seremos forçados a dizer que qualquer um desses resultados é igualmente persuasivo como solução. Contudo, se a teoria dos jogos é considerada uma teoria explicativa e/ou normativa do raciocínio estratégico, isso parece deixar algo de fora: jogadores sensatos com

informações perfeitas convergiriam certamente para s1-t1? (Note que isso **não** é como a situação no DP, em que a situação socialmente superior não é atingível porque ela não é um EN. No caso do jogo acima, ambos os jogadores têm todos os motivos para tentar convergir no EN em que se dão melhor.)

Isso ilustra o fato de que EN é um conceito de solução relativamente **fraco** (logicamente), que falha de modo frequente em prever soluções intuitivamente sensatas porque, caso aplicado sozinho, não permite que os jogadores usem princípios de seleção de equilíbrio que, se não forem **exigidos** pela racionalidade econômica - ou o conceito de racionalidade de um filósofo mais ambicioso -, ao menos parecem ser sensatos e computacionalmente acessíveis. Considere outro exemplo de (KREPS, 1990, p.397):

		II	
		t1	t2
I	s1	10,0	5,2
	s2	10,1	2,0

Figura 7

Aqui, nenhuma estratégia domina estritamente a outra. Contudo, a linha superior do Jogador I, s1, domina **fracamente** s2, uma vez que I está ao menos tão bem ao usar tanto s1 quanto s2 para qualquer resposta do Jogador II, e em uma resposta de II (t2), I se dá melhor com s1. Desse modo, não deveriam os jogadores (e o analista) deletar a linha fracamente dominada s2? Quando eles o fazem, a coluna t1 é, assim, estritamente dominada, e o EN s1-t2 é selecionado como a única solução. Contudo, como Kreps continua a mostrar ao usar esse exemplo, a ideia de que estratégias fracamente dominadas deveriam ser deletadas assim como aquelas estritamente dominadas possui consequências estranhas. Suponha que alteremos as recompensas do jogo apenas um pouco, como se segue:

		II	
		t1	t2
I	s1	10,0	5,2
	s2	10,11	2,0

Figura 8

s2 é ainda fracamente dominada; contudo, dos nossos dois EN, s2-t1 é agora o mais atrativo para ambos os jogadores; então por que o analista deveria eliminar sua possibilidade? (Note que esse jogo, novamente, **não** replica a lógica do DP. Lá, faz sentido eliminar o resultado mais atrativo, a recusa conjunta em confessar, porque ambos os jogadores possuem incentivos para se desviar unilateralmente dele, portanto ele não é um EN. Isso não é verdadeiro sobre s2-t1 no presente jogo. Você deve estar começando a ver claramente porque chamamos o jogo do DP de "incomum".) O argumento **para** eliminar estratégias fracamente dominadas é que o Jogador I pode estar nervoso, temendo que não haja completa **certeza** de que o Jogador II seja economicamente racional (ou que o Jogador II teme que o Jogador I não seja completa e confiavelmente economicamente racional, ou que o Jogador II teme que o Jogador I teme que o Jogador II não seja completa e confiavelmente economicamente racional, e assim *ad infinitum*) e possa jogar t2 com alguma probabilidade positiva. Se a possibilidade de afastamento da confiável racionalidade econômica é tomada seriamente, então temos um argumento para eliminar estratégias fracamente dominadas: desse modo, o Jogador I se assegura contra seu pior resultado, s2-t2. É claro que ele paga um custo por esse seguro, reduzindo sua recompensa esperada de 10 para 5. Por outro lado, nós poderíamos imaginar que os jogadores pudessem se comunicar antes de jogar o jogo e concordar em jogar **estratégias correlacionadas** a fim de se **coordenarem** em s2-t1, removendo, por esse meio, alguma, a maior parte ou toda a incerteza que encoraja a eliminação da linha fracamente dominada s1, eliminando, em vez disso, s1-t2 como uma solução viável!

Qualquer princípio proposto para solucionar jogos que possam ter o efeito de eliminar um ou mais EN de consideração como solução é referido como um **refinamento** do EN. No caso que acabamos de discutir, a eliminação de estratégias fracamente dominadas é um possível refinamento, uma vez que ele

remove o EN s2-t1, e a correlação é outro, uma vez que remove o outro EN, s1-t2.Desse modo, qual refinamento é mais apropriado como um conceito de solução? As pessoas que pensam a teoria dos jogos enquanto teoria explicativa e/ou normativa da racionalidade estratégica têm produzido uma literatura substancial em que os méritos e as desvantagens de um grande número de refinamentos são discutidos. Em princípio, não parece haver limite no número de refinamentos que poderiam ser considerados, visto que pode também não haver limite no conjunto de intuições filosóficas sobre quais princípios um agente racional poderia ou não achar adequado seguir, ou temer, ou esperar que os outros jogadores estejam seguindo.

Façamos agora uma breve digressão para observar algo sobre a terminologia. Os teóricos que adotam a interpretação da preferência revelada das funções de utilidade em teoria dos jogos são por vezes referidos na literatura da filosofia da economia como "behavioristas". Isso reflete o fato de que as abordagens de preferência revelada igualam as escolhas a ações economicamente consistentes, ao invés de vê-las se referindo a a constructos mentais. Historicamente, havia uma relação de alinhamento confortável, embora não uma co-construção teórica direta, entre preferência revelada em economia e o behaviorismo metodológico e ontológico que dominou a psicologia científica durante as décadas intermediárias do século XX. Contudo, é cada vez mais provável que esse uso cause confusão graças ao mais recente surgimento da **teoria comportamental dos jogos** (CAMERER, 2003). Esse programa de pesquisa visa incorporar diretamente aos modelos jogo-teóricos generalizações, derivadas principalmente de experimentos com pessoas, sobre maneiras em que elas se diferenciam de agentes puramente econômicos nas inferências que extraem de informações (*"framing"*). As aplicações também incorporam comumente pressuposições especiais sobre funções de utilidade, também derivadas de experimentos. Por exemplo, os jogadores podem ser vistos como dispostos a fazer escolhas entre as magnitudes de suas próprias recompensas e as desigualdades na distribuição de recompensas entre os jogadores. Voltaremos a algumas discussões sobre teoria comportamental dos jogos na **Seção 8.1**, **Seção 8.2** e na **Seção 8.3**. No momento, note que esse uso da teoria dos jogos depende crucialmente de pressuposições sobre representações psicológicas de

valor que se pensam ser comuns entre as pessoas. Portanto, seria enganoso se referir à teoria comportamental dos jogos como "behaviorista". Mas então apenas traria confusão continuarmos a nos referir à teoria econômica convencional dos jogos que se baseia em preferência revelada como uma teoria "behaviorista" dos jogos. Portanto, vamos nos referir a ela como teoria dos jogos "não-psicológica". Queremos referir com isso ao tipo de teoria dos jogos usado pela maior parte dos economistas que não são economistas comportamentais **revisionistas**. (Usamos o qualificador "revisionista" para refletir a complicação adicional que cada vez mais economistas que aplicam conceitos de preferência revelada realizam experimentos, e alguns deles chamam a si mesmos de "economistas behavioristas"! Para um novo conjunto de convenções proposto para reduzir esse caos de rotulação, *vide* ROSS, 2014, p. 200-201.) Esses economistas do *establishment* tratam a teoria dos jogos como a matemática abstrata da interação estratégica, em vez de uma tentativa de caracterizar diretamente disposições psicológicas especiais que poderiam ser comuns em humanos.

Teóricos dos jogos não-psicológicos tendem a desaprovar grande parte do programa de refinamento. Isso ocorre pela razão óbvia de que ele se baseia em **intuições** sobre quais tipos de inferências as pessoas **deveriam** achar sensatas. Como a maior parte dos cientistas, os teóricos não-psicológicos dos jogos suspeitam da força e da base de pressuposições filosóficas como guias para a modelagem empírica e matemática.

A teoria comportamental dos jogos pode ser entendida como um refinamento da teoria dos jogos, apesar de não o ser necessariamente de seus conceitos de solução, em um sentido diferente. Ela restringe os axiomas subjacentes da teoria para aplicação a uma classe especial de agentes, indivíduos, humanos psicologicamente normais. Ela motiva essa restrição por referência a inferências, juntamente com as preferências, que as pessoas **de fato** acham **natural***s*, independentemente de parecerem **racionais**, o que frequentemente não parecem. As teorias dos jogos não psicológica e a teoria comportamental dos jogos têm em comum que nenhuma delas se destina a ser normativa - embora ambas sejam frequentemente usadas para **descrever** normas que prevalecem em grupos de jogadores, bem como **explicar** por que as normas podem persistir em grupos de jogadores mesmo quando não parecem ser

totalmente racionais para as intuições filosóficas. Ambas veem o trabalho da teoria dos jogos **aplicada** como sendo o de prever resultados de jogos empíricos **dada** alguma distribuição de disposições estratégicas, e alguma distribuição de expectativas sobre as disposições estratégicas de outros, que são moldadas pela dinâmica nos ambientes dos jogadores, incluindo-se pressões e estruturas institucionais e seleção evolutiva. Portanto, vamos agrupar teóricos dos jogos não psicológicos e teóricos comportamentais dos jogos como teóricos **descritivos** dos jogos apenas para propósitos de contraste com os teóricos **normativos** dos jogos.

Os teóricos descritivos dos jogos estão frequentemente inclinados a duvidar que o objetivo de buscar uma teoria **geral** da racionalidade faça sentido enquanto um projeto. Instituições e processos evolutivos constroem muitos ambientes, e o que conta como um procedimento racional em um ambiente pode não ser favorecido em outro. Por outro lado, uma entidade que não satisfaz, ao menos estocasticamente (ou seja, talvez de maneira perturbada, mas estatisticamente mais frequentemente do que não perturbada), as restrições mínimas da racionalidade econômica não pode ser precisamente caracterizada como visando maximizar uma função de utilidade, exceto por acidente. Por princípio, a teoria dos jogos não possui aplicação para essas entidades.

Isso não implica que os teóricos dos jogos não psicológicos renunciem a todos os meios de restringir conjuntos de EN a subconjuntos baseados em suas probabilidades relativas de surgirem. Em particular, os teóricos dos jogos não psicológicos tendem a ser simpáticos a abordagens que trocam a ênfase em racionalidade por considerações da dinâmica informacional de jogos. Talvez não devêssemos nos surpreender que a análise de EN sozinha frequentemente falhe em nos dizer muito sobre o interesse empírico e aplicado dos jogos de formato estratégico (como a Figura 6 acima), em que a estrutura informacional é suprimida. As questões de seleção de equilíbrio são frequentemente abordadas de maneira mais proveitosa no contexto de jogos de formato extensivo.

2.6. Perfeição em Subjogos

Para aprofundar a nossa compreensão dos jogos de formato extensivo, precisamos de um exemplo com uma estrutura mais interessante do que a oferecida

pelo DP. Considere o jogo descrito por esta árvore:

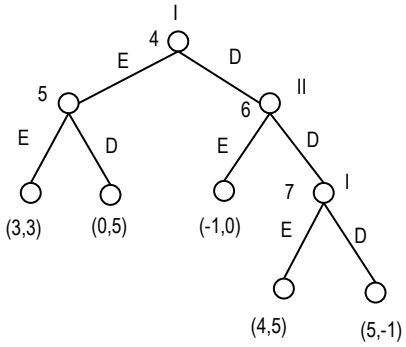


Figura 9

Esse jogo não se destina a ser encaixado em alguma situação preconcebida; ele é só um objeto matemático em busca de uma aplicação. As letras E e D denotam "esquerda" e "direita" respectivamente. Considere agora o formato estratégico do seguinte jogo:

		II			
		EE	ED	DE	DD
I	EE	3,3	3,3	0,5	0,5
	ED	3,3	3,3	0,5	0,5
	DE	-1,0	4,5	-1,0	4,5
	DD	-1,0	5,-1	-1,0	5,-1

Figura 10

Se você ficou confuso, lembre-se que uma estratégia deve dizer a um jogador o que fazer em **todo** conjunto informacional em que esse jogador tem uma ação. Uma vez que aqui cada jogador escolhe entre duas ações em cada um dos dois conjuntos informacionais, cada jogador tem quatro estratégias no total. A primeira letra em cada designação de estratégia diz a cada jogador o que fazer se ele ou ela alcançar seu primeiro conjunto informacional, e, a segunda, o que fazer se seu

segundo conjunto informacional é alcançado. Ou seja, ED para o Jogador II diz a ele para jogar E se o conjunto informacional 5 é alcançado e D se o conjunto informacional 6 é alcançado.

Se examinar a matriz na Figura 10, você descobrirá que (EE, DE) está entre os EN. Isso é um pouco intrigante, uma vez que se o Jogador I alcançar seu segundo conjunto informacional (7) no jogo de formato extensivo, ele dificilmente desejaria jogar E nesse nó; ele recebe uma recompensa maior jogando D. A mera análise de EN não discrimina isso porque o EN é insensível ao que acontece **fora do caminho do jogo**. O Jogador I garante que o nó 7 não será alcançado ao escolher E no nó 4; isso é o que significa dizer que ele está "fora do caminho do jogo". Contudo, ao se analisar jogos de formato extensivo, **devemos** nos importar com o que acontece fora do caminho do jogo, porque a consideração disso é crucial para o que acontece **no** caminho do jogo. Por exemplo, é o fato de que o Jogador I **jogaria** D se o nó 7 fosse alcançado que **causaria** o Jogador II jogar E se o nó 6 fosse alcançado, e essa é razão pela qual o Jogador I não escolherá D no nó 4. Estamos jogando fora informações relevantes para a solução de jogos se ignoramos os resultados fora do caminho do jogo, como faz a mera análise via EN. Note que essa razão para duvidar que o EN é um conceito de equilíbrio totalmente satisfatório em si mesmo não tem nada a ver com intuições sobre a racionalidade, como no caso dos conceitos de refinamento discutidos na **Seção 2.5**.

Aplique agora os algoritmo de Zermelo ao formato extensivo do exemplo atual. Comece, novamente, com o último subjogo, aquele descendente do nó 7. Esse é o movimento do Jogador I, e ele escolheria D porque ele prefere seu resultado de 5 ao resultado de 4 que recebe ao jogar E. Portanto, atribuímos o resultado (5, -1) ao nó 7. Assim, no nó 6, II enfrenta uma escolha entre (-1, 0) e (5, -1). Ele escolhe E. No nó 5, II escolhe D. No nó 4, I está, assim, escolhendo entre (0, 5) e (-1, 0), e então joga E. Note que, como no DP, um resultado aparece em um nó final - (4, 5) do nó 7 - que é Pareto superior ao EN. Contudo, novamente, a dinâmica do jogo o impede de ser alcançado.

O fato de que o algoritmo de Zermelo escolhe o vetor de estratégia (ED, DE) como a única solução para o jogo mostra que ele está produzindo algo diferente do que apenas um EN. Na verdade, ele está gerando o **equilíbrio perfeito de subjogo** (EPS) do jogo. Ele dá um resultado que produz um EN não

apenas em **todo** o jogo mas em todos os subjogos também. Esse é um conceito de solução persuasivo porque, novamente, diferentemente dos refinamentos da **Seção 2.5**, ele não demanda uma racionalidade "extra" dos agentes, no sentido da expectativa de que tenham e usem intuições filosóficas sobre "o que faz sentido". Mas ele assume de fato que os jogadores não só saibam tudo o que for estrategicamente relevante às suas situações, mas que também **usem** toda essa informação. Em argumentos sobre os fundamentos da economia, isso é frequentemente referido como um aspecto da racionalidade, como na frase "expectativas racionais". No entanto, como notado anteriormente, é melhor ter cuidado para não confundir a ideia normativa geral de racionalidade com o poder de cálculo e a posse de estimativas, de tempo e energia, para aproveitá-la ao máximo.

Um agente que joga uma estratégia perfeita em um subjogo simplesmente escolhe, em cada nó que alcançar, o caminho que lhe traz a maior recompensa **no subjogo que emana desse nó**. O EPS prevê o resultado de um jogo apenas caso os jogadores prevejam que todos eles farão isso ao solucionar o jogo.

Um valor principal ao analisar jogos de formato extensivo para EPS é que isso pode nos ajudar a localizar barreiras estruturais à otimização social . Em nosso exemplo atual, o Jogador I se sairia melhor e o Jogador II não se sairia pior no nó esquerdo que emana do nó 7 do que no resultado de EPS. Mas a racionalidade econômica do Jogador I, e a ciência do Jogador II disso, bloqueia o resultado socialmente eficiente. Se os nossos jogadores desejam reealizar o resultado mais socialmente eficiente (4,5), eles devem fazê-lo redesenhando suas instituições de modo a mudar a estrutura do jogo. O empreendimento de mudar estruturas institucionais e informacionais de modo a tornar os resultados eficientes mais prováveis nos jogos em que os agentes (isto é, pessoas, corporações, governos, etc.) realmente jogam é conhecido como **design do mecanismo**, e é uma das principais áreas de aplicação da teoria dos jogos. A principais técnicas são revistas em (HURWICZ; REITER, 2006), tendo sido Hurwicz agraciado com o Prêmio Nobel por seu trabalho pioneiro na área.

2.7. Sobre a Interpretação de Recompensas: Moralidade e Eficiência em Jogos

Muitos leitores, especialmente filósofos, podem questionar que o design do mecanismo não seria necessário no caso do exemplo na seção anterior a não ser que os jogadores sejam sociopatas morbidamente egoístas. Certamente, os jogadores podem ser capazes de claramente **ver** que o resultado (4, 5) é socialmente e moralmente superior; e visto que todo o problema também toma como certo que eles também podem ver o caminho de ações que leva a esse resultado eficiente, quem é o teórico dos jogos para anunciar que ele é inatingível a não ser que seu jogo seja alterado? Essa objeção, que aplica a ideia distinta de racionalidade incitada por Immanuel Kant, indica o principal caminho pelo qual muitos filósofos querem dizer muito mais com "racionalidade" do que dizem os teóricos descritivos dos jogos. Esse tema é explorado com grande vivacidade e força polêmica por Binmore (1994, 1998).

Essa pesada controvérsia filosófica sobre a racionalidade é por vezes tornada confusa pela má interpretação do significado de "utilidade" na teoria dos jogos não-psicológicos. Para erradicar esse erro, considere o Dilema do Prisioneiro novamente. Vimos que no único EN do DP, ambos os jogadores ganham menos utilidade do que eles poderiam ganhar por meio da cooperação mútua. Isso pode ser visto por você como algo perverso, mesmo caso você não seja um kantiano (como de fato tem sido visto por muitos comentadores). Certamente, você pode pensar que ele simplesmente resulta de uma combinação de egoísmo e paranoia da parte dos jogadores. A princípio, eles não têm consideração pelo bem social e, desse modo, atiram no próprio pé por serem desconfiados demais para respeitar acordos.

Essa maneira de pensar é muito comum em discussões populares e confusas. Primeiramente, vamos introduzir alguma terminologia para falar sobre resultados para dissipar essa má influência. Comumente, os economistas do bem-estar mensuram o bem social em termos da **eficiência de Pareto**. Uma distribuição de utilidade β é dita **Pareto superior** a outra distribuição δ apenas no caso de haver a partir de δ uma redistribuição possível de utilidade até β tal que ao menos um jogador está melhor em β do que em δ e nenhum jogador está pior.

A falha em se mover de uma distribuição Pareto inferior para uma distribuição Pareto superior é **ineficiência**, porque a existência de β como uma possibilidade, pelo menos em princípio, mostra que em δ alguma utilidade é desperdiçada. Agora, o resultado (3, 3) que representa cooperação mútua em nosso modelo do DP é claramente Pareto superior à defecção mútua; em (3, 3) **ambos** os jogadores estão melhores do que em (2, 2). Logo, é verdadeiro que os DPs levam a resultados ineficientes. Isso também era verdadeiro em nosso exemplo da **Seção 2.6**.

No entanto, a ineficiência não deve ser associada à imoralidade. Uma função de utilidade para um jogador deve representar **tudo com o que o jogador se importa**, o que pode ser qualquer coisa. Como descrevemos a situação dos nossos prisioneiros, eles de fato só se importam com suas próprias sentenças de prisão, mas não há nada de essencial nisso. O que faz um jogo uma instância do DP é estritamente e apenas sua estrutura de recompensa. Assim, poderíamos ter aqui dois tipos de Madre Teresa, e ambas se importam pouco consigo mesmas e só desejam alimentar crianças famintas. Mas suponha que a Madre Teresa original deseja alimentar as crianças de Calcutá enquanto que a Madre Juanita deseja alimentar as crianças de Bogotá. E suponha que a agência de ajuda internacional maximizará sua doação se as duas santas nomearem a mesma cidade, darão a segunda maior quantia se elas nomearem as cidades uma da outra, e a menor quantia se cada uma delas nomear a própria cidade. Nossas santas estarão em um Dilema do Prisioneiro, apesar de dificilmente serem egoístas ou despreocupadas com o bem social.

Para retornar aos nossos prisioneiros, suponha que eles **realmente** valorizem o bem-estar um do outro assim como o seu próprio, contrariamente às nossas pressuposições. Isso então deve ser refletido em suas funções de utilidade, e, consequentemente, em suas recompensas. Se suas estruturas de recompensas forem alteradas para que, por exemplo, eles se sintam tão mal por contribuir para a ineficiência que eles prefeririam passar anos extra na prisão do que suportar essa vergonha, então eles não estarão mais em um DP. Mas tudo o que isso mostra é que nem toda situação possível é um DP; ela **não** mostra que o egoísmo está entre as pressuposições da teoria dos jogos. É a **lógica** da situação dos prisioneiros, não a sua psicologia, que os prende no resultado ineficiente, e,

se essa **é** realmente a situação deles, então eles estão presos nela (excetuando-se complicações posteriores a serem discutidas abaixo). Os agentes que desejam evitar resultados ineficientes são aconselhados a evitar que certos jogos surjam; o defensor da possibilidade da racionalidade kantiana está realmente propondo que eles tentem se safar de tais jogos se transformando em diferentes tipos de agentes.

Desse modo, um jogo é em geral parcialmente **definido** pelas recompensas atribuídas aos jogadores. Em qualquer aplicação, tais atribuições devem ser baseadas em evidências empíricas sólidas. Se uma solução proposta envolve alterar tacitamente essas recompensas, então essa "solução" é, de fato, uma maneira disfarçada de alterar o assunto e de evitar as implicações de melhores práticas de modelagem.

2.8. Mãos Trêmulas e Equilíbrio de Resposta Discreta

Nosso último ponto acima abre o caminho para um problema filosófico, um de vários que ainda preocupam aqueles autores apreensivos com os fundamentos lógicos da teoria dos jogos. Ele pode ser levantado com relação a um vasto número de exemplos, mas utilizaremos um exemplo elegante de C. Bicchieri (1993).Considere o seguinte jogo:

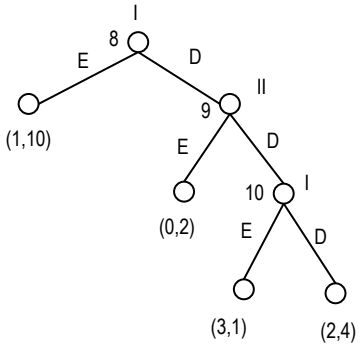


Figura 11

O resultado EN está no único nó mais à esquerda descendente do nó 8. Para ver isso, utilizamos a indução reversa novamente. No nó 10, I jogaria E por uma recompensa de 3, dando a II uma recompensa de 1. II pode fazer melhor do que isso jogando E no nó 9, dando a I uma recompensa de 0. I pode fazer melhor do que isso jogando E no nó 8; então isso é o que I faz, e o jogo termina sem que II consiga se mover. Um problema é então levantado por Bicchieri (juntamente com outros autores, incluindo Binmore (1987) e Pettit e Sugden (1989)) por meio do seguinte raciocínio. O Jogador I joga E no nó 8 porque sabe que o Jogador II é economicamente racional, e então jogaria E, no nó 9, porque o Jogador II sabe que o Jogador I é economicamente racional e então jogaria E no nó 10. Mas agora temos o seguinte paradoxo: o Jogador I deve supor que o Jogador II, no nó 9, preveria a jogada economicamente racional do Jogador I no nó 10, apesar de ter chegado em um nó (9) que só poderia ser alcançado se o Jogador I não fosse economicamente racional! Se o Jogador I não for economicamente racional, então o Jogador II não está justificado em prever que o Jogador I não jogará D no nó 10, caso em que não está claro que o Jogador II não deveria jogar D em 9; e se o Jogador II jogar D em 9, então o Jogador I tem garantia de uma recompensa melhor do que ganha caso jogue E no nó 8. Ambos os jogadores usam a indução reversa para solucionar o jogo; a indução reversa requer que o Jogador I saiba que o Jogador II saiba que o Jogador I é economicamente racional; mas o Jogador II pode solucionar o jogo apenas ao usar um argumento por indução reversa que toma como premissa a falha do Jogador I em se comportar de acordo com a racionalidade econômica. Esse é o **paradoxo da indução reversa**.

Na literatura sobre o tema, uma maneira *standard* de contornar o paradoxo é invocar a assim chamada "mão trêmula" (SELTEN, 1975). A ideia aqui é que uma decisão e seu ato consequente podem se "colapsar" com alguma probabilidade diferente de zero, por menor que seja. Isto é, um jogador pode pretender tomar uma ação, mas então escorregar na execução e enviar o jogo para outro caminho em vez disso. Se houver sequer uma possibilidade remota de que um jogador possa cometer um erro - de que sua "mão possa tremer" -, então nenhuma contradição é introduzida por um jogador usando um argumento de indução reversa que requer a pressuposição hipotética de que o outro jogador tenha tomado um caminho que um jogador economicamente racional não poderia

tomar. Em nosso exemplo, o Jogador II poderia raciocinar sobre o que fazer no nó 9 condicionado à pressuposição de que o Jogador I escolheu E no nó 8, mas, depois disso, escorregou.

Gintis (2009a) aponta que o aparente paradoxo não surge meramente da nossa suposição de que ambos os jogadores são economicamente racionais. Ele se baseia crucialmente na premissa adicional de que cada jogador deve saber, e raciocina com base nesse saber, que o outro jogador é economicamente racional. Essa é a premissa com a qual cada uma das conjecturas dos jogadores sobre o que aconteceria fora do caminho de equilíbrio do jogo é inconsistente. Um jogador tem razões para considerar possibilidades fora de equilíbrio caso ele acredite que seu oponente é economicamente racional mas que sua mão pode tremer, **ou** ele atribua alguma probabilidade diferente de zero à possibilidade de que seu oponente não seja economicamente racional, ou ele atribua alguma dúvida à sua conjectura sobre a função de utilidade de seu oponente. Como Gintis também destaca, esse problema com a solução de jogos de formato extensivo para EPS pelo algoritmo de Zermelo pode ser generalizado: um jogador não tem razões para jogar nem mesmo uma estratégia de equilíbrio de **Nash** a menos que ele espere que outros jogadores joguem estratégias de equilíbrio de Nash. Retornaremos a esse problema na **Seção 7**.

O paradoxo da indução reversa, como os problemas levantados pelo refinamento de equilíbrio, é principalmente um problema para quem vê a teoria dos jogos como uma contribuição para uma teoria normativa da racionalidade (específicamente, como uma contribuição para aquela teoria maior, a teoria da racionalidade **estratégica**). O teórico dos jogos não-psicológicos pode oferecer um tipo diferente de explicação da jogada aparentemente “irracional” e da prudência que ela encoraja. Isso envolve um apelo ao fato empírico de que agentes reais, incluindo pessoas, devem **aprender** as estratégias de equilíbrio dos jogos que eles jogam, ao menos sempre que os jogos forem complicados como um todo. Uma pesquisa mostra que mesmo um jogo simples como o Dilema do Prisioneiro exige aprendizagem por parte das pessoas (LEDYARD, 1995; SALLY, 1995; CAMERER, 2003, p. 265). O que quer dizer que as pessoas devam aprender estratégias de equilíbrio é que devemos ser um pouco mais sofisticados do que o indicado anteriormente ao construirmos funções de utilidade

a partir do comportamento na aplicação da Teoria da Preferência Revelada. Em vez de construir funções de utilidade com base em episódios singulares, devemos fazê-lo com base em sequências observadas de comportamento **assim que estiverem estabilizados**, o que significa maturidade de aprendizado por parte dos sujeitos e do jogo em questão. De novo, o Dilema do Prisioneiro é um bom exemplo. As pessoas encontram poucas versões de lance único do Dilema do Prisioneiro na vida cotidiana, mas elas encontram muitos DPs **repetidos** com não estranhos. Como resultado, quando definido no que se destina a ser uma versão de lance único do DP no laboratório experimental, as pessoas tendem a jogar inicialmente como se o jogo fosse uma única rodada de um DP repetido. O DP repetido tem muitos equilíbrios de Nash que envolvem cooperação em vez de defecção. Desse modo, os sujeitos experimentais tendem a cooperar de início nessas circunstâncias, mas aprendem a abandonar o acordo após algum número de rodadas. O experimentador não pode inferir que ele induziu com sucesso uma versão de lance único do DP com seu sistema de experimento até que veja esse comportamento se estabilizar.

Se os jogadores percebem que os outros jogadores podem precisar aprender estruturas e equilíbrios de jogo a partir da experiência, isso lhes dá razões para levar em conta o que acontece fora dos caminhos de equilíbrio de jogos de formato extensivo. Obviamente, se um jogador teme que os outros jogadores não aprenderam os equilíbrios, isso pode bem remover seu incentivo para jogar uma estratégia de equilíbrio. Há aqui um conjunto de problemas profundos sobre a aprendizagem social (FUDENBERG; LAVINE, 1998). Como jogadores ignorantes aprendem a jogar equilíbrios se jogadores sofisticados não os mostram, pois esses não são incentivados a jogar estratégias de equilíbrio até que o ignorante tenha aprendido? A resposta crucial no caso das aplicações da teoria dos jogos às interações entre pessoas é que pessoas jovens são **socializadas** ao crescerem em redes de **instituições**, incluindo **normas culturais**. A maioria dos jogos complexos que as pessoas jogam já estão em progresso entre as pessoas que foram socializadas antes delas - isto é, aprenderam estruturas e equilíbrios de jogo (ROSS, 2008a). Os novatos devem então apenas imitar aqueles jogadores cujas jogadas parecem ser esperadas e compreendidas por outros jogadores. As instituições e as normas são ricas em

lembretes, incluindo-se sermões e regras de ouro facilmente lembradas, para ajudar as pessoas a se lembrarem do que estão fazendo (CLARK, 1997).

Como notado na **Seção 2.7** acima, quando o comportamento observado **não** estabiliza em torno do equilíbrio em um jogo, e não há evidências de que a aprendizagem ainda está em progresso, o analista deve inferir que ele modelou incorretamente a situação que estuda. É bastante provável que ele tenha especificado errado as funções de utilidade dos jogadores, as estratégias ou as informações a eles disponíveis. Dada a complexidade de muitas das situações que os cientistas sociais estudam, não deve surpreeender que a especificação errada de modelos ocorra com frequência. Os teóricos dos jogos aplicados devem aprender muito, assim como seus sujeitos.

O paradoxo da indução reversa é um membro de uma família de paradoxos que surgem caso se insira a posse e o uso de informações literalmente completas em um conceito de racionalidade. (Considere, por analogia, o paradoxo do mercado de ações que surge se supomos que o investimento economicamente racional incorpora literalmente expectativas racionais: suponha que nenhum investidor individual possa vencer o mercado no longo prazo porque o mercado sempre sabe tudo o que o investidor sabe; então ninguém possui incentivo para adquirir conhecimento sobre valores de ativos; então ninguém jamais adquirirá tais informações e, portanto, da suposição de que o mercado sabe tudo , segue-se que o mercado não pode saber nada!) Como veremos em detalhes em várias discussões abaixo, a maioria das aplicações de teoria dos jogos incorpora explicitamente incerteza e perspectivas de aprendizagem pelos jogadores. Os jogos de formato extensivo com EPS que vimos acima são realmente ferramentas conceituais para nos ajudar a preparar conceitos para aplicação em situações nas quais informações completas e perfeitas são incomuns. Não podemos evitar o paradoxo se pensamos, como alguns filósofos e teóricos normativos dos jogos o fazem, que uma das ferramentas conceituais que queremos usar em teoria dos jogos para aprimorá-la é uma ideia completamente geral de racionalidade ela mesma. Mas isso não é uma preocupação para economistas e outros cientistas que aplicam a teoria dos jogos na modelagem empírica. Em casos reais, a menos que os jogadores tenham experienciado jogar em equilíbrio uns com os outros no passado, mesmo se todos forem

economicamente racionais e todos acreditarem nisso uns sobre os outros, devemos prever que eles atribuirão alguma probabilidade positiva à conjectura de que a compreensão das estruturas do jogo entre alguns jogadores é imperfeita. Isso explica por que as pessoas, mesmo se elas forem agentes economicamente racionais, podem frequentemente, ou mesmo normalmente, jogar como se acreditassem em mãos trêmulas.

A aprendizagem sobre equilíbrio pode tomar várias formas para diferentes agentes e para jogos de diferentes níveis de complexidade e risco. Assim, incorporá-la em modelos jogo-teóricos de interações introduz um novo e extenso conjunto de tecnicalidades. Para a teoria geral mais desenvolvida, o leitor pode conferir (FUDENBERG; LEVINE, 1998); os mesmos autores fornecem uma visão geral não técnica das questões em (FUDENBERG; LEVINE, 2016). Uma primeira distinção importante é entre aprender parâmetros específicos entre rodadas de um jogo repetido (*vide* **Seção 4**) com jogadores comuns e aprender sobre expectativas estratégicas gerais em diferentes jogos. Esta última aprendizagem pode incluir aprender sobre jogadores se o aprendiz está atualizando expectativas com base em seus modelos de **tipos** de jogadores que ele recorrentemente encontra. Desse modo, podemos distinguir entre aprendizagem **passiva**, em que o jogador **meramente** atualiza suas probabilidades subjetivas prévias com base na sua observação de movimentos e resultados, e na escolha estratégica que ele infere desses movimentos e resultados, e aprendizagem **ativa**, em que ele examina - costuma-se dizer: **peneira** - informações sobre as estratégias de outros jogadores, escolhendo estratégias que testam suas conjecturas sobre o que ocorrerá fora do que ele acredita ser o caminho do equilíbrio do jogo. Uma grande dificuldade para ambos os jogadores e modeladores é que movimentos de peneiramento podem ser mal interpretados se os jogadores são também incentivados a realizar movimentos para **sinalizar** informações uns para os outros (*vide* **Seção 4**). Em outras palavras: sob certas circunstâncias, tentar aprender sobre estratégias pode interferir nas habilidades dos jogadores em aprender sobre equilíbrio. Finalmente, a discussão até o momento assumiu que toda a aprendizagem possível em um jogo é sobre a estrutura do próprio jogo. Wilcox (2008) mostra que, se os jogadores aprendem novas informações sobre processos causais ocorrendo fora

de um jogo enquanto estão simultaneamente tentando atualizar expectativas sobre as estratégias de outros jogadores, o modelador pode se ver ultrapassando os limites do conhecimento técnico.

Eu disse acima que as pessoas podem **normalmente** jogar como se acreditassem em mãos trêmulas. Uma razão bastante geral para isso é que quando as pessoas interagem, o mundo não lhes fornece cartões de sinalização avisando-as acerca das estruturas dos jogos que elas estão jogando. Elas devem fazer e testar conjecturas a esse respeito a partir de seus contextos sociais. Às vezes, os contextos são fixados por regras institucionais. Por exemplo, quando uma pessoa entra em uma loja de varejo e vê uma etiqueta de preço em algo que gostaria de ter, ela sabe, sem precisar conjecturar ou aprender qualquer coisa, que está envolvida em um simples jogo de "pegar ou largar". Em outros mercados, ela pode saber que se espera que ela barganhe, e sabe as regras para isso também.

Dada a relação complexa e não resolvida entre a teoria da aprendizagem e a teoria dos jogos, o raciocínio acima pode parecer implicar que a teoria dos jogos nunca pode ser aplicada às situações envolvendo jogadores humanos que são novas para eles. Contudo, felizmente, nós não enfrentamos tal impasse. Em um par de artigos influentes em meados dos anos 1990, McKelvey e Palfrey (1995, 1998) desenvolveram o conceito de solução de **equilíbrio de resposta discreta** (ERD). O ERD não é um refinamento do EN, no sentido de ser um esforço filosoficamente motivado de fortalecer o EN por referência aos padrões normativos de racionalidade. Em vez disso, ele é um método para calcular as propriedades de equilíbrio de escolhas feitas por jogadores cujas conjecturas sobre erros possíveis nas escolhas de outros jogadores são incertas. O ERD é, assim, um equipamento *standard* no kit de ferramentas de economistas experimentais que buscam estimar a distribuição de funções de utilidade em populações de pessoas reais colocadas em situações modeladas como jogos. O ERD não teria sido praticamente utilizável dessa maneira antes do desenvolvimento de pacotes de econometria tal como o Stata (TM), que permitiu o cálculo de ERD a partir de registros de observação adequadamente poderosos de jogos interessantemente complexos. O ERD é raramente utilizado por economistas comportamentais e quase nunca é utilizado por psicólogos na

análise de dados de laboratório. Em consequência disso, muitos estudos por pesquisadores desses tipos fazem observações dramáticas e retóricas ao "descobrir" que pessoas reais frequentemente falham em convergir no EN em jogos experimentais. Contudo, apesar de ser um conceito de solução minimalista em certo sentido, porque abstrai de muita estrutura informacional, o EN é simultaneamente uma expectativa empírica exigente caso seja imposta categoricamente (isto é, caso se espere que os jogadores joguem como se todos eles estivessem certos de que todos os outros estão jogando estratégias EN). Prever jogadas consistentes com ERD é consistente com - aliás, é motivado pela - a visão de que o EN capta o conceito geral central de um equilíbrio estratégico. Uma maneira de enquadrar a relação filosófica entre o EN e o ERD é como se segue. O EN define um princípio **lógico** que é bem adaptado para disciplinar o pensamento e para conceber novas estratégias para a modelagem genérica de novas classes de fenômenos sociais Para propósitos de estimar dados empíricos reais, precisa-se de ser capaz de definir equilíbrio **estatisticamente**. O ERD representa uma maneira de fazer isso, consistentemente com a lógica do EN. A ideia é suficientemente rica para que suas profundezas permaneçam um domínio aberto de investigação por teóricos dos jogos. O estado atual de entendimento do ERD é compreensivelmente revisto em (GOEREE; HOLT; PALFREY, 2016).

3. Incerteza, Risco e Equilíbrio Sequencial

Todos os jogos que modelamos até agora envolveram jogadores que escolhem entre **estratégias puras**, nas quais cada um busca um único curso de ação otimizado em cada nó que constitua a melhor resposta às ações de outros. Contudo, a utilidade de um jogador é frequentemente otimizada através do uso de uma estratégia **mista**, em que ele lança uma moeda ponderada entre as várias ações possíveis. (Veremos que há uma interpretação alternativa de mistura que não envolve randomização em um conjunto informacional particular; mas começaremos aqui com a interpretação do lançamento de moedas e a partir disso construiremos essa interpretação na **Seção 3.1**.) Misturar é necessário sempre que nenhuma estratégia pura maximiza a utilidade de um jogador contra todas as estratégias adversárias. O nosso jogo de atravessar o rio da **Seção 1** exemplifica

isso. Como vimos, o problema naquele jogo consiste no fato de que se o raciocínio do fugitivo seleciona uma ponte particular como a melhor, deve-se assumir que perseguidor seja capaz de replicar esse raciocínio. O fugitivo pode escapar somente se seu perseguidor não pode prever de maneira confiável qual ponte ele [o fugitivo] usará. A simetria do poder de raciocínio lógico da parte dos dois jogadores garante que o fugitivo possa surpreender o perseguidor somente se for possível, para ele, surpreender **a si mesmo**.

Vamos ignorar as pedras e as cobras por um momento e imaginar que as pontes são igualmente seguras. Suponha que o fugitivo não tenha nenhum conhecimento especial acerca de seu perseguidor que possa levá-lo a arriscar uma distribuição de probabilidade especialmente conjecturada sobre as estratégias disponíveis do perseguidor. Nesse caso, o melhor caminho para o fugitivo é lançar um dado de três lados, em que cada lado representa uma ponte diferente (ou, mais convencionalmente, um dado de seis lados em que cada ponte é representada por dois lados). Ele deve se pré-comprometer a utilizar qualquer ponte que for selecionada por esse **dispositivo randômico**. Isso corrige as chances de sobrevivência independentemente do que o perseguidor faça; mas como o perseguidor não tem razões para preferir qualquer estratégia pura ou mista, e como, em qualquer caso, estamos presumindo que sua posição epistêmica seja simétrica à do fugitivo, podemos supor que ele lançará um dado de três lados por conta própria. O fugitivo tem agora uma probabilidade de 2/3 de escapar e o perseguidor uma probabilidade de 1/3 de pegá-lo. Nem o fugitivo nem o perseguidor podem melhorar suas chances dada a mistura randomizada do outro, portanto as duas estratégias de randomização estão em equilíbrio de Nash. Note que se **um** jogador randomiza, então o outro se sai igualmente bem em qualquer combinação de probabilidades sobre pontes; e portanto há infinitamente muitas combinações de melhores respostas. Contudo, cada jogador deve se preocupar se alguma estratégia que não seja randômica possa ser coordenada com algum fator que o outro jogador pode detectar e explorar. Uma vez que qualquer estratégia não-randômica é explorável por outra estratégia não-randômica, em um jogo de soma zero, tal como no nosso exemplo, apenas o vetor de estratégias randomizadas é um EN.

Agora, vamos reintroduzir os fatores paramétricos, isto é, o deslizamento

de pedras na ponte #*2* e as cobras na ponte #*3*. Novamente, suponha que o fugitivo está certo de atravessará com segurança a ponte #*1*, tenha 90% de chance de atravessar com segurança a ponte #*2*, e 80% de chance de atravessar com segurança a ponte #*3*. Podemos resolver esse novo jogo se fizermos certas pressuposições acerca das funções de utilidade dos dois jogadores. Suponha que o Jogador I, o fugitivo, se importa apenas com viver ou morrer (preferindo a vida à morte) enquanto que o perseguidor simplesmente deseja ser capaz de reportar que o fugitivo está morto, preferindo isso a reportar que ele escapou. (Em outras palavras, nenhum dos jogadores se importa com **como** o fugitivo vive ou morre.) Suponha também, por enquanto, que nenhum dos jogadores recebe qualquer utilidade ou desutilidade por assumir mais ou menos riscos. Nesse caso, o fugitivo simplesmente pega sua fórmula de randomização original e a pondera de acordo com os diferentes níveis de perigo paramétrico nas três pontes. Cada ponte deve ser pensada como uma **loteria** sobre os possíveis resultados do fugitivo, em que cada loteria tem uma **recompensa esperada** diferente em termos dos itens em sua função de utilidade.

Considere as questões do ponto de vista do perseguidor. Ele estará usando sua estratégia EN quando escolher a mistura de probabilidades sob as três pontes que torna o fugitivo indiferente quanto às suas possíveis estratégias puras. A ponte com pedras é 1.1 vezes mais perigosa para ele do que a ponte segura. Portanto, ele será indiferente entre as duas quando o perseguidor tiver 1.1 vezes mais chances de estar esperando na ponte segura do que na ponte com pedras. A ponte com cobras é 1.2 vezes mais perigosa para o fugitivo do que a ponte segura. Portanto, ele será indiferente entre essas duas pontes quando a probabilidade do perseguidor esperar na ponte segura for 1.2 vezes maior do que a probabilidade de ele [o perseguidor] esperar na ponte com cobras. Suponha que usemos s1, s2 e s3 para representar as taxas de sobrevivência paramétricas do fugitivo em cada ponte. Assim, o perseguidor minimiza a taxa de sobrevivência líquida em qualquer par de pontes ajustando as probabilidades p1 e p2 de que ele esperará nelas tal que

$$s1(1 - p1) = s2(1 - p2).$$

Uma vez que $p1 + p2 = 1$, podemos reescrever isso como

$$s1 \times p2 = s2 \times p1$$

Assim,

$$p1/s1 = p2/s2.$$

Desse modo, o perseguidor encontra sua estratégia EN resolvendo as seguintes equações simultâneas:

$$1(1 - p1) = 0.9(1 - p2)$$
$$= 0.8(1 - p3)$$
$$p1 + p2 + p3 = 1.$$

Portanto,

$$p1 = 49/121$$
$$p2 = 41/121$$
$$p3 = 31/121$$

Agora, permita que f1, f2 e f3 representem as probabilidades com as quais o fugitivo escolhe cada respectiva ponte. O fugitivo então encontra sua estratégia EN resolvendo:

$$s1 \times f1 = s2 \times f2$$
$$= s3 \times f3$$

Assim,

$$1 \times f1 = 0.9 \times f2$$
$$= 0.8 \times f3$$

Simultaneamente com

$$f1 + f2 + f3 = 1.$$

Portanto,

$$f1 = 36/121$$
$$f2 = 40/121$$
$$f3 = 45/121$$

Esses dois conjuntos de probabilidades dizem a cada jogador como ponderar seu dado antes de jogá-lo. Note o resultado - talvez surpreendente - de que o fugitivo utiliza pontes mais arriscadas com uma **maior** probabilidade, embora, por hipótese, não sinta prazer nenhum por jogos de azar. Essa é a única maneira de tornar o perseguidor indiferente quanto a qual ponte em que ele aposta, o que, por sua vez, é o que maximiza a probabilidade de sobrevivência do fugitivo.

Fomos capazes de resolver esse jogo diretamente porque definimos as funções de utilidade de modo a torná-lo **soma zero** ou **estritamente competitivo**. Isto é, todo ganho em utilidade esperada por um jogador representa uma perda precisamente simétrica pelo outro. Contudo, essa situação pode frequentemente não ser o caso. Suponha que as funções de utilidade sejam mais complexas. O perseguidor prefere mais um resultado em que ele atira no fugitivo e então reivindica crédito por sua prisão do que um em que ele [o fugitivo] morre por deslizamento de pedra ou por mordida de cobra; e o perseguidor prefere esse segundo resultado à fuga do fugitivo. O fugitivo prefere uma morte rápida por tiro do que a dor de ser esmagado ou o terror de um encontro com uma cobra. Acima de tudo, é claro, ele prefere escapar. Plausivelmente, suponha que o fugitivo se importa mais **intensamente** em sobreviver do que com ser morto de uma maneira em vez de outra. Como antes, não podemos resolver esse jogo simplesmente com base em conhecer as funções de utilidade ordinal dos jogadores, uma vez que as **intensidades** de suas respectivas preferências serão agora relevantes para suas estratégias.

Antes do trabalho de von Neumann e Morgenstern (1947), situações desse tipo eram inerentemente desconcertantes para os analistas. Isso ocorre porque a utilidade não denota uma variável psicológica escondida tal como o **prazer**. Como discutimos na **Seção 2.1**, a utilidade é meramente uma medida de disposições comportamentais relativas dadas certas pressuposições de consistência sobre as relações entre preferências e escolhas. Portanto, não faz sentido imaginar comparar entre si as preferências **cardinais** - isto é, sensíveis à intensidade -, visto que não há um padrão de medida independente e interpessoalmente constante que possamos usar. Como então podemos modelar jogos em que informações cardinais são relevantes? Afinal, como vimos, modelar jogos requer que as utilidades de todos os jogadores sejam levadas em conta

simultaneamente.

Um aspecto crucial do trabalho de von Neumann e Morgenstern (1947) foi a solução para esse problema. Aqui, forneceremos um breve esboço de sua técnica engenhosa para construir funções de utilidade cardinais a partir de funções de utilidade ordinal. Enfatiza-se que o que se segue é meramente um **esboço**, de modo a tornar a utilidade cardinal não misteriosa para você enquanto estudante que está interessado em saber sobre os fundamentos filosóficos da teoria dos jogos, e sobre o alcance dos problemas a que ela pode ser aplicada. Fornecer um manual que você pudesse seguir na **construção** de suas próprias funções de utilidade requereria muitas páginas. Manuais desse tipo estão disponíveis em muitos livros didáticos.

Suponha que nós agora atribuamos a seguinte função ordinal ao fugitivo que atravessa o rio:

Escapar $\gg 4$
Morte por tiro $\gg 3$
Morte por deslizamento de pedras $\gg 2$
Morte por mordida de cobra $\gg 1$

Supomos que a preferência do fugitivo por escapar sob **qualquer** forma de morte é mais forte do que suas preferências entre as causas de morte. Isso deve ser refletido no seu comportamento de escolha da seguinte maneira. Em uma situação tal como o jogo de atravessar o rio, ele deve estar disposto a correr maiores riscos para aumentar a probabilidade relativa de fuga em vez de tiro do que ele está em aumentar a probabilidade relativa de tiro em vez de mordida de cobra. Esse bocado de lógica é o *insight* crucial por trás da solução de von Neumann e Morgenstern (1947) para o problema da cardinalização.

Suponha que perguntemos ao fugitivo para selecionar o **melhor** e o **pior** resultado dentre o conjunto de resultados disponíveis. "Melhor" e "pior" são definidos em termos de recompensas esperadas como ilustrado em nosso atual exemplo do jogo de soma zero: um jogador maximiza sua recompensa esperada se, quando escolhe entre loterias que só contêm dois prêmios possíveis, ele sempre escolhe de modo a maximizar a probabilidade do melhor resultado - chame isso de **W -** e a minimizar a probabilidade do pior resultado - chame isso de

L. Agora, imagine expandir o conjunto de prêmios possíveis de modo que ele inclua prêmios que o agente valorize como intermediários entre **W** e **L**. Para um conjunto de resultados contendo tais prêmios, encontramos uma loteria sobre eles tal que o nosso agente é indiferente entre essa loteria e uma loteria incluindo apenas **W** e **L**. No nosso exemplo, essa é uma loteria que inclui ser baleado e ser esmagado por pedras. Chame essa loteria de **T**. Definimos a função de utilidade $q = u(\boldsymbol{T})$ dos resultados para a linha numérica real (oposta à ordinal) tal que, se q é o prêmio esperado em **T**, o agente é indiferente entre ganhar **T** e ganhar uma loteria **T*** em que **W** ocorre com probabilidade $u(\boldsymbol{T})$ e **L** ocorre com probabilidade $1 - u(\boldsymbol{T})$. Assumindo-se que o comportamento do agente respeita o princípio de **redução de loterias compostas** (RLC) - isto é, ele não ganha ou perde utilidade ao considerar loterias mais complexas em vez de simples - o conjunto de mapeamentos de resultados em **T** para u**T*** dá uma função de utilidade von Neumann-Morgenstern (fuvNM) com estrutura cardinal sobre todos os resultados em **T**.

O que fizemos aqui, exatamente? Demos escolhas entre loterias ao nosso agente, em vez de escolhas entre resultados resolvidos, e observado o quanto de risco extra de morte ele está disposto a correr para mudar as chances de receber uma forma de morte relativamente à outra. Note que isso cardinaliza a estrutura de preferência do agente apenas relativamente aos pontos de referência específicos **W** e **L** do agente; o procedimento não revela nada sobre preferências extraordinais comparativas **entre** agentes, o que ajuda a tornar claro que construir uma fuvNM não introduz um elemento psicológico potencialmente objetivo. Além disso, dois agentes em um jogo, ou um agente sob diferentes tipos de circunstâncias, podem exibir atitudes variadas em relação ao risco. Talvez no jogo de atravessar o rio, o perseguidor, cuja vida não está em jogo, gostará de apostar a sua glória enquanto que o fugitivo será cauteloso. Contudo, ao analisar o jogo de atravessar o rio, não **temos** de ser capazes de comparar as utilidades cardinais do perseguidor com as do fugitivo. Afinal de contas, ambos os agentes podem encontrar suas estratégias EN caso possam estimar as probabilidades que cada um atribuirá às ações do outro. Isso significa que cada um deve conhecer ambas as fuvNM; mas nenhum precisa tentar avaliar comparativamente os resultados entre os quais estão escolhendo.

Podemos agora preencher o restante da matriz do jogo de atravessar o rio que começamos a esboçar na **Seção 2**. Se ambos os jogadores são neutros quanto ao risco e suas preferências reveladas respeitam a RLC, então temos informações suficientes para sermos capazes de atribuir utilidades esperadas, exprimidas pela multiplicação das recompensas originais pelas probabilidades relevantes, como resultados na matriz. Suponha que o perseguidor espere na ponte com cobras com probabilidade *x* e na ponte com pedras com probabilidade *y*. Uma vez que suas probabilidades de atravessar as três pontes devem somar 1, isso implica que ele deve esperar na ponte segura com a probabilidade $1-(x+y)$. Então, continuando a atribuir ao fugitivo uma recompensa de 0 caso ele morra e de 1 caso ele escape, e ao perseguidor as recompensas inversas, nossa matriz é como se segue:

		Perseguidor		
		Ponte Segura	Ponte de Pedras	Ponte de Cobras
Fugitivo	Ponte Segura	0,1	1, 0	1,0
	Ponte de Pedras	0.9, 0.1	0, 1	0.9, 0.1
	Ponte de Cobras	0.8, 0.2	0.8, 0.2	0,1

Figura 12

Agora, podemos ler os seguintes fatos sobre o jogo diretamente da matriz. Nenhum par de estratégias puras é um par de respostas melhores do que outro. Portanto, o único EN do jogo requer que ao menos um jogador use uma estratégia mista.

3.1. Crenças e Probabilidades Subjectivas

Em todos os nossos exemplos e trabalhos até aqui pressupomos que as crenças dos jogadores sobre probabilidades em loterias correspondem a probabilidades objetivas. Contudo, em situações reais de escolhas interativas, os agentes devem frequentemente confiar em suas estimativas subjetivas ou percepções de probabilidades. Em uma das maiores contribuições às ciências comportamentais e sociais do século XX, Savage (1954) mostrou como incorporar

probabilidades subjetivas, e suas relações com as preferências sobre os riscos, dentro do *framework* da teoria da utilidade esperada de von Neumann-Morgenstern. De fato, a conquista de Savage equivale à conclusão formal da TUE [Teoria da Utilidade Esperada]. Pouco mais de uma década depois, Harsanyi (1967) mostrou como resolver jogos envolvendo maximizadores da utilidade esperada de Savage. Isso é considerado frequentemente como tendo marcado a verdadeira maturidade da teoria dos jogos como uma ferramenta para aplicação nas ciências comportamentais e sociais, e foi reconhecida como tal quando Harsanyi se juntou a Nash e Selten como ganhador do primeiro prêmio Nobel concedido a teóricos dos jogos em 1994.

Como observamos ao considerar a necessidade para as pessoas que jogam jogos de aprender o equilíbrio das mãos trêmulas e ERQ, quando modelamos as interações estratégicas de pessoas, devemos levar em conta o fato de que elas são comumente incertas sobre os modelos umas das outras. Essa incerteza é refletida em suas escolhas de estratégias. Além disso, algumas ações podem ser realizadas especificamente pelo motivo de aprender sobre a precisão das conjecturas de um jogador acerca de outros jogadores. A extensão de Harsanyi da teoria jogos incorpora esses elementos cruciais.

Considere o jogo de informações imperfeitas com três jogadores abaixo, conhecido como “o cavalo de Selten” (devido ao seu inventor, também premiado com o Nobel, Reinhard Selten, e por causa da forma de sua árvore; retirado de (KREPS, 1990, p. 426)):

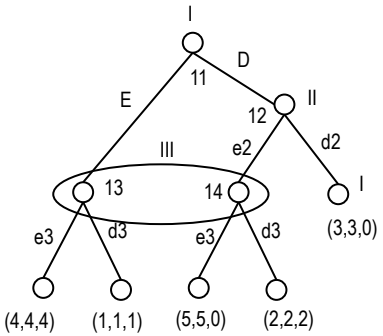


Figura 13

Esse jogo tem quatro EN: (E, e_2, e_3), (E, d_2, e_3), (D, d_2, e_3) e (D, d_2, d_3). Considere o quarto desses EN. Ele surge porque, quando o Jogador I joga D e o Jogador II joga d_2, o conteúdo informacional inteiro do Jogador III está fora do caminho de jogo e não importa para o resultado o que o Jogador III faz. Mas o Jogador I não jogaria D se o Jogador III pudesse dizer a diferença entre estar no nó 13 e estar no nó 14. A estrutura do jogo incentiva esforços do Jogador I para fornecer ao Jogador III informações que abririam seu conjunto de informações fechadas. O Jogador III deve acreditar nessa informação porque a estrutura do jogo mostra que o Jogador I tem incentivos para comunicá-la com veracidade. A solução do jogo seria então que o EPS do (agora) jogo de informação perfeita: (E, d_2, e_3).

Os teóricos que encaram a teoria dos jogos como parte de uma teoria normativa da racionalidade geral, como é o caso da maioria dos filósofos, e economistas entusiastas do programa de refinamento, têm buscado uma estratégia que identificaria os princípios gerais de tal solução. Observe o que o Jogador III no Cavalo de Selten pode se perguntar ao selecionar sua estratégia. "Dado que eu tenho um movimento, meu nó de ação foi alcançado do nó 11 ou do nó 12?" Em outras palavras, o que são as **probabilidades condicionais** de que o Jogador III está no nó 13 ou 14, dado que ele tem um movimento? Se o Jogador III se pergunta sobre probabilidades condicionais, então sobre o que os Jogadores I e II podem conjecturar quando eles selecionam **suas** estratégias são as **crenças** do Jogador III sobre essas probabilidades condicionais. Nesse caso,

o Jogador I deve conjecturar sobre as crenças do Jogador II sobre as crenças do Jogador III, e as crenças do Jogador III sobre as crenças do Jogador II, e assim por diante. As crenças relevantes aqui não são meramente estratégicas, uma vez que não só dizem respeito ao que os jogadores **farão**, dado um conjunto de recompensas e estruturas de jogo, mas, antes, sobre com qual entendimento de probabilidades condicionais eles devem esperar que os outros jogadores operem.

Quais crenças sobre probabilidade condicional são razoáveis para os jogadores esperarem uns dos outros? Se seguirmos Savage (1954), sugeriríamos como um princípio normativo que eles devem raciocinar e ter expectativas de acordo com a **regra de Bayes**. Isso lhes diz como computar a probabilidade de um evento F dada informação E (escrito "$pr(F/E)$"):

$$pr(F/E) = [pr(E/F) \times pr(F)]/pr(E)$$

Se pressupormos que as crenças dos jogadores são sempre consistentes com essa equação, então podemos definir um **equilíbrio sequencial**. Um ES tem duas partes: (1) um perfil de estratégia $\S$ para cada jogador e (2) **um sistema de crenças** μ para cada jogador. μ atribui a cada conjunto informacional h uma distribuição de probabilidade sobre os nós em h, com a interpretação de que essas são as crenças do jogador $i(h)$ sobre a posição na qual ele está em seu conjunto informacional, dado que o conjunto informacional h foi alcançado. Assim, um equilíbrio sequencial é um perfil de estratégias $\S$ e um sistema de crenças μ consistentes com a regra de Bayes tal que, começando de todo conjunto informacional h na árvore, o jogador $i(h)$ joga de forma otimizada a partir de então, dado que o que ele acredita ter acontecido anteriormente é dado por $\mu(h)$ e o que acontecerá nos movimentos subsequentes é dado por $\S$.

Vamos aplicar esse conceito de solução ao Cavalo de Selten. Considere novamente o EN (D, d_2, d_3). Suponha que o Jogador III atribui $pr(1)$ à sua crença de que estará no nó 13 caso ele faça um movimento. Então, dado um $\mu(I)$ consistente, o jogador I deve acreditar que o Jogador III jogará e_3, onde, no caso, sua única estratégia ES é E. Assim, embora (D, d_2, e_3) seja um EN, ele não é um ES.

O uso do requerimento de consistência nesse exemplo é de alguma

maneira trivial; assim, considere agora um segundo caso (também retirado de (KREPS, 1990, P. 429)):

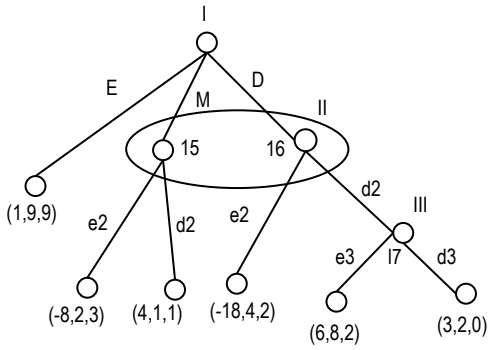


Figura 14

Suponha que o Jogador I jogue E, que o Jogador II jogue e_2 e que o Jogador III jogue e_3. Suponha também que $\mu(II)$ atribui $pr(0.3)$ ao nó 16. Nesse caso, e_2 não é uma estratégia ES para o Jogador II, uma vez que e_2 retorna uma recompensa esperada de $(0.3 \times 4) + (0.7 \times 2) = 2.6$, enquanto que d_2 traz uma recompensa esperada de 3.1. Note que se mexermos no perfil de estratégia do Jogador III enquanto mantivermos tudo o mais fixo, e_2 poderia **se tornar** uma estratégia ES para o Jogador II. Se $\S(III)$ produzisse uma jogada de e_3 com $pr(0.5)$ e d_3 com $pr(0.5)$, então, se o Jogador II jogar d_2, sua recompensa esperada será agora de 2.2 e (Ee_2e_3) seria um ES. Agora, imagine colocar $\mu(III)$ de volta como era antes, mas mudar $\mu(II)$ de tal modo que o Jogador II pense a probabilidade condicional de estar no nó 16 como maior que 0.5; nesse caso, e_2 não é, novamente, uma estratégia ES.

Felizmente, a ideia de ES é agora clara. Podemos aplicá-la ao jogo de atravessar o rio de uma maneira que evite a necessidade do perseguidor lançar qualquer moeda caso modifiquemos um pouco o jogo. Suponha agora que o perseguidor possa mudar de ponte duas vezes durante a passagem do fugitivo e que irá pegá-lo apenas no caso em que o encontrar enquanto ele [o fugitivo] sai da ponte. Então a estratégia ES do perseguidor é dividir seu tempo nas três

pontes de acordo com a proporção dada pela equação no terceiro parágrafo da **Seção 3** acima.

Como a regra de Bayes não pode ser aplicada aos eventos com probabilidade 0, deve-se notar que a sua aplicação ao ES requer que os jogadores atribuam probabilidades diferentes de zero a todas as ações disponíveis em formato extensivo. Esse requerimento é cumprido pressupondo-se que todos os perfis de estratégia sejam **estritamente mistos**, isto é, que toda ação em todo conjunto informacional seja tomada com probabilidade positiva. Você verá que isso é justamente equivalente a pressupor que todas as mãos às vezes tremem, ou, alternativamente, que nenhuma expectativa é completamente certa. Diz-se que um ES é mão trêmula perfeito se todas as estratégias jogadas em equilíbrio são as melhores respostas às estratégias que são estritamente mistas. Não deve surpreendê-lo também ser dito que nenhuma estratégia fracamente dominada pode ser mão trêmula perfeita, uma vez que a possibilidade de mãos trêmulas dá aos jogadores a razão mais convincente para evitar tais estratégias.

Como pode o teórico não-psicológico dos jogos entender o conceito de um EN que é um equilíbrio tanto de ações quanto de crenças? Décadas de estudo experimental têm mostrado que, quando os seres humanos jogam jogos, especialmente jogos que idealmente pedem o uso da regra de Bayes para fazer conjecturas sobre as crenças de outros jogadores, devemos esperar uma **heterogeneidade** significativa nas respostas estratégicas. Múltiplas espécies de canais de informação conectam comumente diferentes agentes às estruturas de incentivo em seus ambientes. Na verdade, alguns agentes podem calcular o equilíbrio, com mais ou menos erros. Outros podem se estabelecer dentro dos intervalos de erro que giram em torno, aleatoriamente, dos valores de equilíbrio por meio da aprendizagem condicionada que pode ser mais ou menos míope. Outros ainda podem selecionar padrões de resposta copiando o comportamento de outros agentes, ou seguindo regras de senso comum atreladas às estruturas culturais e institucionais e que representam a aprendizagem coletiva histórica. Note que a questão aqui é específica da teoria dos jogos, ao invés de ser meramente uma reiteração de um ponto mais geral, que se aplicaria a qualquer ciência comportamental, de que as pessoas se comportam de modo ruidoso

quando consideradas da perspectiva da teoria ideal. Em um determinado jogo, se seria racional mesmo para um agente treinado, ciente de si, e computacionalmente bem provido jogar EN, dependeria da frequência com que ele esperasse que os outros fizessem o mesmo. Se ele espera que alguns outros jogadores se desviem do jogo EN, isso pode dar-lhe razões para se desviar. Em vez de prever que os jogadores humanos exibirão estratégias EN, o experimentador ou modelador experiente antecipa que haverá uma relação entre seu jogo e os custos esperados de afastamento do EN. Consequentemente, a estimativa de máxima *likelihood* das ações observadas identificam comumente um ERQ como fornecendo um ajuste melhor do que qualquer EN.

Ao lidar com dados empíricos dessa maneira, um analista não deve ser interpretado como "testando a hipótese" de que os agentes sob análise são "racionais". Em vez disso, ele conjectura que eles são agentes, isto é, que há uma relação sistemática entre alterações em padrões estatísticos dos seus comportamentos e algumas classificações cardinais ponderadas pelo risco de possíveis estados finais. Se os agentes são pessoas ou grupos de pessoas institucionalmente estruturadas que monitoram umas às outras e são incentivadas a tentar agir coletivamente, então essas conjecturas serão frequentemente consideradas razoáveis pelos críticos, ou mesmo como pragmaticamente fora de questão, ainda que sempre anuláveis, dada a possibilidade diferente de zero de ocorrerem circunstâncias desconhecidas bizarras do tipo que os filósofos por vezes consideram (por exemplo, as pessoas aparentes são simulacros mecânicos não-inteligentes e pré-programados que seriam revelados como tais apenas caso o ambiente incentivasse respostas não escritas em seu programas). O analista pode assumir que todos os agentes respondem às mudanças de incentivo de acordo com a teoria da utilidade esperada de Savage, particularmente se os agentes são empresas que aprenderam contingências de respostas sob condições normativamente exigentes de competição de mercado com muitos jogadores. Se os sujeitos do analista são pessoas individuais, e, especialmente, se elas estão em um ambiente não *standard* relativamente à sua experiência cultural e institucional, ele [o analista] estimaria mais sabiamente um modelo misto de máxima *likelihood* que permita que uma gama de diferentes estruturas de utilidade governe diferentes subconjuntos de seus dados de escolha. Tudo

isso é para dizer que o uso da teoria dos jogos não força o cientista a aplicar empiricamente um modelo que tende a ser preciso e limitado demais em suas especificações para ajustar plausivelmente as confusas complexidades da interação estratégica real. Um bom e aplicado teórico dos jogos deve ser também um econometrista bem formado .

4. Jogos Repetidos e Coordenação

Até o momento, restringimos a nossa atenção a jogos **de lance único**, isto é, jogos em que as preocupações estratégicas dos jogadores não se estendem para além dos nós terminais de sua única interação. No entanto, jogos são frequentemente jogado com partidas **futuras** em mente, e isso pode alterar significativamente os seus resultados e as estratégias de equilíbrio. Nosso tópico nessa seção são **jogos repetidos**, isto é, jogos em que conjuntos de jogadores esperam se deparar uns com os outros em situações similares em múltiplas ocasiões. Abordamos esses jogos primeiramente por meio do contexto limitado dos repetidos dilemas do prisioneiro.

Vimos que o único EN no jogo de lance único do DP era a defecção mútua. Mas isso pode não mais se manter caso os jogadores esperem se encontrar novamente em futuros DPs. Imagine que quatro empresas, todas fabricando *widgets*, concordam em manter os preços altos restringindo conjuntamente o fornecimento do produto (isto é, elas formam um cartel). Isso funcionará apenas se cada uma mantiver a cota de produção acordada. Comumente, cada empresa pode maximizar seu lucro se afastando de sua cota enquanto que as outras observam o acordo, uma vez que ela depois vende mais unidades pelo maior preço de mercado provocado pelo cartel quase intacto. No caso do lance único, todas as empresas compartilhariam desse incetivo de abandonar o acordo, e o cartel entraria em colapso imediatamente. Contudo, as empresas esperam se deparar umas com as outras em competição por um período longo de tempo. Nesse caso, cada empresa sabe que, se ela quebrar o acordo do cartel, as outras podem puni-la com preços mais baixos por um período longo o suficiente para mais do que eliminar seu ganho de curto prazo. Obviamente, as empresas punidoras também terão perdas no curto prazo durante

seu período de preços baixos. Mas essas perdas podem valer a pena caso sirvam para reestabelecer o cartel e maximizar os preços no longo prazo.

Uma estratégia simples e famosa (porém **não** necessariamente otimizada, contrariamente ao mito bastante difundido) para preservar a cooperação em DPs repetidos é chamada de **olho por olho**. Essa estratégia diz a cada jogador para se comportar do seguinte modo:

1. Sempre coopere na primeira rodada.
2. Depois disso, realize qualquer ação que o seu oponente realizou na rodada anterior.

Um grupo de jogadores em que **todos** estão jogando olho por olho nunca verá qualquer defecção. Uma vez que o olho por olho é a resposta racional para cada jogador em uma população em que os outros jogam essa estratégia, todos jogando olho por olho é um EN. Você deve frequentemente ouvir pessoas que sabem **um pouco** (mas não o suficiente) de teoria dos jogos falarem como se isso fosse o fim da história. Mas não é.

Há duas complicações. Primeiramente, os jogadores devem estar incertos quanto ao prazo final das suas interações. Suponha que os jogadores sabem quando é a última rodada. Nessa rodada, abandonar o acordo será um maximizador de utilidade para os jogadores, visto que nenhuma punição será possível. Agora, considere a penúltima rodada. Nela, os jogadores também não enfrentam punição alguma por abandonarem o acordo, visto que esperam fazê-lo na última rodada de qualquer modo. Então, eles abandonam o acordo na penúltima rodada. Mas isso significa que eles não enfrentam nenhuma ameaça de punição na antepenúltima rodada, e o abandonam lá também. Nós podemos simplesmente repetir isso para trás ao longo da árvore do jogo até atingirmos a primeira rodada. Como a cooperação não é uma estratégia de EN nessa rodada, o olho por olho não é mais um equilíbrio de Nash no jogo repetido, e obtemos o mesmo resultado - defecção mútua - que o da versão de lance único do DP. Portanto, a cooperação só é possível em DPs repetidos em que o número esperado de repetições é indeterminado. (Obviamente, isso se aplica de fato a muitos jogos da vida real.) Note que, nesse contexto, qualquer quantidade de incerteza em expectativas, ou possibilidade de mãos trêmulas, será propício à

cooperação, ao menos por um tempo. Quando as pessoas em experimentos jogam DPs repetidos com pontos finais conhecidos, elas realmente tendem a cooperar por algum tempo, porém aprendem a abandonar o acordo mais cedo à medida que ganham experiência.

Agora, introduziremos uma segunda complicação. Suponha que a habilidade dos jogadores em distinguir defecção de cooperação seja imperfeita. Considere nosso caso do cartel do *widget*. Suponha que os jogadores observem uma queda no preço de mercado de *widgets*. Talvez isso tenha ocorrido porque um membro do cartel trapaceou. Ou talvez ela tenha sido resultado de um declínio exógeno na demanda. Se os jogadores de olho por olho confundirem o segundo caso com o primeiro, eles abandonarão o acordo, provocando assim uma reação em cadeia de defecções mútuas da qual eles nunca se recuperarão, uma vez que cada jogador irá responder com defecção à primeira defecção com a qual se depara, gerando, desse modo, mais defecções, e assim por diante.

Se os jogadores sabem que é possível haver tal má comunicação, eles possuem incentivos para recorrer a estratégias mais sofisticadas. Em particular, eles podem estar dispostos a arriscar, de vez em quando, seguir defecções com cooperação a fim de testar suas inferências. Contudo, se eles forem **muito** indulgentes, então os outros jogadores podem explorá–los através de defecções adicionais. Em geral, estratégias sofisticadas têm um problema. Por elas serem mais difíceis para os outros jogadores inferirem, seu uso aumenta a probabilidade de má comunicação. Mas a má comunicação é o que faz com que o equilíbrio cooperativo de jogos repetidos comece a falhar, em primeiro lugar. As complexidades que circundam a sinalização, a peneiragem e a inferência de informações em DPs repetidos ajudam a explicar intuitivamente o **teorema popular** [*folk theorem*], assim chamado porque ninguém tem certeza de quem primeiro o reconheceu, que, em DPs repetidos, para **qualquer** estratégia S, existe uma possível distribuição de estratégias entre outros jogadores tal que o vetor de S e essas outras estratégias são um EN. Portanto, não há nada de especial, afinal, no olho por olho.

Dramas sociais e políticos reais e complexos raramente são instanciações diretas de jogos simples como DPs. Hardin (1995) oferece uma análise de dois casos políticos tragicamente reais, a guerra civil Iugoslava de

1991-95 e o genocídio em Ruanda em 1994, como DPs que foram classificados como **jogos de coordenação**.

Um jogo de coordenação ocorre sempre que a utilidade de dois ou mais jogadores é maximizada por eles fazerem a mesma coisa que o outro, e onde tal correspondência é mais importante para eles do que o que quer que seja que ambos façam em particular. Um exemplo *standard* surge com as regras de direção: "todos dirigem à esquerda" e "todos dirigem à direita" são ambos resultados que são ENs, e nenhum é mais eficiente do que o outro. Em jogos de coordenação "pura", nem mesmo ajuda usar um critério de equilíbrio mais seletivo. Por exemplo, suponha que requeremos que nossos jogadores raciocinem de acordo com a regra de Bayes (*vide* **Seção 3**). Nessas circunstâncias, qualquer estratégia que seja uma resposta melhor a qualquer vetor de estratégias mistas disponíveis em EN é chamada de **racionalizável**. Isto é, um jogador pode encontrar um conjunto de sistemas de crenças para os outros jogadores tal que qualquer história do jogo ao longo de um caminho de equilíbrio seja consistente com esse conjunto de sistemas. Jogos de coordenação pura são caracterizados por vetores não-únicos de estratégias racionalizáveis. O vencedor do prêmio Nobel, Thomas Schelling (1978), conjecturou e demonstrou empiricamente que em tais situações os jogadores podem tentar prever o equilíbrio procurando por **pontos focais**, isto é, aspectos de algumas estratégias que eles acreditam que serão proeminentes para outros jogadores, e que eles acreditam que os outros jogadores acreditarão ser proeminentes para eles. Por exemplo, se duas pessoas querem se encontrar em um determinado dia em uma cidade grande mas não podem se comunicar para organizar um horário e um local específicos, ambos poderiam, de maneira sensata, ir para a praça central mais proeminente da cidade ao meio-dia. Em geral, os melhores jogadores conhecem uns aos outros, ou, quanto mais frequentemente eles forem capazes de observar o comportamento estratégico uns dos outros, mais inclinados estarão em obter sucesso na busca por pontos focais com os quais se coordenar.

A coordenação foi, de fato, o primeiro tópico de aplicação da teoria dos jogos que veio à atenção difundida dos filósofos. Em 1969, o filósofo David Lewis (1969) publicou seu livro **Convention**, em que o *framework* conceitual da teoria dos jogos foi aplicado a uma das questões fundamentais da epistemologia do

século XX, a natureza e extensão das convenções que governam a semântica e sua relação com a justificação de crenças proposicionais. A intuição básica pode ser entendida com um exemplo simples. A palavra "galinha" denota galinhas e "avestruz" denota avestruzes. Nós não estaríamos melhor ou pior se "galinha" denotasse ostras e "avestruz" denotasse galinhas; contudo, **estaríamos** pior se metade de nós usasse o par de palavras da primeira maneira e a outra metade da segunda maneira, ou se todos nós randomizássemos entre elas para se referir de modo geral a pássaros que não voam. Obviamente, esse *insight* precede bastante Lewis; mas o que ele reconheceu é que essa situação tem a forma lógica de um jogo de coordenação. Assim, enquanto convenções particulares podem ser arbitrárias, as estruturas interativas que as estabilizam e as mantêm não o são. Além disso, os equilíbrios envolvido na coordenação de significados de substantivos parece ter um elemento arbitrário apenas porque não podemos colocá-los em uma hierarquia de Pareto; mas Millikan (1984) mostra implicitamente que, a esse respeito, eles são casos atípicos de coordenação linguística. Eles são certamente casos atípicos de convenções coordenadoras em geral, um ponto no qual Lewis se enganou por supervalorizar "intuições semânticas" acerca do "significado" de "convenção" (BACHARACH, 2006; ROSS, 2008a).

Ross e LaCasse (1995) apresentam o seguinte exemplo de um jogo de coordenação na vida real em que o EN não é Pareto indiferente, mas o EN Pareto inferior é mais frequentemente observado. Em uma cidade, os motoristas devem se coordenar em um de dois EN no que diz respeito ao seu comportamento quanto às luzes de semáforos. Ou todos devem seguir a estratégia de acelerar para tentar passar rapidamente as luzes que ficam amarelas (ou âmbar) e parar antes de seguir adiante quando as luzes vermelhas mudam para verdes, ou todos devem seguir a estratégia de reduzir a velocidade nos amarelos e imediatamente acelerar nas mudanças para o verde. Ambos os padrões são EN: uma vez que a comunidade tenha se coordenado com um deles, então nenhum indivíduo possui incentivos para se comportar diferente: aqueles que reduzem a velocidade nos amarelos enquanto outros passam em disparada terão a traseira batida, enquanto aqueles que passam em disparada nos amarelos no outro equilíbrio arriscarão uma colisão com aqueles que aceleram imediatamente nos verdes. Portanto,

assim que o padrão de trânsito de uma cidade se estabelece em um desses equilíbrios, ele tenderá a permanecer. E, de fato, esses são os dois padrões que são observados nas cidades do mundo. Contudo, os dois equilíbrios não são Pareto indiferentes, uma vez que o segundo EN permite mais carros virarem à esquerda em cada ciclo em uma jurisdição de direção pela mão esquerda, e à direita em cada ciclo em uma jurisdição de direção pela mão direita, o que reduz a principal causa de congestionamento em redes rodoviárias urbanas e permite a todos os motoristas esperarem maior eficiência na circulação. Infelizmente, por razões acerca das quais nós podemos apenas especular na espera por mais análises e trabalhos empíricos, muito mais cidades estão presas no EN de Pareto inferior do que no de Pareto superior. A teoria condicional dos jogos (*vide* **Seção 5**) fornece recursos promissores para modelar casos tais como esse, em que a manutenção de equilíbrio de jogos de coordenação deve ser apoiada por normas sociais estáveis, porque os jogadores são anônimos e encontram oportunidades regulares para obter vantagens isoladas abandonando o acordo de ajudar o equilíbrio prevalecente. Esse trabalho está atualmente em andamento.

Convenções acerca de *standards* de evidência e de racionalidade científica, os tópicos de filosofia da ciência que ofereceram o contexto para a análise de Lewis, são provavelmente do tipo classificáveis Paretamente. Enquanto vários arranjos poderiam ser EN no jogo social da ciência, como os seguidores de Thomas Kuhn gostam de nos relembrar, é altamente improvável que todos esses [arranjos] residam em uma única curva de indiferença Pareto. Fortemente representados na epistemologia, na filosofia das ciências e na filosofia da linguagem contemporâneas, esses temas são todos, ao menos implicitamente, aplicações de teoria dos jogos. (O leitor poderá encontrar uma ampla amostra de aplicações e referências a uma literatura mais ampla em (NOZICK, 1998)).

A maior parte dos jogos de coordenação sociais e políticos jogados por pessoas também possui esse aspecto. Infelizmente, para todos nós, armadilhas de eficiência representadas por EN Pareto inferior são aí extremamente comuns. E, por vezes, dinâmicas desse tipo dão origem ao mais terrível de todos os comportamentos coletivos humanos recorrentes. A análise de Hardin de dois episódios de genocídio recentes se baseia na ideia de que as propriedades biologicamente superficiais pelas quais as pessoas se classificam em grupos

raciais e étnicos servem de maneira altamente eficiente como pontos focais em jogos de coordenação, o que, por sua vez, produzem DPs mortais entre eles.

De acordo com Hardin, nem o desastre Iugoslavo nem o desastre de Ruanda eram DPs, para começo de conversa. Isto é, em nenhuma situação, em ambos os lados, a maioria das pessoas começou por preferir a destruição do outro à cooperação mútua. Contudo, a lógica mortal da coordenação, deliberadamente instigada por políticos egoístas, **criou** DPs dinamicamente. Alguns indivíduos sérvios (Hutus) foram encorajados a perceber seus interesses individuais como melhor atendidos por meio da identificação com interesses de grupos sérvios (Hutus). Isto é, eles descobriram que algumas de suas circunstâncias, tais como aquelas que envolvem a competição por trabalho, tinham a forma de jogos de coordenação. Assim, eles agiram de maneira a criar situações em que isso fosse verdadeiro para outros sérvios (Hutus) também. Eventualmente, assim que sérvios (Hutus) em número suficiente identificaram o interesse próprio com o interesse de grupo, a identificação se tornou quase universalmente **correta**, porque (1) o objetivo mais importante para cada sérvio (Hutu) era fazer, grosso modo, tudo o que os outros sérvios (Hutus) fariam, e (2) a coisa mais distintamente **sérvia** a se fazer, cuja execução sinalizou a coordenação, era excluir os Croatas (Tutsi). Isto é, as estratégias que envolvem tal comportamento excludente foram selecionados como um resultado de se ter pontos focais eficientes. Essa situação fez com que o interesse próprio individual - e individualmente ameaçado - de um croata (Tutsi) fosse maximizado pela coordenação em uma identidade de grupo assertiva dos croatas (Tutsi), o que aumentou ainda mais a pressão nos sérvios (Hutus) para coordenar, e assim por diante. Note que não faz parte dessa análise sugerir que os sérvios ou Hutus começaram as coisas; o processo poderia ter sido (mesmo que não tenha sido de fato) perfeitamente recíproco. Mas o resultado é medonho: sérvios e croatas (Hutus e Tutsis) parecem progressivamente mais ameaçadores uns aos outros à medida em que se reúnem para autodefesa, até que ambos vejam isso como um imperativo para antecipar os seus rivais e atacar antes de serem atacados. Se Hardin está certo - e o ponto aqui não é reivindicar que ele **está** certo mas, em vez disso, de assinalar a importância mundana em determinar quais jogos os agentes estão, de fato, jogando -, então a mera presença de um executor externo

(OTAN?) não teria mudado o jogo, apesar da análise de Hobbes, uma vez que o executor não poderia ter ameaçado nenhum dos lados com nada pior do que o que cada um temia do outro. O que era necessário era uma recalibração da avaliação de interesses, o que (discutivelmente) aconteceu na Iugoslávia quando o exército croata começou a vencer decisivamente, a tal ponto que os sérvios da Bósnia decidiram que seu interesse próprio/de grupo seria melhor atendido com a chegada das forças de paz da OTAN. O genocídio ruandês também terminou com uma solução militar; nesse caso, a vitória Tutsi. (Mas isso se tornou a semente para a guerra internacional mais mortal da Terra desde 1945, a Guerra do Congo de 1998-2006.)

Obviamente, não é o caso que a maior parte dos jogos repetidos levam a desastres. A base biológica da amizade em pessoas e outros animais é, parcialmente, uma função da lógica de jogos repetidos. A importância de recompensas alcançáveis via cooperação em jogos futuros leva aqueles que esperam interagir neles a serem menos egoístas do que a tentação encorajaria em jogos presentes.O fato de que tal equilíbrio se torna estável por meio de aprendizagem dá aos amigos o caráter lógico de investimentos acumulados, que a maior parte das pessoas sente um grande prazer em sentimentalizar. Além disso, cultivar interesses e sentimentos compartilhados fornece redes de pontos focais ao redor dos quais a coordenação pode ser progressivamente facilitada.

5. Raciocínio em Equipe e Jogos Condicionais

Seguindo a introdução de Lewis (1969) de jogos de coordenação na literatura filosófica, a filósofa Margaret Gilbert (1989) argumentou, contra Lewis, que a teoria dos jogos é o tipo errado de técnica analítica para se pensar sobre as convenções humanas. Dentre outros problemas, ela é muito “individualista”, ao passo que as convenções são essencialmente fenômenos sociais. Mais diretamente, ela afirmou que as convenções não são meramente produtos de decisões de muitas pessoas individuais, como poderia ser sugerido por um teórico que modelou convenção como um equilíbrio de um jogo de *n*-pessoas em que cada jogador era uma única pessoa. Preocupações semelhantes acerca dos fundamentos alegadamente individualistas da teoria dos jogos foram repetidos

pelo filósofo Martin Hollis (1998) e pelos economistas Robert Sugden (1993, 2000, 2003) e Michael Bacharach (2006). Em particular, isso motivou Bacharach a propor uma teoria do **raciocínio em equipe**, que foi completada por Sugden, juntamente com Nathalie Gold, após a morte de Bacharach. Essa teoria é central para a correta apreciação do valor de uma grande extensão recente à teoria dos jogos, a teoria dos **jogos condicionais** de de Wynn Stirling (2012).

Considere novamente a versão de lance único do Dilema do Prisioneiro como discutida na **Seção 2.4** e apresentada, com uma matriz invertida para facilitar a discussão posterior, como se segue:

		II C	D
I	C	2,2	0, 3
	D	3,0	1,1

(C é a estratégia de cooperar com o oponente (isto é, se recusar a confessar) e D é a estratégia de defecção com o oponente (isto é, confessar)). Muitas pessoas acham incrível quando um teórico dos jogos lhes diz que nesse jogo os jogadores designados com o honorífico "racional" devem escolher de tal modo a produzir o resultado (D, D). A explicação parece requerer apelo a formas muito fortes de individualismo tanto descritivo quanto normativo. Afinal de contas, se os jogadores atrelassem um valor maior ao bem social (para a sua sociedade de ladrões e com duas pessoas) do que ao seu bem-estar individual, eles poderiam se sair melhor individualmente também; a "racionalidade" jogo-teórica, objeta-se, produz um comportamento que é perverso mesmo do ponto de vista da otimização individual. Alguém poderia argumentar que os jogadores prejudicam seu próprio bem-estar porque eles se recusam, obstinadamente, a prestar qualquer atenção ao contexto social de suas escolhas. Sugden (1993) parece ter sido o primeiro a sugerir que os jogadores que merecem verdadeiramente ser chamados de "racionais", incluindo-se aqueles não-altruístas, iriam raciocinar **como um grupo** na versão de lance único do DP, isto é, cada um chegaria às suas escolhas de estratégias se perguntando "O que é o melhor para ***nós***?" em vez de "O que é o melhor para **mim**?".

Binmore (1994) argumenta vigorosamente que essa linha de criticismo

confunde a teoria dos jogos enquanto matemática com questões sobre quais modelos jogo-teóricos são mais aplicáveis comumente às situações em que as pessoas se encontram. Se os jogadores valorizam a utilidade de uma equipe da qual fazem parte acima de seus interesses mais estritamente individualistas, então isso deve ser representado nas recompensas associadas com um modelo jogo-teórico de suas escolhas. Na situação modelada acima como um DP, se a preocupação dos dois jogadores pela "equipe" forem fortes o suficiente para induzir uma mudança nas estratégias de D para C, então as recompensas na célula superior esquerda (interpretada cardinalmente) teria de ser aumentada para pelo menos 3. (**Em** 3, os jogadores seriam indiferentes quanto a cooperar ou a abandonar o acordo.) Portanto, obtemos a seguinte transformação do jogo:

		II	
		C	D
I	C	4,4	0, 3
	D	3,0	1,1

Esse não é mais um DP; é um **jogo da confiança**, que tem dois EN em (C, C) e (D, D), com o primeiro sendo Pareto superior ao último. Assim, se os jogadores encontram esse equilíbrio, não devemos dizer que jogaram estratégias que não são EN em um DP. Em vez disso, devemos dizer que o DP era o modelo errado para a situação deles.

O que está em questão aqui é a melhor escolha de uma convenção para aplicar a matemática às descrições empíricas. Binmore está claramente certo, e a maioria dos comentadores reconheceu isso, caso interpretemos as recompensas de jogos por referência às funções de utilidade com domínio irrestritos. Essa é a prática *standard* predominante tanto em economia quanto em teoria formal da decisão. Por alguns anos, essa questão foi considerada fechada na literatura sobre o assunto. Contudo,Sugden (2018) argumenta, em um trabalho recente, que há razões, independentes de considerações técnicas sobre quais convenções são as mais convenientes para representar interações empíricas como jogos, para evitar apelar a preferências em domínios irrestritos ao se analisar o bem-estar (isto é, ao fazer economia normativa). Com base nesse argumento, Sugden volta a usar modelos jogo-teóricos em que as recompensas estão

restritas a métricas objetivamente especificáveis, tal como retornos monetários. As questões substanciais em economia do bem-estar sobre as quais Sugden lança luz são interessantes demais para um crítico recusar razoavelmente a se engajar com elas por mera teimosia quanto a aderir a uma convenção de interpretação de representações de jogos. É muito cedo para avaliar se os avanços em análises de bem-estar que Sugden busca são sustentáveis sob pressão crítica. Caso não o sejam, então sua motivação por uma convenção alternativa na interpretação de recompensas se dissolverá. Todavia, penso ser mais provável que um período de inovação intensiva em economia de bem-estar está logo à nossa frente, e que, no decurso disso, os economistas e outros analistas se sentirão confortáveis em operar duas convenções representacionais diferentes a depender dos contextos do problema. Se esse for, de fato, o nosso futuro, então podemos antecipar um estágio mais adiante em que um novo formalismo é exigido para permitir que ambas as convenções cooperem sem confusão em uma única aplicação, porque os contextos de problemas tendem a não permanecer convenientemente isolados uns dos outros. Tais especulações, no entanto, se situam bem à frente do estado atual da teoria.

Retornemos à linha do desenvolvimento da teoria que se seguiu à acomodação generalizada da crítica de Binmore. Os executores científicos de Bacharach, Sugden e Gold (em BACHARACH, 2006, p. 171-173), diferentemente de Hollis e Sugden (1993), usam a convenção *standard* para a interpretação de recompensas, sob a qual os jogadores podem apenas ser modelados como cooperando em uma versão de lance único de DP se ao menos um jogador comete um erro. (Para algumas especificações de erro, (C, C) poderia surgir consistentemente com ERQ como o conceito de solução.) Sob essa pressuposição, Bacharach, Sugden e Gold argumentam que os jogadores humanos de jogos evitarão frequentemente elaborar situações de tal maneira que uma versão de lance único de DP seja o modelo correto de suas circunstâncias. Uma situação que agentes "individualistas" elaborariam como um DP poderia ser elaborada por agentes de "raciocínio em equipe" como na transformação do jogo de confiança acima. Note que o bem estar da equipe poderia fazer uma diferença para recompensas (cardinais) sem fazer diferença **suficiente** para superar a atratividade da defecção unilateral. Suponha que ele as tenha aumentado para

2.5 para cada jogador; então o jogo permaneceria um DP. Esse ponto é importante, uma vez que, em experimentos em que os sujeitos jogam sequências de DPs na versão de lance único (**não** DPs repetidos, visto que os oponentes nos experimentos mudam de rodada para rodada), a maioria dos sujeitos começa por cooperar, mas aprende a abandonar o acordo à medida que os experimentos progridem. No relato de Bacharach desse fenômeno, esses sujeitos elaboraram inicialmente o jogo como pensadores de equipe. Contudo, uma minoria dos sujeitos o elabora como pensadores individualistas e abandonam o acordo, tomando os lucros dos caroneiros. Desse modo, os pensadores de equipe reelaboram a situação para se defender. Isso introduz um aspecto crucial do relato de Bacharach. Pensadores individualistas e pensadores de equipe não são considerados tipos diferentes de pessoas. Ele sustenta que as pessoas oscilam entre ação individualista e participação em ação de equipe.

Considere agora o seguinte jogo de Coordenação Pura:

		II C	**II** D
I	C	1,1	0, 0
	D	0,0	1,1

Podemos interpretar isso como representando uma situação em que os jogadores são estritamente individualistas, e, assim, cada um é indiferente quanto aos dois EN de (C,C) e (D, D), ou são pensadores de equipe mas que não reconheceram que sua equipe se dá melhor caso eles se estabilizem em torno de um EN ao invés do outro. Caso cheguem a tal reconhecimento, talvez por encontrar um ponto focal, então o jogo de Coordenação Pura é transformado no seguinte jogo conhecido como ***Hi-Lo***:

		II t1	**II** t2
I	s1	10,10	0, 0
	s2	0,0	1,1

De modo crucial, a transformação aqui requer mais do que um **mero** raciocínio de

equipe. O jogador também precisa de pontos focais para saber qual dos dois equilíbrios de Coordenação Pura oferece a perspectiva menos arriscada para a estabilização social (BINMORE, 2008). Na verdade, Bacharach e seus executores estão interessados na relação entre jogos de Coordenação Pura e jogos ***Hi-lo*** por uma razão especial. Não parece implicar qualquer criticismo do EN como um conceito de solução que ele não favoreça um vetor de estratégia em detrimento de outro em um jogo de Coordenação Pura. Contudo, o EN **também** não favorece a escolha de (s1,t1) em detrimento de (s2,t2) no jogo *Hi-Lo* apresentado acima, porque (s2,t2) é também um EN. Nesse ponto, Bacharach e seus amigos adotam o raciocínio filosófico do programa de refinamento. É certo, eles se queixam, que a "racionalidade" recomenda (s1,t1). Portanto, eles concluem, os axiomas para o raciocínio de equipe devem ser embutidos nos fundamentos refinados da teoria dos jogos.

Não precisamos endossar a ideia de que conceitos de solução jogo-teóricos devem ser refinados a fim de acomodar um conceito geral intuitivo de racionalidade visando motivar o interesse na contribuição de Bacharach. O teórico dos jogos não-psicológico pode propor uma mudança sutil de ênfase: em vez de se preocupar sobre se nossos modelos deveriam respeitar uma norma de racionalidade centrada na equipe, poderíamos simplesmente apontar para as evidências empíricas de que as pessoas, e talvez outros agentes, parecem frequentemente fazer escolhas que revelam preferências condicionadas ao bem-estar de grupos com os quais elas estão associadas. Nessa medida, as suas ações são parcial ou inteiramente - e talvez aleatoriamente - identificadas com esses grupos, e, quando as modelarmos usando funções de utilidade, isso precisará ser levado em conta. Assim, poderíamos descrever melhor a teoria que queremos como uma teoria de escolhas centradas na equipe ao invés de uma teoria do **raciocínio** de equipe. Note que essa interpretação filosófica é consistente com a ideia de que algumas de nossas evidências, talvez mesmo a nossa melhor evidência, para a existência de escolhas centradas na equipe sejam psicológicas. Também é consistente com a sugestão de que os processos que levam as pessoas a alternar entre ações individualizadas e ações centradas na equipe são frequentemente não deliberativos ou conscientemente representados. O ponto é simplesmente que não precisamos seguir Bacharach em pensar na

teoria dos jogos como um modelo de raciocínio ou racionalidade a fim de sermos persuadidos de que ele identificou uma lacuna que gostaríamos de ter os recursos formais para preencher.

Desse modo, as escolhas das pessoas parecem **realmente** revelar preferências centradas na equipe? Exemplos *standard*, inclusive o próprio exemplo de Bacharach, são retirados de esportes coletivos. Os membros de tais equipes estão sob considerável pressão social para escolher ações que maximizam as perspectivas de vitória em detrimento de ações que aumentam as suas estatísticas pessoais. O problema com esses exemplos é que eles incorporam difíceis problemas de identificação no que diz respeito à estimativa de funções de utilidade; um jogador interessado estritamente em si próprio e que quer ser popular com os fãs poderia agir de maneira idêntica ao jogador centrado na equipe. Soldados em condições de batalha fornecem exemplos mais persuasivos. Embora tentar convencer soldados a sacrificar suas vidas em prol dos interesses de seus países seja frequentemente ineficaz, a maioria dos soldados pode ser levada a cometer riscos extraordinários na defesa de seus companheiros ou quando os inimigos ameaçam diretamente suas cidades natais e suas famílias. É fácil pensar em outros tipos de equipes com as quais a maioria das pessoas se identifica em parte ou na maior parte do tempo: grupos de projetos, pequenas companhias, comitês eleitorais políticos, sindicatos locais, clãs e laços familiares. Teorias sociais fortemente individualistas tentam construir tais equipes como equilíbrios em jogos entre pessoas individuais, mas nenhuma pressuposição embutida na teoria dos jogos (ou, nesse caso, a teoria econômica dominante) força essa perspectiva (*vide* GUALA, 2016 para uma revisão crítica de opções). Em vez disso, podemos supor que as equipes são frequentemente fundidas de modo exógeno por processos psicológicos e institucionais complexos inter-relacionados. Isso leva o teórico dos jogos a se engajar em uma tarefa matemática que consiste, não em modelar o raciocínio em equipe, mas, ao invés disso, em modelar escolhas que são condicionais à existência de dinâmicas de equipe.

Isso nos leva para a extensão de Stirling (2012) à teoria dos jogos para cobrir tais interações condicionais. O foco de Stirling é formalizar e derivar condições de equilíbrio para uma noção de preferência de grupo que, por um

lado, não seja uma mera aglomeração de preferências individuais, mas que também, por outro, não pressuponha simplesmente a existência de uma vontade coletiva transcendente que seja imposta aos indivíduos. O alvo intuitivo que Stirling tem em mente é aquele de processos pelos quais as pessoas derivam suas preferências reais parcialmente com base nas consequências comparativas para o bem-estar de grupo de diferentes perfis possíveis de preferências que os membros poderiam indiviualmente revelar. Uma restrição chave que Stirling respeita é que os conceitos de solução da teoria (ou seja, seus equilíbrios) devem **generalizar** formalmente os conceitos de solução *standard* (EN, EPS, ERQ), não **substituí-los**. A teoria condicional dos jogos deve ser uma "verdadeira" teoria dos jogos, não uma "pseudo" teoria dos jogos.

Vamos desenvolver a ideia intuitiva de condicionalização de preferência com mais detalhes. As pessoas podem frequentemente - e talvez tipicamente - postergar a resolução total de suas preferências até que obtenham mais informações sobre as preferências de outros que são seus companheiros de equipe atuais ou potenciais. O próprio Stirling fornece um exemplo simples (certamente muito simples) de Keeney e Raiffa (1976), em que um fazendeiro forma uma preferência clara entre diferentes condições climáticas para uma compra de terra somente após, e, em parte, à luz de, saber as preferências de sua esposa. Esse pequeno experimento de pensamento é plausível, mas não ideal enquanto uma ilustração, porque ele é facilmente combinado com noções vagas que poderíamos considerar sobre **fusão** de ações no casamento ideal - e é importante distinguir a dinâmica da condicionalização de preferências em equipes de agentes distintos do simples **colapso** de agências individuais. Então vamos construir um exemplo melhor. Imagine um(a) presidente(a) corporativo(a) consultando seu conselho de aversão de risco sobre se eles deveriam levar adiante uma perigosa e hostil tentativa pública de aquisição. Compare dois procedimentos possíveis que ele(a) poderia usar: (i) ele(a) envia um e-mail com essa ideia a cada membro do conselho uma semana antes da reunião; (ii) ele(a) propõe a ideia ao conselho **na** própria reunião. A maior parte das pessoas concordará que os dois procedimentos poderiam produzir resultados diferentes, e que a principal razão para isso é que no procedimento (i), mas não no (ii), alguns membros poderiam se fechar em opiniões pessoais que eles não teriam tempo de

elaborar caso recebessem informações sobre a disposição uns dos outros de desafiar o(a) presidente(a) em público assim que ouvissem a proposta pela primeira vez. Em ambos os procedimentos imaginados, no momento da votação, há conjuntos de preferências individuais a serem agregadas pelo voto. Mas é mais provável que algumas preferências no conjunto gerado pelo segundo processo eram **condicionais** quanto às preferências de outros. Como Stirling a define, uma preferência condicional de um agente é uma preferência influenciada por informações acerca das preferências de outros agentes (especificados).

Uma segunda noção formalizada na teoria de Stirling é a de **concordância**. Ela diz respeito à extensão de controvérsia ou discórdia que um conjunto de preferências, incluindo um conjunto de preferências condicionais, geraria se o equilíbrio entre elas fosse implementado. Membros e líderes de equipes nem sempre querem maximizar a concordância ao projetar todos os jogos internos como jogos de Garantia ou Hi-lo (embora, de modo semelhante, sempre quererão eliminar DPs). Por exemplo, um gerente poderia querer encorajar um grau de competição entre setores que geram lucro em uma empresa, enquanto espera que os setores de gastos se identifiquem completamente com a equipe como um todo.

Stirling define formalmente teoremas de representação para três tipos de funções de utilidade ordenada: utilidade condicional, utilidade concordante e utilidade concordante condicional. Elas podem ser aplicadas recursivamente, isto é, para indivíduos, para equipes, e para equipes de equipes. Portanto, o núcleo do desenvolvimento formal é a teoria que agrega preferências concordantes condicionais de indivíduos para construir modelos de escolhas de equipe que não são impostas de modo exógeno aos membros da equipe, mas, ao invés disso, derivam de suas várias preferências. Aqui, ao enunciar o procedimento de aglomeração de Stirling, é útil alterar a sua terminologia, e, portanto, parafraseá-lo ao invés de citá-lo diretamente. A razão disso é que Stirling se refere a “grupos” em vez de “equipes”. O trabalho inicial de Stirling na TCJ foi inteiramente independente do trabalho de Bacharach, de modo que não foi configurado dentro do contexto do raciocínio de equipe (ou o que poderíamos reinterpretar como escolhas centradas na equipe). No entanto, as ideias Bacharach fornecem um cenário natural no qual enquadrar a conquista técnica de Stirling enquanto um

enriquecimento da aplicabilidade da teoria dos jogos na ciência social (*vide* HOFMEYR; ROSS, 2019). Assim, podemos parafrasear suas cinco restrições à aglomeração como se segue:

1. **Condicionamento**: A ordem de preferência de um companheiro de equipe pode ser influenciada pelas preferências dos outros membros da equipe, ou seja, pode ser condicional. (A influência pode ser marcada como zero, e, nesse caso, a ordem de preferência condicional colapsa na ordem de preferência categorial para a TPR *standard*.)
2. **Endogenia**: Uma ordenação concordante para uma equipe deve ser determinada por interações sociais de suas subequipes. (Essa condição garante que as preferências da equipe não são simplesmente impostas às preferências individuais.)
3. **Aciclicidade**: As relações de influência social não são recíprocas. (Isso provavelmente parecerá uma restrição estranha à primeira vista: certamente, a maioria das relações de influência social **são** recíprocas entre pessoas de qualquer nível. No entanto, como notado anteriormente, precisamos manter a preferência condicional distinta da fusão de agente, e essa condição ajuda a fazer isso. Mais importante, enquanto questão de matemática, ela permite que as equipes sejam representadas por grafos diretos. A condição não é tão restritiva a respeito da flexibilidade de modelagem, como alguém poderia pensar a princípio, por duas razões. Primeiramente, ela só nos impede de representar um agente *j* influenciado por um outro agente *i* de influenciá-lo **diretamente**. Somos livres para representar *j* influenciando *k*, que, por sua vez, influencia *i*. Em segundo lugar, e mais importante, à luz da restrição de intercambialidade (abaixo), a aglomeração é insensível à ordenação de pares de jogadores entre os quais existe uma relação de influência social.)
4. **Intercambialidade**: As ordenações de preferência concordante são invariantes sob transformações representacionais equivalentes em relação às informações sobre preferências condicionais.
5. **Monotonicidade:** Se uma subequipe prefere a alternativa de escolha A ao invés de B, e todas as outras subequipes são indiferentes quanto a A e B, então a equipe não prefere B ao invés de A.

Sob essas restrições, Stirling prova um teorema da aglomeração que se segue de um resultado geral para atualizar a utilidade à luz de novas informações que foi desenvolvido por Abbas (2003). Cada membro individual da equipe calcula a preferência da equipe agregando preferência concordantes condicionais. O analista aplica então a **marginalização**. Sejam X^n uma equipe e $X^m = \{X_{j1}, ..., X_{jm}\}$ e $X^k = \{X_{i1}, ..., X_{ik}\}$ subequipes disjuntas de Xn. A **utilidade concordante marginal** de X^m em relação à subequipe $\{X^m, X^k\}$ é obtida pelo somatório de $\mathcal{A}^k$, produzindo

$$U_{x_m}(\alpha_m) = \sum_{\alpha_k} Ux_m x_k(\alpha_m, \alpha_k)$$

e a utilidade marginal do membro da equipe individual X_i é dada por

$$U_{x_m}(\alpha_m) = \sum_{\sim a_i} Ux_n(a_1, ..., a_n)$$

onde a notação $\sum_{\sim a_i}$ significa que a soma é de todos os argumentos com exceção de a_i (STIRLING, 2012, p. 62). Essa operação produz as preferências **não-condicionais** de indivíduos i *ex post* - isto é, atualizados à luz de suas preferências concordantes condicionais e informações a que são condicionadas, nomeadamente, as preferências concordantes condicionais da equipe. Uma vez que todas as preferências *ex post* de agentes tenham sido computadas, os jogos resultantes podem ser resolvidos via análise *standard.*

A construção de Stirling é, como ele diz, uma verdadeira generalização da teoria da utilidade *standard* de modo a tornar a utilidade não-condicionada ("categórica") um caso especial. Ela fornece uma base para a formalização de utilidade de equipe, que pode ser comparada com qualquer uma das seguintes: a utilidade categórica pré-condicionada de um indivíduo ou subequipe; a utilidade condicional de um indivíduo ou subequipe; ou a utilidade concordante condicional de um indivíduo ou subequipe. Uma vez que todas as preferências individuais em um problema de escolha da equipe tenham sido marginalizadas, as análises EN, EPS ou ERQ podem ser propostas como soluções para o problema desde que haja informação total sobre influências sociais. Situações de informações incompletas podem ser resolvidas usando-se os equilíbrios Byes-Nash ou

sequencial.

Caso o leitor tenha se esforçado para seguir o ponto geral das construções técnicas acima, podemos resumir a conquista da teoria condicional dos jogos (TCJ) de maneira sofisticada como se segue. A TCJ modela a propagação de fluxos de influência pela aplicação da sintaxe formal da teoria da probabilidade (por meio da operação de marginalização) à teoria dos jogos, e pela construção de representações teóricas em grafos. Uma preferência pode surgir na medida em que a influência social se propaga por meio de um grupo e os jogadores modulam suas preferências com base nas preferências de outros jogadores. Preferências de grupo não são uma base direta para a ação, mas encapsulam um modelo social ao incorporar as relações e interdependências entre os agentes. A TCJ nos mostra como derivar uma ordenação coordenada para um grupo que combina as preferências condicionais e categóricas de seus membros de maneira muito semelhante à maneira como, na teoria da probabilidade, a probabilidade conjunta de um evento é determinada por probabilidades condicionais e marginais. Desse modo, assim como a aplicação convencional da sintaxe de probabilidade é um meio de expressar a incerteza epistemológica do agente cognitivo a respeito da crença, estender essa sintaxe à teoria dos jogos nos permite representar a incerteza prática de um agente a respeito da preferência.

Se esse fosse o fim da história, a TCJ seria um pouco mais do que um mecanismo de pré-processamento para identificar jogos *standard*. A verdadeira inovação está em representar a influência de considerações de concordância na determinação de equilíbrio. O modelo social pode ser usado para gerar uma definição operacional de preferência de grupo, e para definir escolhas verdadeiramente coordenadas. Não há nenhuma pressuposição de que grupos necessariamente otimizem suas preferências ou de que agente individuais coordenem as suas escolhas. O ponto é meramente que podemos representar formalmente condições sob as quais agentes em jogos podem fazer o que pessoas reais frequentemente parecem fazer: adaptar e estabelecer suas preferências individuais à luz tanto do que os outros preferem quanto do que promove estabilidade e eficiência de um grupo. A ação de equipe é assim incorporado à teoria dos jogos ao invés de ser deixada como um constructo

psicológico exógeno que o analista deve investigar para acelerar a construção de um modelo jogo-teórico de agentes socialmente inseridos.

Em trabalhos posteriores, Stirling (2016) estende a TCJ para incorporar escolhas estratégicas sob incerteza. Stirling e Ross estão atualmente envolvidos em um projeto conjunto para aplicar a TCJ para modelar a estabilização estratégica e a manutenção de normas sociais no sentido de Bicchieri (2006).

6. Comprometimento

Em alguns jogos, um jogador pode melhorar seu resultado ao realizar uma ação que torne impossível para ele realizar, no jogo de movimento simultâneo correspondente, o que seria sua melhor ação. Essas ações são chamadas de **comprometimentos** e podem servir como alternativas aos reforços externos em jogos que, de outra forma, decidiriam por equilíbrios Pareto ineficientes.

Considere o seguinte exemplo hipotético (que **não** é um DP). Suponha que você possua um pedaço de terra adjacente ao meu, e que eu gostaria de comprá-lo de modo a expandir meu terreno. Infelizmente, você não quer vendê-lo pelo preço que estou disposto a pagar. Se agirmos simultaneamente - você anuncia um valor de venda, e eu, de modo independente, dou ao meu corretor um valor de oferta, não haverá venda alguma. Eu poderia então tentar mudar seus incentivos ao realizar uma jogada de abertura em que anuncio que construirei uma estação de tratamento de esgoto com um cheiro pútrido no meu terreno atrás do seu a menos que você o venda, induzindo-lhe, dessa forma, a abaixar seu preço. Há agora um jogo de movimento sequencial. Contudo, esse movimento ainda não muda nada. Se você se recusar a vender mesmo diante de minha ameaça, então não é do meu interesse concretizá-la, porque estarei provocando dano em mim mesmo ao provocar dano em você. Como você sabe disso, você deve ignorar a minha ameaça. Minha ameaça não é crível, é um caso de conversa fiada.

Todavia, eu poderia tornar minha ameaça crível ao me **comprometer** com ela. Por exemplo, eu poderia assinar um contrato com alguns fazendeiros com a promessa de supri-los com dejetos tratados (fertilizantes) da minha estação, mas incluir uma cláusula de rescisão no contrato, livrando-me da minha

obrigação somente caso eu possa dobrar o tamanho do meu terreno e colocá-lo para algum outro uso. Agora, a minha ameaça é crível: se você não vender seu terreno, eu estou comprometido a construir uma estação de tratamento de esgoto. Como você sabe disso, você tem agora um incentivo para me vender sua terra a fim de escapar de sua ruína.

Esse tipo de caso expõe uma de muitas diferenças fundamentais entre a lógica da maximização não-paramétrica e da maximização paramétrica. Em situações paramétricas, um agente nunca pode ficar pior por ter mais opções. (Mesmo se uma opção nova seja pior do que as opções com as quais começou, ele pode simplesmente ignorá-la.) No entanto, quando as circunstâncias são não-paramétricas, a estratégia de um agente pode ser influenciada em favor de outra caso as opções forem visivelmente restritas. O incêndio, provocado por Cortez, de seus próprios barcos (*vide* **Seção 1**) é uma instância disso, que serve para tornar literal a metáfora comum.

Outro exemplo ilustrará esse ponto, bem como ilustrará a aplicabilidade de princípios através de tipos de jogos. Para esse fim, construiremos uma situação imaginária que não é um DP - visto que apenas um jogador tem incentivos para a defecção -, mas que é um dilema social na medida em que seu EN na ausência de comprometimento é Pareto inferior em relação a um resultado alcançável **com** um dispositivo de comprometimento. Suponha que dois de nós desejem caçar um raro antílope de um parque nacional a fim de vender o troféu. Um de nós deve atrair o animal na direção da segunda pessoa, que espera de tocaia para atirar nele e colocá-lo em um caminhão. Certamente, você promete compartilhar os rendimentos comigo, mas a sua promessa não é crível. Assim que você tiver o animal, você não tem motivos para não levá-lo embora e embolsar o valor total. Afinal de contas, não posso fazer uma queixa policial sem ser preso também. Mas agora suponha que eu adicione o seguinte movimento de abertura ao jogo. Antes de nossa caçada, eu coloco no caminhão um alarme que só pode ser desligado digitando um código. Apenas eu sei o código. Se você tentar ir embora sem mim, o alarme soará e nós dois seremos pegos. Sabendo disso, você tem agora um incentivo para me esperar. O que é crucial aqui é que você **prefere** que eu ligue o alarme, uma vez que isso torna mais crível sua promessa de compartilhar os rendimentos. Se eu não fizer isso, deixando sua

promessa **não** crível, nós seremos incapazes de concordar em tentar o crime em primeiro lugar, e nós dois perderemos nossa chance de lucrar com a venda do troféu. Portanto, você se beneficia do meu impedimento de que você faça o que lhe é melhor em um subjogo.

Podemos agora combinar nossa análise de DPs e dispositivos de comprometimento na discussão da aplicação que primeiro tornou a teoria dos jogos famosa fora da comunidade acadêmica. O impasse nuclear entre as superpotências durante a Guerra Fria foi intensamente estudado pela primeira geração de teóricos dos jogos, muitos dos quais receberam suporte financeiro direta ou indiretamente do exército dos Estados Unidos. Poundstone (1992) fornece uma história relativamente "higienizada" desse envolvimento que já se tornou disponível para o historiador casual que se baseia em fontes secundárias além das reminiscências públicas dos teóricos. Recentemente, um estudo histórico mais ceticamente alerta e profissional foi realizado por Amadae (2016), que fornece contexto acadêmico para memórias ainda mais arrepiantes de um pioneiro da teoria dos jogos aplicada, participante no desenvolvimento da estratégia nuclear da Guerra Fria, responsável também pelo vazamento de arquivos secretos do Pentágono na guerra do Vietnã, Daniel Ellsberg (ELLSBERG, 2017). Uma história consistente com esses relatos mas que estimula uma menor dilatação de pupila no leitor é encontrada em ERICKSON, 2015.

Na narrativa convencional, o impasse nuclear entre os Estados Unidos da América (EUA) e a União das Repúblicas Socialistas Soviéticas (URSS) atribui a seguinte política a ambas as partes. Cada uma ameaçou responder um primeiro ataque da outra com um contra-ataque devastador. Esse par de estratégias recíprocas, que realmente significaria no final da década de 1960 explodir o mundo, era conhecido como "Destruição Mutuamente Assegurada" [*Mutually Assured Destruction* ou "MAD"]. Os teóricos dos jogos naquele tempo objetaram que MAD era realmente louca, porque armou um DP como resultado do fato de que as ameaças recíprocas não eram críveis. O raciocínio por trás desse diagnóstico ocorreu como se segue. Suponha que a URSS lance um primeiro ataque contra os EUA. Nesse ponto, o presidente americano encontra seu país já destruído. Ele não o traz de volta à vida ao explodir o mundo, então ele não tem

incentivos para levar adiante sua ameaça original de retaliação, que manifestadamente falha em atingir seu propósito. Uma vez que os russos podem antecipar isso, eles devem ignorar a ameaça de retaliação e atacar primeiro. Obviamente, os americanos estão em uma posição exatamente simétrica e também devem atacar primeiro. Cada potência reconhece esse incentivo por parte da outra, e então antecipa o ataque caso elas não se apressem em impedí-lo. Isso é o que devemos esperar, pois esse é o único EN do jogo, uma corrida entre as duas potências para serem as primeiras a atacar. A implicação clara disso é a destruição do mundo.

Essa análise causou preocupação e medo genuínos em ambos os lados durante a Guerra Fria, e tem a fama de ter produzido algumas tentativas extremas de produzir dispositivos de comprometimento. Algumas anedotas, por exemplo, contam que o presidente Nixon fez a CIA tentar convencer os russos de que ele era insano ou se encontrava frequentemente bêbado, assim eles acreditariam que ele lançaria um ataque retalhador mesmo quando não seria mais de seu interesse fazê-lo. Similarmente, alega-se por vezes que a KGB soviética, durante os últimos anos de Brejnev, fabricou relatórios médicos exagerando a extensão de sua senilidade tendo o mesmo objetivo em mente. Mesmo se essas histórias não forem reais, a sua persistente circulação indica compreensão da lógica de comprometimento estratégico. Em última análise, a simetria estratégica que preocupou os analistas do Pentágono era complicada e talvez quebrada por mudanças nas táticas de implementação de mísseis americanos. Eles equiparam uma frota mundial de submarinos com mísseis suficientes para lançarem sozinhos um contra-ataque devastador. Isso tornou a confiabilidade das redes de comunicações militares dos EUA menos direta, e, ao fazerem isso, introduziram um elemento de incerteza estrategicamente relevante. O presidente dos EUA provavelmente poderia estar menos certo de ser capaz de alcançar os submarinos e cancelar suas ordens de atacar se as perspectivas de sobrevivência americana tivessem se tornado desesperançosas. Obviamente, o valor disso na quebra da simetria dependia dos russos estarem cientes desse problema potencial. No clássico filme de Stanley Kubrick, **Dr. Fantástico**, o mundo é destruído por acidente porque os russos construíram uma máquina do juízo final que aciona automaticamente um ataque retaliador independentemente da

determinação de seu líder em seguir até o fim a ameaça implícita de MAD, **mas então a mantém em segredo**. Como resultado, quando um coronel americano inequivocamente louco lança mísseis contra a Rússia por conta própria, e o presidente americano tenta convencer sua contraparte soviética de que o ataque foi não-intencional, o primeiro ministro russo lhe conta timidamente sobre a máquina do juízo final. Agora, os dois líderes não podem fazem nada a não ser assistir, pesarosos, enquanto o mundo é explodido graças a um erro jogo-teórico.

Esse exemplo do impasse da Guerra Fria, embora famoso e de importância considerável na história da teoria dos jogos e da sua recepção pública, baseou-se então em análises que não eram muito sutis. Em primeiro lugar, os teóricos dos jogos militares estavam quase certamente enganados na medida em que modelaram a Guerra Fria em uma versão de lance único do DP. Por um lado, o jogo de equilíbrio nuclear estava emaranhado em jogos maiores, jogos de poder global de grande complexidade. Por outro, está longe de ser claro que, para qualquer uma das super potências, aniquilar a outra e evitar a autoaniquilação era, de fato, o resultado mais bem classificado. Caso não o fosse, em qualquer um ou em ambos os lados, então o jogo não era um DP. Um cínico poderia sugerir que os pesquisadores do setor de operações de ambos os lados estavam jogando uma estratégia astuta em um jogo sobre financiamento, um jogo que envolvia a cooperação mútua a fim de convencer seus respectivos políticos a alocar mais recursos para armas.

Em circunstâncias mais mundanas, a maioria das pessoas explora o ubíquo dispositivo de comprometimento que Adam Smith tornou a peça central de sua teoria da ordem social há muito tempo: para as pessoas, o valor de suas próprias **reputações**. Mesmo que eu seja avarento, posso desejar fazer com que as pessoas pensem que eu seja generoso ao oferecer gorjetas em restaurantes, inclusive em restaurantes em que nunca mais pretendo comer. Quanto mais eu faço esse tipo de coisa, mais invisto em uma reputação valorosa que eu poderia danificar gravemente por meio de um único ato de óbvia, e observada, mesquinhez. Assim, a minha reputação duramente conquistada de generosidade funciona como um mecanismo de comprometimento em jogos específicos, reforçando, ele mesmo, um reinvestimento continuado. Com o tempo, a minha benevolência pode se tornar habitual, e, consequentemente, insensível à

variações circunstanciais, ao ponto em que um analista não possui nenhuma justificação empírica restante para continuar a me modelar como tendo uma preferência pela avareza. Há uma boa parcela de evidências de que a hiper-sociabilidade dos seres humanos é apoiada por disposições biológicas evoluídas (encontradas na maioria das pessoas, mas não em todas) em sofrer emocionalmente por fofocas negativas e por ter medo dessas fofocas. As pessoas são também naturalmente dispostas a **gostar** de fofocar, o que significa que punir os outros por espalhar as notícias quando seus dispositivos de compromisso falham é uma forma de policiamento social que elas alegremente assumem, além de não acharem custoso. Uma boa característica dessa forma de punição é que ela pode ser removida sem deixar danos de longo prazo naquele que foi punido, diferentemente de (digamos) bater em pessoas com pedaços de pau. Essa é uma propriedade feliz de um dispositivo que tem como seu objetivo a manutenção de incentivos para contribuir para projetos sociais em conjunto; a colaboração é, geralmente, mais frutífera com companheiros de equipe cujos ossos não estejam quebrados. Assim, convenções de perdão desempenham também um papel estratégico nesse elegante mecanismo de comprometimento que a seleção natural construiu para nós. Finalmente, as **normas** são expectativas mútuas culturalmente evoluídas em um grupo de pessoas (ou, talvez, em alguns outros animais sociais inteligentes) que possui a propriedade adicional de que indivíduos que as violam podem punir **a si mesmos** se sentindo culpados ou envergonhados. Assim, eles podem frequentemente realizar ações cooperativas contra seus próprios e limitados interesses, mesmo quando ninguém mais está prestando atenção. As histórias religiosas, ou as filosóficas envolvendo uma "racionalidade" moral kantiana, são especialmente propensas a serem contadas como explicação de normas porque a base jogo-teórica subjacente não ocorre às pessoas, e as normas em questão podem funcionar mais efetivamente por essa mesma razão.

Embora as assim chamadas "emoções morais" sejam extremamente úteis para manter comprometimentos, elas não são necessárias para tanto. Instituições humanas maiores são, de modo notório, altamente obtusas em questões morais; contudo, o comprometimento é comumente crucial às suas lógicas funcionais. Por exemplo, um governo tentado a negociar com terroristas

para assegurar a liberação de reféns em uma ocasião particular pode se comprometer com uma estratégia "linha dura" para manter uma reputação de firmeza visando reduzir os incentivos dos terroristas em realizar futuros ataques. Um diferente tipo de exemplo é oferecido pela companhia aérea australiana Qantas. A Qantas nunca tinha sofrido um acidente fatal, e por um tempo (até ela ter sofrido alguns acidentes não-fatais embaraçosos para os quais, provavelmente, temia chamar a atenção) fez muito uso disso em sua propaganda. Isso significa que seus aviões **eram**, provavelmente, mais seguros do que a média pelo menos durante aquele período, ainda que a vantagem inicial fosse meramente um pouco de boa sorte estatística, porque o valor de sua capacidade para reivindicar um registro perfeito aumentava à medida que isso durava, e assim deu à companhia aérea incentivos contínuos para incorrer em maiores custos a fim de garantir a segurança. Ela provavelmente ainda possui incentivos para tomar um cuidado extra para evitar que o seu registro de fatalidades ultrapasse a linha de reputação mágica entre 0 e 1.

Certas condições têm de ser cumpridas se os efeitos de reputação devem subscrever o compromisso. A reputação de uma pessoa pode ter um valor permanente em uma série de jogos que ela jogar, mas, nesse caso, sua preocupação pelo valor da reputação deve ser levada em consideração nas recompensas em que se determina cada jogo específico em que ela entra. A reputação pode ser construída **por meio** de jogadas apenas no caso de um jogo repetido. Então, o valor da reputação deve ser maior para o seu cultivador do que o valor de sacrificá-la em **qualquer** rodada particular do jogo repetido. Assim, os jogadores podem estabelecer um comprometimento reduzindo o valor de cada rodada de modo que a tentação para quebrar o acordo em qualquer uma delas nunca seja alta o suficiente para constituir uma tentação de difícil resistência. Por exemplo, as partes de um contrato podem mudar as suas obrigações com pequenos incrementos para reduzir os incentivos em ambos os lados para voltar atrás. Assim, empreiteiros em projetos de construção podem ser pagos em parcelas semanais ou mensais. Similarmente, o Fundo Monetário Internacional frequentemente concede empréstimos para governos em pequenas parcelas, reduzindo, desse modo, os incentivos dos governos em violar condições de empréstimos uma vez que o dinheiro esteja em mãos; na verdade, os governos

podem preferir tais arranjos com o objetivo de remover a pressão política interna para o uso desviante do dinheiro. Obviamente, todos nós estamos familiarizados com casos em que a recompensa de uma defecção em uma rodada atual se torna muito alta relativamente ao valor de longo prazo de reputação para a cooperação futura e ficamos sabendo que o tesoureiro da sociedade fugiu durante a noite com os fundos. O comprometimento por meio da preocupação pela reputação é o cimento da sociedade, mas qualquer agente com tal liga natural estará longe de ser perfeitamente eficaz.

7. Teoria Evolutiva dos Jogos

Gintis (2009a, 2009b) se sente justificado em afirmar que "a teoria dos jogos é uma linguagem universal para a unificação das ciências comportamentais." Existem bons exemplos desse trabalho unificador. Binmore (1998, 2005a) modela a história social como uma série de convergências em equilíbrios cada vez mais eficientes em jogos de transação comumente encontrados, interrompida por episódios em que algumas pessoas tentam se mover para um novo equilíbrio afastando-se de caminhos de equilíbrios estáveis, o que resulta em catástrofes periódicas. (Por exemplo, Stalin tentou mudar sua sociedade para um conjunto de equilíbrio em que as pessoas se preocupavam mais com o futuro poder industrial, militar e político de seu Estado do que com suas próprias vidas. Ele não foi bem sucedido; mas seus esforços certamente criaram uma situação em que, por algumas décadas, muitos soviéticos davam muito menos importância à vida de **outras pessoas** do que o usual.) Uma perspectiva jogo-teórica parece, de fato, amplamente útil para compreender fenômenos nas ciências sociais. Na **Seção 4**, por exemplo, consideramos o reconhecimento de Lewis de que cada língua humana equivale a uma rede de equilíbrios de Nash em jogos de coordenação em torno da transmissão de informação.

Dada a época de seu trabalho, Lewis restringiu sua atenção à teoria dos jogos estática, em que os agentes são modelados como **escolhendo** deliberadamente estratégias dadas as funções de utilidade fixas de modo exógeno. Como um resultado dessa restrição, a sua explicação incitou alguns

filósofos a fazer uma busca equivocada por uma teoria analítica geral da racionalidade de convenções (como notado por Bickhard (2008)). Embora Binmore tenha criticado esse foco repetidamente ao longo de uma valorosa carreira de contribuições (*vide* a bibliografia deste capítulo para uma seleção), Gintis (2009a) recentemente isolou o problema subjacente com uma claridade e uma tenacidade particulares. O EN e o EPS são conceitos **frágeis** de soluções quando aplicados a mecanismos computacionais naturalmente evoluídos como cérebros animais (incluindo humanos). Como vimos na **Seção 3**, em jogos de coordenação (e outros) com múltiplos EN, o que é economicamente racional para um jogador fazer é altamente sensível aos estados de aprendizagem de outros jogadores. Em geral, quando os jogadores se encontram em jogos em que eles não possuem estratégias estritamente dominantes, eles só têm incentivos descomplicados para jogar estratégias EN ou EPS na medida em que pode ser esperado que outros jogadores encontrem **suas** estratégias EN ou EPS. Pode-se racionalmente esperar que uma teoria **geral** da racionalidade estratégica, do tipo que os filósofos têm buscado, cubra as contingências resultantes? Recorrer aos princípios bayesianos de raciocínio, como revimos na **Seção 3.1**, é o caminho *standard* para tentar incorporar tal incerteza em teorias da decisão racional e estratégica. Contudo, como Binmore (2009) argumenta seguindo o exemplo de Savage (1954), princípios bayesianos são plausíveis somente **como princípios da racionalidade** nos assim chamados "mundos pequenos", isto é, ambientes em que distribuições de risco são quantificadas em um conjunto de parâmetros conhecidos e enumeráveis, como na solução do nosso jogo de atravessar o rio da **Seção 3**. A ideia de que a regra de Bayes diz aos jogadores como "ser racional" é um tanto implausível em mundos grandes, nos quais funções de utilidade, conjuntos de estratégias e estrutura informacional são difíceis de estimar e estão sujeitas a mudança por influências exógenas contingentes. Mas então por que devemos esperar que os jogadores escolham estratégias EN, EPS ou de equilíbrio sequencial em uma ampla variedade de interações sociais?

Como Binmore (2009) e Gintis (2009a) enfatizam, se a teoria dos jogos deve ser usada para modelar comportamentos reais e naturais e a sua história, do lado de fora dos cenários de mundos pequenos em que microeconomistas (mas não macroeconomistas, cientistas políticos, sociólogos ou filósofos da ciência)

transitam, então precisamos de alguma explicação sobre o que seria atrativo nos equilíbrios em jogos, mesmo quando nenhuma análise pode identificá-los ao domar toda a incerteza de tal modo que ela possa ser representada como puro risco. Para fazer referência novamente ao tópico de Lewis, quando a linguagem humana se desenvolveu, não havia nenhum árbitro externo para se preocupar com e providenciar uma eficiência de Pareto fornecendo pontos focais para coordenação. Ainda assim, de alguma forma as pessoas em geral concordaram em usar dentro de comunidades linguísticas as mesmas palavras e construções para dizer coisas similares. Parece improvável que qualquer escolha explícita e deliberada de estratégia por parte de alguém tenha desempenhado um papel nesses processos. No entanto, a teoria dos jogos acabou fornecendo conceitos essenciais para compreender a estabilização de linguagens. Esse é um ponto marcante que apoia fortemente o otimismo de Gintis acerca do alcance da teoria dos jogos. Para compreendê-lo, devemos voltar a nossa atenção para jogos **evolutivos**.

A teoria dos jogos tem sido aplicada proveitosamente na biologia evolutiva, em que espécies e/ou genes são tratados como jogadores, desde o trabalho pioneiro de Maynard Smith (1982) e seus colaboradores. A teoria evolutiva (ou **dinâmica**) dos jogos constitui agora uma nova e significativa extensão matemática aplicável a muitos cenários além do biológico. Skyrms (1996) usa a teoria evolutiva dos jogos para tentar responder a questões que Lewis não pôde nem mesmo formular sobre as condições sob as quais a linguagem, os conceitos de justiça, a noção de propriedade privada e outros fenômenos gerais que interessam ao filósofos seriam suscetíveis de surgir. O que é novo acerca da teoria evolutiva dos jogos é que os movimentos não são escolhidos por meio de deliberação de agentes individuais. Em vez disso, os agentes são comumente geneticamente programados para estratégias particulares, e o sucesso de uma estratégia é definido em termos do número de cópias de si mesma que ela deixará para jogar nos jogos de sucessivas gerações, dada uma população em que outras estratégias com as quais ela interage são distribuídas em frequências particulares. Nesse tipo de cenário, as próprias estratégias são os jogadores, e os indivíduos que desempenham essas estratégias são seus meros executores, que recebem os custos e benefícios

imediatos associados aos resultados.

A discussão aqui seguirá de perto a de Skyrms. Começaremos por introduzir **a dinâmica do replicador**. Primeiramente, considere como a seleção natural funciona para alterar linhagens de animais, modificando, criando e destruindo espécies. O mecanismo básico é a **reprodução diferencial**. Qualquer animal com características **hereditárias** que aumenta seu **número esperado de descendentes** em um dado ambiente tenderá a deixar mais descendentes do que outros, desde que o ambiente permaneça relativamente estável. Esses descendentes serão mais propensos a herdar as características em questão. Portanto, a proporção dessas características na população aumentará gradualmente à medida que as gerações passam. Algumas dessas características podem **ser fixadas**, isto é, eventualmente tomar conta de toda a população (até que o ambiente mude).

Como a teoria dos jogos entra nisso? Frequentemente, um dos mais importantes aspectos do ambiente de um organismo será o das tendências comportamentais de outros organismos. Podemos pensar em cada linhagem como "tentando" maximizar sua aptidão reprodutiva (= frequências futuras de suas estruturas genéticas características) por meio da descoberta de estratégias que são otimizadas, dadas as estratégias de outras linhagens. Portanto, a teoria evolutiva é outro domínio de aplicação para a análise não-paramétrica.

Na teoria evolutiva dos jogos, não pensamos mais em indivíduos como escolhendo estratégias à medida que se movem de um jogo para outro. A razão disso é que os nossos interesses são diferentes. Estamos mais preocupados agora com a descoberta de quais equilíbrios são estáveis, e como eles mudam com o tempo, do que em encontrar o equilíbrio em jogos singulares. Portanto, modelamos agora **as próprias estratégias** como jogando umas contra as outras. Uma estratégia é "melhor" do que outra se for mais provável que deixe mais cópias de si mesma na próxima geração, quando o jogo será jogado novamente. Estudamos as mudanças em distribuição de estratégias na população à medida que o jogo se desenrola.

Para a teoria evolutiva dos jogos, introduzimos um novo conceito de equilíbrio, graças a Maynard Smith (1982). Um conjunto de estratégias, em alguma proporção particular (por exemplo, 1/3:2/3, 1/2:1/2, 1/9:8/9, 1/3:1/3:1/6:1/6

- sempre somando 1), está em um equilíbrio de **EEE** (Estratégia Evolutiva Estável) somente no caso em que (1) nenhum indivíduo jogando uma estratégia poderia aprimorar sua aptidão reprodutiva trocando para uma das outras estratégias na proporção, e (2) nenhum mutante jogando uma estratégia completamente diferente poderia se estabelecer na ("invadir a") população.

Os princípios da teoria evolutiva dos jogos são melhor explicados através de exemplos. Skyrms começa por investigar as condições sob as quais um senso de justiça - compreendido, para propósitos de sua análise específica, como uma disposição para ver divisões iguais de recursos como justas a não ser que considerações de eficiência em casos especiais sugiram o contrário - poderia surgir. Ele nos pede para considerar uma população em que os indivíduos se encontrem regularmente e devem negociar recursos. Comece com três tipos de indivíduos:

a. **Justos** sempre demandam exatamente a metade dos recursos.

b. **Gananciosos** sempre demandam mais da metade dos recursos. Quando um ganancioso encontra outro ganancioso, eles desperdiçam os recursos lutando entre si para obter esses recursos.

c. **Modestos** sempre demandam menos da metade dos recursos. Quando um modesto encontra outro modesto, eles pegam menos do que o total dos recursos disponíveis e desperdiçam alguns.

Cada encontro em que as demandas totais somam 100% é um EN daquele jogo individual. Similarmente, pode haver muitos equilíbrios dinâmicos. Suponha que os gananciosos demandem 2/3 dos recursos e que o modestos demandem 1/3. Desse modo, dado um pareamento aleatório para interação, as duas proporções seguintes são EEEs:

i. Metade da população é gananciosa e metade é modesta. Podemos calcular o resultado médio aqui. O modesto ganha 1/3 dos recursos em todo encontro. O ganancioso ganha 2/3 quando encontra um modesto, mas não ganha nada quando encontra outro ganancioso. Então, sua recompensa média é também 1/3. Isso é uma EEE porque o justo não pode invadir. Quando o justo encontra um modesto, ele recebe 1/2. Mas quando o justo encontra um ganancioso, ele não recebe nada. Então sua recompensa média é de apenas 1/4. Nenhum modesto

tem incentivos para mudar de estratégia, nem tampouco qualquer ganancioso. Um justo mutante, que surgisse na população, faria o pior de tudo, e então a seleção não encorajará a propagação de nenhum mutante do tipo.
ii. Todos os jogadores são justos. Todo mundo sempre recebe metade dos recursos, e ninguém pode se sair melhor trocando de estratégia. Os gananciosos que entram nessa população encontram justos e recebem uma recompensa média de 0. Modestos ganham 1/3 como antes, mas isso é menos do que a recompensa do justo de 1/2.

Note que o equilíbrio (i) é ineficiente, visto que a recompensa média ao longo de toda a população é menor. Contudo, assim como resultados ineficientes podem ser EN de jogos estáticos, eles também podem ser EEEs nos evolutivos.

Referimo-nos aos equilíbrios em que mais de uma estratégia ocorre como **polimorfismos**. Em geral, no jogo de Skyrms, qualquer polimorfismo em que o ganancioso demanda x e o modesto demanda $1 - x$ é um EEE. A questão que interessa ao estudante da justiça diz respeito à **likelihood** relativa com o qual esses diferentes equilíbrios surgem.

Isso depende das proporções de estratégias no estado populacional original. Se a população começa com mais de um justo, então há alguma probabilidade de que os justos encontrarão um ao outro, e receberão a recompensa média mais alta possível. Os modestos, por si só, não inibem a propagação de justos: apenas os gananciosos o fazem. Mas os gananciosos, por si só, dependem de ter modestos por perto para serem viáveis. Então, quanto mais justos existirem na população relativamente aos **pares** de gananciosos e modestos, melhor o justo se sairá na média. Isso implica um efeito de limiar. Se a proporção de justos cair abaixo de 33%, então a tendência será que eles entrem em extinção, porque não encontram uns aos outros com frequência suficiente. Se a população de justos subir acima de 33%, então a tendência será que eles se tornem fixos, porque seus ganhos extras quando encontram uns aos outros compensa suas perdas quando encontram gananciosos. Você pode ver isso ao notar que todos têm uma recompensa média de 1/3 quando cada estratégia é adotada por 33% da população. Portanto, qualquer elevação acima desse limiar por parte dos justos tenderá a torná-los fixos.

Esse resultado mostra que e como, dadas certas condições relativamente gerais, a justiça como a definimos **pode** surgir dinamicamente. A nova para os fãs da justiça se torna ainda melhor se introduzirmos **jogadas correlacionadas**.

O modelo que acabamos de considerar assume que as estratégias não são correlacionadas, isto é, que a probabilidade com que cada estratégia encontra toda outra estratégia é uma função simples de suas relativas frequências na população. Agora, examinaremos o que acontece em nosso jogo dinâmico de divisão de recursos quando introduzimos a correlação. Suponha que os justos tenham uma pequena habilidade de distinguir e procurar outros justos como parceiros de interação. Nesse caso, em média, o justos se saem melhor, e isso deve ter o efeito de abaixar seu limiar para se tornarem fixos.

Um modelador de jogos evolutivos estuda os efeitos da correlação e de outras restrições paramétricas por meio da realização de grandes simulações de computador em que as estratégias competem umas com as outras, rodada após rodada, no ambiente virtual. As proporções iniciais de estratégias, e qualquer grau escolhido de correlação, podem simplesmente ser definidas no programa. Pode-se então assistir as suas dinâmicas se desenrolarem no tempo e medir a proporção de tempo que elas permanecem em qualquer equilíbrio. Essas proporções são representadas pelos tamanhos relativos das **bacias de atração** para diferentes equilíbrios possíveis. Equilíbrios são pontos de atração em um espaço dinâmico; uma bacia de atração para cada ponto desses é então o conjunto de pontos no espaço ao qual a população convergirá até o equilíbrio em questão.

Ao introduzir a correlação nesse modelo, Skyrms, primeiramente, define o grau de correlação muito pequeno em 0.1. Isso faz a bacia de atração para o equilíbrio (i) encolher pela metade. Quando o grau de correlação é definido em 0.2, a bacia polimórfica se reduz ao ponto em que a população inicia um polimorfismo. Assim, aumentos muito pequenos na correlação produzem grandes aumentos proporcionais na estabilidade do equilíbrio em que todos desempenham o papel do justo. Uma pequena quantia de correlação é uma pressuposição razoável na maioria das populações, dado que vizinhos tendem a interagir uns com os outros e a imitar uns aos outros (seja geneticamente ou por causa de tendências em copiar deliberadamente uns aos outros), e porque animais genética e culturalmente similares são mais propensos a viver em

ambientes comuns. Desse modo, se a justiça pode surgir em todos, ela tenderá a ser dominante e estável.

Muito da filosofia política consiste em tentativas de produzir argumentos dedutivos normativos destinados a convencer um agente injusto de que ele tem razões para agir de modo justo. A análise de Skyrms sugere uma abordagem um tanto diferente. O justo se sairá melhor do que todos no jogo dinâmico caso ele realize passos ativos para preservar a correlação. Portanto, há uma pressão evolutiva para emergir tanto uma **aprovação moral da justiça** quanto **instituições justas**. A maioria das pessoas pode pensar que divisões 50-50 são "justas" e que valem a pena serem mantidas por meio de recompensas e sanções morais e institucionais **porque** somos os produtos de um jogo dinâmico que promoveu nossa tendência de pensar dessa maneira.

O tópico que tem recebido a maior parte da atenção por parte de teóricos evolutivos dos jogos é o **altruísmo**, definido como qualquer comportamento de um organismo que diminui sua própria aptidão esperada em uma única interação mas aumenta aquela do outro com quem interage. Ele é discutivelmente comum na natureza. Mas como ele pode surgir, dada a competição darwiniana?

Skyrms estuda essa questão usando a dinâmica do Dilema do Prisioneiro em seu exemplo. Trata-se simplesmente de uma série de jogos DP jogados por uma população, alguns membros da qual são desertores e outros são cooperadores. Como sempre em jogos evolutivos, as recompensas são medidas em termos do número de cópias esperadas de cada estratégia em gerações futuras.

Seja $\boldsymbol{U}(A)$a aptidão média da estratégia A na população e $\boldsymbol{U}$ a aptidão média da população inteira. Então, a proporção da estratégia A na próxima geração é apenas a razão $\boldsymbol{U}(A)/\boldsymbol{U}$. Desse modo, se A possui uma aptidão maior do que a média da população, então A aumenta. Se A possui uma aptidão menor do que a média da população, então A diminui.

Na dinâmica do DP, em que a interação é aleatória (ou seja, não há correlação), os desertores se saem melhor do que a média da população desde que os cooperadores estejam por perto. Isso se segue do fato de que, como vimos na **Seção 2.4**, a defecção é sempre a estratégia dominante em um único jogo. Portanto, 100% de defecção é o EEE no jogo dinâmico sem correlação, o

que corresponde ao EN na versão de lance único do DP estático.

Contudo, introduzir a possibilidade de correlação muda radicalmente o quadro. Precisamos agora calcular a aptidão média de uma estratégia **dada a sua probabilidade de encontrar cada outra estratégia possível**. No DP evolutivo, os cooperadores cujas probabilidades de encontrar outros cooperadores são altas se saem melhor do que os desertores cujas probabilidades de encontrar outros desertores são altas. Assim, a correlação favorece a cooperação.

A fim de ser capaz de dizer algo mais preciso sobre essa relação entre correlação e cooperação (e a fim de ser capaz de relacionar a teoria evolutiva dos jogos com problemas em teorias da decisão, um assunto que está além do escopo deste artigo), Skyrms introduz um novo conceito técnico. Ele chama uma estratégia de **adaptativamente ratificável** caso haja uma região ao redor do seu ponto de fixação no espaço dinâmico tal que haverá fixação a partir de qualquer lugar nessa região. Os tamanhos relativos das bacias de atração são altamente sensíveis aos mecanismos particulares pelos quais a correlação é alcançada. Para ilustrar esse ponto, ele constrói vários exemplos.

Um dos modelos de Skyrms introduz a correlação por meio de um **filtro** de pareamento para interação. Suponha que, na rodada 1 de um DP dinâmico, os indivíduos inspecionam uns aos outros e interagem, ou não, a depender do que eles encontram. Nas segunda rodada e nas subsequentes, todos os indivíduos que não parearam na rodada 1 são pareados aleatoriamente. Nesse jogo, a bacia de atração para desertores é grande **a não ser** que haja uma alta proporção de cooperadores na rodada um. Nesse caso, os desertores falham em parear na rodada 1, então ficam emparelhados, principalmente uns com os outros, na rodada 2 e levam uns aos outros à extinção. Um modelo que é mais interessante, porque o seu mecanismo é menos artificial, não permite que os indivíduos escolham seus parceiros, mas exige que eles interajam com aqueles que lhes são próximos. Devido ao parentesco genético (ou à aprendizagem cultural por cópia), os indivíduos são mais propensos a se assemelhar aos seus vizinhos do que a não se assemelhar. Se essa população (finita) está ordenada em uma única dimensão (isto é, em uma linha), e ambos cooperadores e desertores são colocados em posições ao longo dela de modo aleatório, então nós temos as seguintes dinâmicas. Cooperadores isolados têm aptidões esperadas menores do

que os desertores em volta e são levados localmente à extinção. Membros de grupos de dois cooperadores têm uma probabilidade de 50% de cada um interagir com um desertor. Como resultado disso, as suas aptidões médias esperadas permanecem menores do que as de seus vizinhos desertores, e eles enfrentam também uma provável extinção. Grupos de três cooperadores formam um ponto instável do qual tanto a extinção quanto a expansão são igualmente prováveis. Contudo, em grupos de quatro ou mais cooperadores, é garantido ao menos um encontro de um cooperador com outro suficiente para substituir o grupo original. Sob essa circunstância, os cooperadores enquanto grupo se saem melhor do que os desertores em volta e aumentam às suas custas. Eventualmente, os cooperadores **quase** se tornam fixos - mas não exatamente. Os desertores, sozinhos na periferia da população, atacam os cooperadores nas extremidades e sobrevivem como pequenas "comunidades criminosas". Assim, vemos que o altruísmo não apenas pode ser mantido pela dinâmica dos jogos evolutivos, mas, com correlação, podem mesmo se propagar e colonizar populações que não são originalmente altruístas.

Desse modo, a dinâmica darwiniana oferece boas novas qualificadas para a cooperação. Mas note que isso ocorre apenas enquanto os indivíduos estiverem presos à sua programação natural ou cultural e não puderem reavaliar suas utilidades por si mesmos. Se os nossos agentes ficarem muito espertos e flexíveis, eles podem notar que estão em DPs e que cada um se sairia melhor caso abandonasse o acordo. Nesse caso, eles eventualmente levarão a si mesmos à extinção - a não ser que desenvolvam normas morais estáveis e efetivas que funcionem para reforçar a cooperação. Mas, é claro, isso é justamente o que esperaríamos que evoluísse em populações de animais cujos níveis de aptidão média estão intimamente ligados às suas capacidades para uma cooperação social bem sucedida.Mesmo com isso, essas populações serão extintas a menos que elas se preocupem com as gerações futuras por alguma razão. Mas não há nenhuma razão não sentimental que já não pressuponha moralidade altruística para explicar por que os agentes **deveriam** se preocupar com as gerações futuras se cada nova geração substitui inteiramente a anterior em cada mudança de associados. Por esse motivo, os economistas usam modelos de "gerações sobrepostas" quando modelam jogos de distribuição

intertemporal. Os indivíduos na geração 1 que durarão até a geração 5 economizam recursos para os indivíduos da geração 3 com quem vão querer cooperar; na geração 3, os novos indivíduos se importam com a geração 6; e assim por diante.

Gintis (2009a) argumenta que, quando nos propusemos a usar a teoria evolutiva dos jogos para unificar as ciências comportamentais, deveríamos começar por usá-la para unificar a própria teoria dos jogos. Afirmamos em vários pontos neste artigo que EN e EPS são conceitos de solução problemáticos em muitas aplicações em que faltam regras institucionais explícitas, porque os agentes só possuem incentivos para jogar EN ou EPS caso estejam confiantes de que os outros agentes farão o mesmo. Na medida em que os agentes não possuem tal confiança - e isso, a propósito, é por si mesmo uma intuição graças à teoria dos jogos -, o que deve ser previsto é a desordem geral e a confusão social. Gintis mostra em detalhes como a chave para esse problema é a existência do que ele chama de um "coreógrafo". Ele quer dizer com isso algum elemento exógeno que informa os agentes sobre quais estratégias de equilíbrio eles devem esperar que os outros joguem. Como discutido na **Seção 6**, normas culturais são provavelmente os coreografos mais importantes para as pessoas. Funções de utilidade interessantes que incorporam normas do tipo relevante são estudadas extensivamente em (BICCHIERI, 2006). Nesse contexto, Gintis mostra mais um elemento unificador de grande importância: se os agentes atribuem uma utilidade positiva em seguir as sugestões do coreografo (isto é, em serem estrategicamente correlacionados com outros unicamente em função disso), então, onde quer que as recompensas potenciais concorrentes não sobrecarreguem esse incentivo, pode-se esperar também que os agentes estimem consistentemente prévias bayesianas, e chegar, desse modo, aos equilíbrios de crenças, como discutido na **Seção 3.1**, em jogos de informações imperfeitas. Finalmente, como discutido na **Seção 5**, a teoria condicional dos jogos promete fornecer os recursos para modelar a emergência endógena do coreografo dentro da dinâmica de jogos.

À luz disso, quando pensamos acerca do valor de modelos jogo-teóricos aplicados ao comportamento humano fora de mercados bem estruturados, muitas coisas dependem do que consideramos plausível e como fontes empiricamente validadas de incentivos das pessoas para se coordenarem umas com as outras.

Recentemente, isso tem sido um assunto de intenso debate, o que iremos rever na **Seção 8.3** abaixo.

8. Teoria dos Jogos e Evidência Comportamental

Em seções anteriores, revimos alguns problemas que surgem quando se trata a teoria dos jogos clássica (não-evolutiva) como uma teoria normativa que diz às pessoas o que elas devem fazer caso desejem ser racionais em situações estratégicas. Como vimos, a dificuldade é que não parece haver nenhum conceito de solução que possamos recomendar inequivocamente para todas as situações, especialmente situações em que os agentes possuem informações privadas. Contudo, na seção anterior, mostramos como o apelo a fundamentos evolutivos lança luz sobre as condições sob as quais funções de utilidade explicitamente elaboradas podem ser aplicadas de maneira plausível aos grupos de pessoas, o que leva a modelos jogo-teóricos com soluções plausíveis e estáveis. No entanto, não revisamos até então nenhuma evidência empírica real de observações ou experimentos comportamentais. A teoria dos jogos tem realmente ajudado os pesquisadores empíricos a fazerem novas descobertas sobre o comportamento (humano ou não-humano)? Se sim, qual tem sido, de modo geral, o conteúdo dessas descobertas?

Um problema epistêmico imediato surge quando abordamos essas questões. Não há forma alguma de aplicar a teoria dos jogos "por si só" independentemente de outras tecnologias modeladoras. Utilizando-se uma terminologia *standard* em filosofia da ciência, pode-se testar um modelo jogo-teórico de um fenômeno somente em conjunto com "suposições auxiliares" sobre o fenômeno em questão. No mínimo, isso se segue caso sejamos rigorosos quanto a tratarmos a teoria dos jogos puramente como matemática, com nenhum conteúdo empírico por si só. Em certo sentido, uma teoria sem nenhum conteúdo empírico nunca está aberta a teste; pode-se querer saber apenas se os axiomas sob as quais a teoria está baseada são mutuamente consistentes. Todavia, uma teoria matemática pode ser avaliada em relação à sua **utilidade** empírica. Um tipo de crítica filosófica que tem sido feita à teoria dos jogos, interpretada como uma ferramenta matemática para modelar fenômenos comportamentais, é que

sua aplicação exige sempre ou frequentemente recorrer a pressuposições falsas, enganadoras ou terrivelmente simplistas acerca desses fenômenos. Na medida em que as suposições auxiliares variam, esperaríamos que essa crítica tivesse diferentes graus de força em diferentes contextos de aplicação.

E assim as coisas acontecem. Não há nenhum domínio interessante em que as aplicações da teoria dos jogos tenham sido completamente incontroversas. Contudo, de modo geral, tem havido geralmente um consenso mais fácil quanto a como usar a teoria dos jogos (tanto clássica quanto evolutiva) para compreender o comportamento animal não-humano do que quanto a como empregá-la para a explicação e a previsão das atividades estratégicas de pessoas. Permita-nos, primeiramente, considerar brevemente problemas filosóficos e metodológicos que surgem em torno da aplicação da teoria dos jogos na biologia não-humana antes de dedicarmos maior atenção à ciência social jogo-teórica.

A modelagem jogo-teórica menos controversa tem aplicado a forma clássica da teoria em considerações de estratégias por meio das quais animais não-humanos buscam adquirir recursos básicos relevantes para a sua competição evolutiva: oportunidades de produzir descendentes propensos a se reproduzirem. A fim de maximizar sua aptidão esperada, os animais devem fazer escolhas otimizadas dentre vários bens intermediários, tais como nutrição, segurança contra predadores e habilidade de superar rivais na disputa por parceiros. Pontos de escolhas eficientes dentre esses bens podem ser frequentemente estimados por determinadas espécies em determinadas circunstâncias ambientais, e, com base nessas estimativas, podem ser derivados equilíbrios tanto paramétricos quanto não-paramétricos. Modelos desse tipo possuem um impressionante histórico na previsão e expliação de dados empíricos independentes a respeito de fenômenos estratégicos como disputas por alimento, seleção de parceiros, nepotismo, rivalidade entre irmãos, arrebanhamento, vigilância e sinalização coletiva contra predadores e mutualismo interespecífico (simbiose). (Para exemplos, *vide* KREBS; DAVIS, 1984, BELL, 1991, DUGATKIN; REEVE, 1998, DUKAS, 1998, e NOE; VAN HOOF; HAMMERSTEIN, 2001). Por outro lado, como Hammerstein (2003) observa, a reciprocidade, sua exploração e sua metaexploração são muito menos observadas em animais sociais não-humanos do que a modelagem jogo-teórica nos levaria a supor. Uma explicação disso,

sugerida por Hammerstein, é que animais não-humanos têm comumente menos habilidade para restringir seus parceiros de interação do que pessoas. A nossa discussão na seção anterior sobre a importância da correlação para estabilizar soluções de jogos dá suporte teórico a essa sugestão.

Por que a teoria dos jogos clássica tem ajudado a prever o comportamento de animais não-humanos mais diretamente do que ela o faz com a maioria dos comportamentos humanos? Presume-se que a resposta esteja nos diferentes níveis de complicação das relações entre suposições auxiliares e fenômenos. Ross (2005a) oferece o seguinte critério. Os problemas de maximização de utilidade e de maximização de aptidão são do domínio da economia. Uma teoria econômica identifica as unidades maximizadoras - agentes econômicos - com campos de preferência inalteráveis. A identificação de indivíduos biológicos inteiros com tais agentes é mais plausível quanto menos sofisticado cognitivamente for o organismo. Assim, insetos (por exemplo) são feitos sob medida para uma aplicação fácil da Teoria da Preferência Revelada (*vide* **Seção 2.1**). Contudo, encontramos animais que aprendem à medida em que os sistemas nervosos se tornam mais complexos. Aprender pode causar um grau suficiente de modificação permanente nos padrões de comportamento de um animal a ponto de podermos preservar a identificação do indivíduo biológico com um único agente em toda a modificação somente sob o custo de obtermos um vazio explicativo (porque as atribuições de funções de utilidade se tornam cada vez mais *ad hoc*) . Além disso, uma complexidade cada vez maior confunde a modelação simples em uma segunda dimensão: animais cognitivamente sofisticados não apenas mudam suas preferências com o tempo, mas também são governados por processos de controle distribuído que os tornam locais de competição entre agentes **internos** (SCHELLING, 1980; AINSLIE, 1992, 2001) Assim, eles não são agentes econômicos simples **nem** mesmo por um momento. Ao tentar modelar o comportamento de pessoas usando qualquer parte da teoria econômica, inclusive teoria dos jogos, devemos reconhecer que a relação entre qualquer pessoa e um agente econômico que construímos para fins de modelação será sempre mais complicada do que a simples identidade.

Não há nenhum ponto de passagem súbito em que um animal se torna cognitivamente sofisticado demais para ser modelado como um único agente

econômico, e, para todos os animais (inclusive humanos), há contextos em que podemos ignorar de maneira útil a dimensão sincrônica de complexidade. Contudo, encontramos uma mudança de fase em dinâmica de modelagem quando passamos de animais antissociais para animais sociais não-eusociais. (Isso se refere a animais que são sociais mas que não alcançam cooperação graças a mudanças fundamentais em sua genética populacional que tornam quase clones os indivíduos dentro dos grupos, como é o caso das formigas, abelhas, vespas, cupins e ratos-toupeira-pelados. Alguns exemplos conhecidos são papagaios, corvídeos, morcegos, ratos, caninos, hienas, porcos, guaxinins, lontras, elefantes, hiracoides, cetáceos, e primatas.) No caso desses animais, a estabilização da dinâmica de controle interno é parcialmente localizada **fora** dos indivíduos, no nível da dinâmica de grupo. Com essas criaturas, modelar um indivíduo como um agente econômico, com uma única e abrangente função de utilidade, é uma idealização drástica, que só pode ser realizada com a maior cautela metodológica e atenção a fatores contextuais específicos relevantes para o exercício de modelagem específico. Nesse caso, aplicações da teoria dos jogos só podem ser empiricamente adequadas na medida em que a modelagem econômica seja empiricamente adequada.

O *H. Sapiens* é aqui o caso extremo. Indivíduos humanos são socialmente controlados em um grau extremo em comparação com a maioria das outras espécies não-eusociais. Ao mesmo tempo, sua grande plasticidade cognitiva lhes permite variar significativamente entre culturas. Desse modo, pessoas são os agentes econômicos menos simples dentre todos os organismos. (Pode ser considerado irônico que elas foram tomadas, originalmente, e por muitos anos, como sendo os exemplos mais emblemáticos de agentes econômicos, em função de sua alegada "racionalidade" superior.) Consideraremos abaixo as implicações disso para as aplicações da teoria dos jogos.

Mas antes são necessários alguns comentários a respeito da adequação empírica da teoria **evolutiva** dos jogos para explicar e prever distribuições de disposições estratégias em populações de agentes. Tal modelagem é aplicada tanto a animais como produtos da seleção natural (HOFBAUER; SIGMUND, 1998)) quanto a animais sociais não-eusociais (mas especialmente humanos)

como produtos da seleção cultural (YOUNG, 1998). Há dois tipos de suposições auxiliares que devem ser justificadas, relativamente a um caso específico, ao se construir essas aplicações. Primeiramente, deve-se possuir fundamentos para a confiança de que as disposições que se busca explicar são **adaptações** (sejam biológicas ou culturais, conforme o caso) - isto é, disposições que foram selecionadas e são mantidas em razão do modo como elas promovem sua própria aptidão ou a aptidão do sistema mais amplo, ao invés de serem acidentes ou subprodutos estruturalmente inevitáveis de outras adaptações. (para uma discussão geral dessa questão, *videvide* DENNETT, 1995) Em segundo lugar, deve-se ser capaz de definir o empreendimento de modelagem no contexto de um conjunto justificado de suposições sobre as interrelações entre os processos evolutivos abrigados em diferentes escalas de tempo. (Por exemplo, no caso de uma espécie com dinâmica cultural, como a lenta evolução genética restringe a rápida evolução cultural? Como a evolução cultural retroalimenta a evolução genética, se é que ela a retroalimenta de fato? Para uma discussão magistral dessas questões, *vide* STERELNY, 2003). Visões conflitantes sobre quais suposições devem ser feitas acerca da evolução humana são a base para acaloradas discussões atuais na modelagem jogo-teórica evolutiva das disposições e instituições do comportamento humano. É aqui que questões na teoria evolutiva dos jogos se encontram com questões no campo em expansão da teoria dos jogos **comportamental-experimentais**. Descreverei primeiramente o segundo campo antes de dar uma ideia das controvérsias mencionadas há pouco, que constituem agora o campo mais acalorado de discussão filosófica acerca dos fundamentos da teoria dos jogos e de suas aplicações.

8.1. Teoria dos Jogos no Laboratório

Os economistas têm testado teorias realizando experimentos em laboratório com humanos e outros animais desde o trabalho pioneiro de Thurstone (1931). Em décadas recentes, o volume de tal trabalho tem se tornado positivamente enorme. A sua grande maioria coloca sujeitos em ambientes de problema microeconômico que não são perfeitamente competitivos. Uma vez que essa é precisamente a condição em que a microeconomia colapsa na teoria dos

jogos, a maior parte da economia experimental tem sido a teoria experimental dos jogos. Desse modo, é difícil distinguir questões experimentalmente motivadas sobre a adequação empírica da teoria microeconômica das questões sobre a adequação empírica da teoria dos jogos.

Aqui, podemos dar apenas uma visão geral de uma enorme e complicada literatura. Os leitores são direcionados para pesquisas críticas em KAGEL; ROTH, 1995; CAMERER, 2003, SAMUELSON, 2005, e GUALA, 2005. Um princípio útil de alto nível para classificar a literatura a indexa em diferentes pressuposições auxiliares com as quais os axiomas da teoria dos jogos são aplicados. Em apresentações populares (por exemplo, ORMEROD, 1994), diz-se frequentemente que os dados experimentais geralmente refutam a hipótese de que pessoas são agentes econômicos racionais. Tais reivindicações são muito imprecisas para serem interpretações sustentáveis dos resultados. Todos os dados são consistentes com a visão de que as pessoas são agentes econômicos **aproximados,** pelo menos por períodos de tempo longos o suficiente para permitir análises jogo-teóricas de cenários particulares, no sentido mínimo de que seu comportamento poder ser modelado compativelmente com a Teoria da Preferência Revelada (*vide* **Seção 2.1**). Contudo, a TPR faz tão poucas exigências empíricas que isso não é tão surpreendente quanto muitos não-economistas supõem (ROSS, 2005a). O que está realmente em questão em muitos dos debates em torno da interpretação geral da evidência empírica é a extensão em que as pessoas são maximizadoras de utilidade esperada. Como vimos na **Seção 3.1**, a teoria da utilidade esperada (TUE) é geralmente aplicada em conjunto com a teoria dos jogos com o objetivo de modelar situações que envolvem incerteza - ou seja, a maioria das situações de interesse em ciência comportamental. Contudo, uma variedade de modelos estruturais alternativos de utilidade se prestam à cardinalização de preferências de Von Neumann -Morgenstern e são definíveis em termos de subconjuntos dos axiomas de utilidade subjetiva de Savage (1954). A utilidade empírica da teoria dos jogos seria colocada em questão apenas caso pensássemos que o comportamento de pessoas não fosse geralmente descrito por meio de fuvNMs cardinais.

O que a literatura experimental realmente aparenta mostrar é um mundo de comportamentos usualmente ruidosos do ponto de vista do teórico. O ruído em

questão surge de uma heterogeneidade substancial, tanto entre pessoas quanto entre vetores (de pessoa, de situação). Não há uma única função de utilidade estrutural tal que todas as pessoas ajam de modo a maximizar a função dessa estrutura em todas as circunstâncias. Frequentemente, as pessoas se comportam como maximizadores de utilidade esperada quando se deparam com problemas bem aprendidos em contextos que não são excessivamente exigentes, ou que são altamente e institucionalmente estruturados. Para revisões gerais de exemplos de problemas e evidências teóricas, *vide* SMITH, 2008, e BINMORE, 2007. Para uma sequência estendida de exemplos de estudos empíricos, *vide* os assim chamados experimentos de leilão duplo contínuo discutidos em PLOTT; SMITH, 1978, e SMITH, 1962, 1964, 1965, 1976, 1982. Como resultado, a teoria dos jogos clássica pode ser usada em tais domínios com alta probabilidade de previsão de comportamentos e de implementação de políticas públicas, como é demonstrado por dúzias de leilões governamentais extremamente bem sucedidos de bens e outros ativos projetados por teóricos dos jogos para aumentar a receita pública (BINMORE; KLEMPERER, 2002).

Em outros contextos, interpretar o comportamento de pessoas como, geralmente, maximização de utilidade esperada requer uma violência indevida à necessidade de generalidade na construção da teoria. Obtemos uma melhor previsão ao usarmos menos restrições específicas de casos se supusermos que os sujeitos maximizam de acordo com uma ou (comumente) **mais** de várias alternativas (que não serão descritas aqui porque elas não dizem respeito diretamente à teoria dos jogos): a teoria da perspectiva cumulativa (TVERSKY; KAHNEMAN, 1992), ou a teoria da utilidade alpha-nu (CHEW; MACCRIMMON, 1979), ou a teoria da utilidade de hierarquia dependente (QUIGGIN, 1982), (YAARI, 1987). (A última alternativa de fato denota uma família de especificações alternativas. Uma dessas, a especificação de Prelec (1998), tem emergido em uma massa acumulada de estimativas empíricas como a função de utilidade humana estatisticamente mais comum.) Harrisson e Rutstrom (2008) mostram como projetar e codificar **modelos de mistura de *likelihood* máxima**, que permite um modelador empírico aplicar uma gama dessas funções de decisão a um único conjunto de dados de escolha. A análise resultante identifica a proporção do conjunto de escolhas total melhor explicado por cada modelo na

mistura. Anderson (2014) leva essa abordagem ao estado atual da arte, demonstrando o valor empírico de incluir um modelo de processos psicológicos não-maximizadores em uma mistura juntamente com modelos econômicos maximizadores. Essa flexibilidade nova e efetiva no que diz respeito à modelagem de decisão que pode ser empregada em aplicações empíricas da teoria dos jogos alivia a maior parte da pressão de buscar ajustes nas estruturas jogo-teóricas elas mesmas. Assim, ela se encaixa bem com a interpretação da teoria dos jogos como parte do kit de ferramentas matemáticas do cientista comportamental, em vez de um modelo empírico de primeira ordem da psicologia humana.

Uma ameaça mais séria à utilidade da teoria dos jogos é a evidência da reversão sistemática de preferências, tanto em humanos quanto em outros animais. Isso é mais sério tanto porque se estende para além do caso humano quanto porque desafia a Teoria da Preferência Revelada (TPR) em vez de desafiar apenas o compromisso desnecessariamente rígido com TUE. Como explicado na **Seção 2.1**, a TPR, diferentemente da TUE, está entre as fundamentações axiomáticas da teoria dos jogos interpretada não-psicologicamente. (Nem todos os escritores concordam que o aparente fenômeno da reversão de preferências ameaça a TPR ao invés da TUE; mas *vide* as discussões em CAMERER, 1995, p.660-665, e ROSS, 2005a, p. 177-181. Uma base para reversões de preferências que parecem ser comuns em animais com cérebros é o **desconto hiperbólico do futuro** (STROTZ, 1956), (AINSLIE, 1992). Esse é o fenômeno através do qual agentes descontam prêmios futuros mais abruptamente em distâncias temporais próximas do ponto de referência atual do que em distâncias temporais mais remotas. Isso é melhor compreendido por contraste com a ideia encontrada na maioria dos modelos econômicos tradicionais de desconto **exponencial**, em que há uma relação linear entre a taxa de mudança na distância de uma recompensa e a taxa em que o valor da recompensa declina do ponto de referência. A figura abaixo mostra curvas exponenciais e hiperbólicas do mesmo intervalo de um ponto de referência a uma recompensa futura. O de baixo representa a função hiperbólica; a forma curvada resulta da mudança na taxa de desconto.

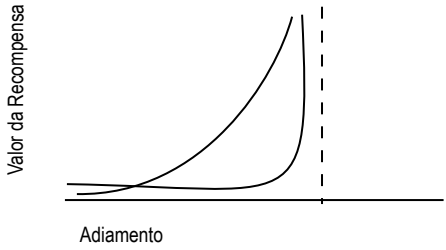


Figura 15

Um resultado disso é que, à medida em que perspectivas posteriores se aproximam mais do ponto de possível consumação, pessoas e animais irão, por vezes, gastar recursos desfazendo as consequências de ações anteriores que também lhes custam recursos. Por exemplo: ao decidir hoje se corrijo uma pilha de redações de graduandos ou se assisto a um jogo de baseball, eu decido procrastinar, apesar de saber que coloco fora de alcance alguma possibilidade mais divertida que possa surgir amanhã (quando haverá um um jogo de baseball igualmente atrativo caso uma melhor opção não surja). Até então, isso pode ser explicado de uma forma que preserve a consistência de preferências: se o mundo poderia terminar hoje à noite, com uma probabilidade bem pequena mas diferente de zero, então há algum nível de aversão ao risco no qual eu preferiria deixar as redações não corrigidas. A figura abaixo compara duas curvas de desconto exponencial, a curva menor para o valor do jogo que assisto antes de terminar minhas correções, e a curva maior para o jogo mais valoroso que aproveito depois de completar meu trabalho. Ambas possuem maior valor do ponto de referência quanto mais próximas estiverem dele; mas as curvas não se cruzam, então minhas preferências reveladas são consistentes ao longo do tempo não importa o quão impaciente eu possa ser.

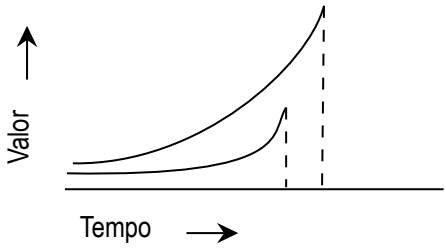


Figura 16

Mas se eu me precaver contra a procrastinação comprando um ingresso para o jogo de amanhã, quando na ausência da terrível tarefa eu não o teria feito, então violei a consistência de preferências intertemporal. Mais vividamente, caso eu estivesse em posição de escolher na semana passada se procrastino hoje, eu teria escolhido não fazê-lo. Nesse caso, minha curva de desconto desenhada do ponto de referência da semana passada cruza a curva desenhada da perspectiva de hoje, e minhas preferências são revertidas. A figura abaixo mostra essa situação.

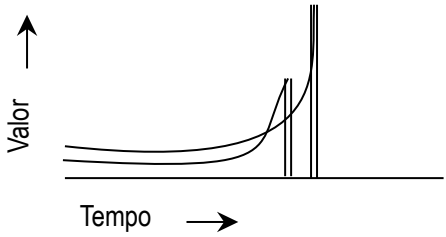


Figura 17

Esse fenômeno complica as aplicações da teoria dos jogos clássica a animais inteligentes. Contudo, ele claramente não a prejudica de modo completo, uma vez que pessoas (e outros animais) **não** revertem frequentemente suas preferências. (Se isso não fosse verdadeiro, os modelos de leilão bem sucedidos e outros assim chamados “projetos de mecanismos” seriam misteriosos.) Curiosamente, as principais teorias que visam explicar porque descontos

hiperbólicos poderiam frequentemente se comportar de acordo com os próprios RPTs apelam para princípios da teoria dos jogos. Ainslie (1992, 2001) produziu uma explicação de pessoas como comunidades de interesses internos de negociação em que subunidades baseadas em interesses de curto, médio e longo prazos enfrentam conflitos que elas devem resolver porque, caso não o façam, e em vez disso gerem um colapso hobbesiano interno (**Seção 1**), os agentes externos que evitam o resultado hobbesiano podem arruiná-las. O dispositivo do tirano hobbesiano não está disponível para o cérebro. Portanto, seu comportamento (quando a insanidade no sistema é evitada) é uma sequência de equilíbrios auto-aplicáveis do tipo estudado pela literatura da escolha pública jogo-teórica sobre negociações de coalizão em legislaturas democráticas. Isto é, a politica interna do cérebro consiste em trocas de favores (STRATMANN, 1997). Desse modo, essas dinâmicas internas são parcialmente reguladas e estabilizadas por jogos sociais mais amplos em que coalizões (pessoas como totalidades de subpartes temporais de suas biografias) são inseridas (ROSS, 2005a, p. 334-353). (Por exemplo: expectativas sociais acerca do papel de alguém como vendedor coloca metas de equilíbrios comportamentais para os processos de trocas de favores em seu cérebro.) Isso potencialmente adiciona elementos relevantes adicionais à explicação por que e como instituições estáveis com regras relativamente transparentes são condições-chave que ajudam pessoas a se assemelharem mais com agentes econômicos simples, tal que a teoria dos jogos clássica seja de modo confiável aplicável a elas como unidades inteiras.

Uma importante nota de cautela é aqui necessária. Muito da literatura comportamental recente toma como certo que descontos temporalmente inconsistentes são o caso *standard* ou predefinido para pessoas. Contudo, Anderson *et al.*(2008) mostra empiricamente que isso surge de (i) assumir que grupos de pessoas são homogêneos no que diz respeito a quais formas funcionais melhor descrevem seu comportamento de desconto, e (ii) falhas em obter e controlar de maneira independente os diferentes níveis de aversão ao risco das pessoas ao estimarem suas funções de desconto. Em uma gama de populações que têm sido estudadas com essas duas considerações em mente, os dados sugerem que descontos temporalmente consistentes descrevem

proporções substancialmente mais altas de escolhas do que escolhas temporalmente inconsistentes o fazem. Portanto, a generalização excessiva de modelos de desconto hiperbólico deve ser evitada.

8.2. Neuroeconomia e Teoria dos Jogos

A ideia de que a teoria dos jogos possa encontrar novas aplicações para a dinâmica interna de cérebros, conforme sugerido na seção anterior, tem sido desenvolvida a partir de motivações independentes pelo programa de pesquisa conhecido como **neuroeconomia** (MONTAGUE; BERNS, 2002),(GLIMCHER, 2003),(ROSS, 2005a, p. 320-334), (CAMERER; LOEWENSTEIN; PRELEC, 2005). Graças às novas tecnologias escaneadoras não invasivas, especialmente à imagem de ressonância magnética funcional (RMF), recentemente tem se tornado possível estudar atividades sinápticas de cérebros em funcionamento enquanto respondem a sinais controlados. Isso tem permitido um novo caminho de acesso - embora um caminho ainda bastante indireto (HARRISON; ROSS, 2010) - à computação do cérebro de valores esperados de prêmios, que são (naturalmente) tomados como tendo um papel crucial na determinação do comportamento. A teoria econômica é usada para enquadrar a derivação de funções maximizadas por computações de nível sináptico desses valores esperados; portanto o nome “neuroeconomia”.

A teoria dos jogos desempenha um papel de liderança na neuroeconomia em dois níveis. Primeiramente, a teoria dos jogos tem sido usada para prever os cálculos que neurônios individuais e grupos de neurônios que servem o sistema de prêmios devem realizar. No melhor exemplo divulgado, Glimcher (2003) e seus colegas escanearam via ressonância magnética macacos que eles treinaram para jogar os assim chamados “jogos de inspeção” contra computadores. Em um jogo de inspeção, um jogador enfrenta uma série de escolhas entre trabalhar por um prêmio, nesse caso, onde ele certamente o receberá, ou de realizar outra ação, mais fácil (“fugir”), caso no qual ele receberá o prêmio somente caso o outro jogador (o “inspetor”) não estiver o monitorando. Suponha que o comportamento do primeiro jogador (o “trabalhador”) revela uma função de utilidade limitada em cada extremidade como se segue: ele trabalhará

em toda ocasião caso o inspetor sempre o monitorar, e ele fugirá em toda ocasião caso o inspetor nunca o monitorar. O inspetor prefere obter a maior quantia possível de trabalho pela menor taxa de monitoramento possível. Nesse jogo, o único EN para ambos os jogadores está em estratégias mistas, uma vez que qualquer padrão na estratégia de um jogador que possa ser detectado pelo outro pode ser explorado. Qualquer par de estratégias em que, em cada tentativa, o trabalhador é indiferente entre trabalhar e fugir, ou o inspetor é indiferente entre monitorar e não monitorar, é um EN, para qualquer par de funções de utilidade específicas para os dois jogadores que atendem às restrições acima.

Aplicar análises de jogos de inspeção a pares ou grupos de agentes requer **ou** que tenhamos justificado de maneira independente suas funções de utilidade sobre todas as variáveis relevantes para seu jogo, caso no qual podemos definir o EN e então testar para ver se elas maximizam de forma bem sucedida a utilidade esperada; **ou** assumir que elas maximizam a utilidade esperada, ou que obedeçam à alguma outra regra tal como uma função de pareamente, e então inferir suas funções de utilidade a partir de seus comportamentos. Qualquer um desses procedimentos pode ser sensato em diferentes contextos empíricos. Mas a vantagem epistemológica aumenta enormemente se a função de utilidade do inspetor for determinada de modo exógeno, como frequentemente é o caso. (Por exemplo, a polícia que implementa inspeções aleatórias nas estradas para pegar motoristas bêbados possui comumente uma incidência máxima de direção alcoolizada atribuída à elas como meta, e um orçamento definido de modo exógeno. Isso determina as suas funções de utilidade, dada uma distribuição de preferências e atitudes quanto ao risco dentre a população de motoristas.) No caso dos experimentos de Glimcher, o inspetor é um computador, então seu programa está sob controle experimental e seu lado da matriz de recompensas é conhecido. Representantes da utilidade esperada dos sujeitos (nesse caso, esguichos de suco de frutas para os macacos), podem ser determinados com antecedência em configurações de teste paramétrico. O computador é então programado com o modelo econômico dos macacos, e pode procurar por padrões explorável snos dados em seus comportamentos em condições de jogo, variando então as suas estratégias. Com essas variáveis fixas, o comportamento EN de maximização da utilidade esperada

por parte dos macacos pode ser calculado e testado manipulando-se a função de utilidade do computador em várias rodadas do jogo.

O comportamento dos macacos após treinamentos rastreia EN de forma bastante robusta (como o comportamento de pessoas que jogam jogos similares por prêmios monetários o fazem (GLIMCHER, 2003, p. 307-308)). Glimcher e seus colegas puderam então realizar os experimentos importantes ao trabalharem com macacos treinados. Os comportamentos dos macacos de trabalhar ou de fugir foram associados por seu treinamento com olhar para a direita ou para a esquerda em uma exibição visual. Em experimentos anteriores, Platt e Glimcher (1999) tinham estabelecido que, em cenários paramétricos, à medida em que prêmios em suco variavam de um bloco de tentativas para outro, as taxas de disparo de cada neurônio parietal que controla os movimentos oculares poderiam ser treinadas para codificar a utilidade esperada para o macaco de cada movimento possível relativamente à utilidade esperada do movimento alternativo. Desse modo, "movimentos que valeram 0.4ml de suco eram representados de maneira duas vezes mais forte [em probabilidades de disparo neural] do que movimentos que valeram 0.2ml de suco" (p. 314). Não surpreendentemente, as taxas de disparo também variavam quando as quantias de suco recompensadas por cada movimento variavam de um bloco de tentativas para outro.

Contra esse plano de fundo, Glimcher e seus colegas puderam investigar o modo como os cérebros dos macacos implementaram o rastreio do EN. Quando os macacos jogaram o jogo de inspeção contra o computador, a meta associada à fuga pôde ser estabelecida como o local ideal, dado o treinamento prévio, para um neurônio específico sob estudo, enquanto a meta de trabalho apareceria em uma localização nula. Isso permitiu a Glimcher testar a resposta à seguinte questão: os macacos mantiveram o EN no jogo ao manterem constante a taxa de disparos do neurônio enquanto o comportamento real e ideal do macaco como um todo variava? Os dados deram a resposta "sim" de maneira robusta. Glimcher interpreta razoavelmente esses dados como sugerindo que as taxas de disparos neurais, a menos nesta região cortical para essa tarefa, codificam a utilidade esperada tanto em configurações paramétricas quanto em não-paramétricas. Aqui, nós temos uma aparente defesa da aplicabilidade empírica da teoria dos jogos clássica em um contexto independente de instituições ou convenções

sociais.

Análises adicionais levaram a hipótese mais a fundo. O computador que interpretava o inspetor foi apresentado com a mesma sequência de resultados que seu oponente macaco havia recebido no jogo do dia anterior, e, para cada movimento, foi solicitado avaliar os valores relativos esperados das ações de fuga e de trabalho disponíveis no próximo movimento. Glimcher relata uma correlação positiva entre pequenas flutuações em torno das taxas de disparo EN estáveis no neurônio individual e os valores esperados estimados pelo computador ao tentar rastrear o mesmo EN. Glimcher comenta sobre essa descoberta que:

> Os neurônios pareceram estar refletindo um cálculo próximo àquele realizado pelo nosso computador em uma base de lance a lance ... [E]m uma ... escala [relativamente] microscópica, fomos capazes de utilizar a teoria dos jogos para começar a descrever os cálculos de decisão por decisão que os neurônios na área LIP estão realizando. (GLIMCHER, 2003, p. 317)

Desse modo, encontramos a teoria dos jogos indo além do seu papel tradicional enquanto uma tecnologia para enquadrar restrições de alto nível à dinâmica evolutiva ou ao comportamento por agentes bem informados que operam sob camisas de força institucionais. Nas mãos de Glimcher, ela é utilizada para modelar diretamente a atividade no cérebro de um macaco. Ross (2005a) argumenta que grupos de neurônios assim modelados não deveriam ser identificados com as unidades de jogos subpessoais encontradas na teoria de Ainslie sobre negociações intrapessoais descrita anteriormente; isso envolveria um tipo de redução direta que a experiência nas ciências comportamentais e da vida têm nos ensinado a não esperar. Essa questão tem surgido desde então em uma disputa direta entre neuroeconomistas a respeito das interpretações rivais das observações RMF de escolhas e descontos intertemporais (MCCLURE *et al.*., 2004; GLIMCHER *et al.*., 2007). O peso das evidências até então favorece a visão de que, se é útil analisar, às vezes, as escolhas das pessoas como equilíbrios em jogos entre agentes subpessoais, os agentes subpessoais em questão não deveriam ser identificados com áreas do cérebro separadas. Infelizmente, a

interpretação contrária ainda é muito comum em literaturas menos especializadas.

Vimos o primeiro nível em que a neuroeconomia aplica a teoria dos jogos. Um segundo nível envolve condicionar variáveis em atividades neurais que poderiam impactar as escolhas de estratégias de pessoas quando elas jogam jogos. Isso envolve comumente repetir protocolos da literatura da teoria comportamental dos jogos com sujeitos de pesquisa deitados em *scanners* de RMF durante o jogo. Harrison (2008) e Ross (2008b) têm argumentado em favor do ceticismo sobre o valor de trabalhos desse tipo, que envolvem vários saltos de inferência desconfortavelmente grandes ao associar o comportamento observado com respostas neurais imputadas específicas. Também se pode questionar se muito conhecimento novo generalizável é obtido na medida em que tais associações podem ser identificadas com sucesso.

Permita-nos fornecer um exemplo desse tipo de "jogo em um *scanner*" - que envolve diretamente interação estratégica. King-Casas (2005) tomou um protocolo *standard* da teoria comportamental dos jogos, o assim chamado jogo "da confiança", e o implementou com sujeitos cujos cérebros eram escaneados em conjunto e usando uma tecnologia para conectar os mapas funcionais de seus respectivos cérebros, conhecida como "*hyperscanning*". Esse jogo envolve dois jogadores. Em seu formato repetido, como usado no experimento de King-Casas, o primeiro jogador é designado como o "investidor" e o segundo como o "administrador". O investidor começa com R$20, do qual ele pode manter para si qualquer porção de sua escolha enquanto investe o restante com o administrador. Nas mãos do administrador, a quantia investida é triplicada pelo experimentador. O administrador pode então devolver ao investidor o mesmo tanto ou uma quantia menor desse lucro que ele julgar adequado. O procedimento é executado por dez rodadas, com as identidades dos jogadores mantidas anônimas uns dos dos outros.

Esse jogo possui um número infinito de ENs. Os dados anteriores da economia comportamental são consistentes com a reivindicação de que o EN modal em jogos humanos **aproxima** ambos os jogadores utilizando estratégias de "olho por olho" (*vide* **Seção 4**) modificadas por defecções ocasionais para sondar informações, e algumas cooperações pós-defecção que manifestam tolerância (limitada) a tais sondagens. Esse é um resultado muito fraco, uma vez que ele é

compatível com uma vasta gama de hipóteses sobre exatamente quais variações de estratégias olho por olho são usadas e mantidas, e, assim, não permite inferências sobre dinâmicas potenciais sob diferentes condições de aprendizagem, de instituições ou de transferências culturais.

Quando executaram esse jogo sob o *hyperscanning*, os pesquisadores interpretaram suas observações como se segue. Pensou-se que os neurônios no núcleo caudado do administrador (geralmente vistos como implementando computações ou *outputs* de sistemas dopaminérgicos do mesencéfalo) mostravam uma forte resposta quando os investidores retribuíram a confiança com benevolência - isto é, responderam à defecção com maior generosidade. À medida que o jogo progredia, acreditava-se que essas respostas passavam de reacionárias a antecipatórias. Assim, os perfis de reputação como previstos por modelos jogo-teóricos clássicos foram inferidos como construídos diretamente pelo cérebro. Um aspecto adicional dos achados não previsíveis somente via modelagem teórica, e cuja observação puramente comportamental não havia sido suficiente para discriminar, foi que as respostas pelos neurônios caudais à reciprocidade malévola - isto é, à generosidade reduzida em resposta à cooperação - era significativamente menor em amplitude. Isso foi especulado como sendo um mecanismo pelo qual o cérebro implementa modificações de olho por olho de modo a impedir defecções ocasionais por sondagem de informações de desfazer a cooperação permanentemente.

O avanço em compreensão que os praticantes desse estilo de neuroeconomia esperam não consiste no que ele nos diz sobre tipos particulares de jogos, mas sim em inferências comparativas que facilita sobre as maneiras pelas quais o enquadramento contextual influencia as conjecturas das pessoas sobre quais jogos elas estão jogando. Conjectura-se que o RMF e outros tipos de sondagens de cérebros em funcionamento poderiam nos possibilitar estimar quantitativamente graus de **surpresas** estratégicas. As expectativas de interação recíprocas sobre surpresa podem elas mesmas estarem sujeitas à manipulação estratégica, mas essa é uma ideia que mal tem começado a ser teoricamente explorada por teóricos dos jogos (*vide* ROSS; DUMOUCHEL, 2004). A visão de alguns neuroeconomistas de que temos agora a perspectiva de testar empiricamente tais teorias novas, contrariamente a só modelá-las por hipótese,

tem estimulado crescimento nessa linha de pequisa.

8.3. Modelos Jogo-Teóricos da Natureza Humana

Os desenvolvimentos revisados na seção anterior nos trazem até a fronteira móvel das aplicações experimentais/comportamentais da teoria dos jogos clássica. Agora, podemos retornar ao ponto de ramificação no qual havíamos parado vários parágrafos atrás, no qual essa corrente de investigação se encontra com aquela vinda da teoria evolutiva dos jogos. Não há nenhuma razão séria para duvidar que, em comparação com outros animais não-eusociais - inclusive nossos parentes mais próximos, os chimpanzés e os bonobos -, os humanos realizaram façanhas prodigiosas de coordenação (*vide* **Seção 4**, e também TOMASELLO *et al.,* 2004). Atualmente, existe uma calorosa controvérsia, com importantes implicações filosóficas e desenvolvida em ambos os lados com argumentos jogo-teóricos, sobre a questão de saber se essa capacidade pode ser completamente explicada por adaptação cultural ou se ela é melhor explicada pela inferência de uma mudança genética no início da carreira do *H. sapiens*.

Henrich *et al.* (2004, 2005) realizaram uma série de jogos experimentais com populações extraídas de quinze sociedades de seres humanos pequenas na América do Sul, África e Ásia, inclusive três grupos de exploradores, seis grupos de horticultores de corte e queima, quatro grupos de pastores nômades, e dois grupos de agricultores de pequena escala. Todos os jogos (Ultimato, Ditatorial, e de Bens Públicos) que eles implementaram colocam os sujeitos em situações que são amplamente semelhantes ao jogo da Confiança que discutimos na seção anterior. Isto é, os jogos de Ultimato e de Bens Públicos são cenários em que tanto o bem-estar social quanto o bem-estar de cada indivíduo são otimizados (tem a eficiência Pareto alcançada) se, e somente se, ao menos alguns jogadores usam estratégias que não são estratégias de equilíbrio perfeito de subjogos (*vide* **Seção 2.4**). Em jogos Ditatoriais, o primeiro estritamente egoísta a se mover capturaria todos os lucros disponíveis. Assim, em cada um dos três tipos de jogos, os jogadores EPS que se preocuparam apenas com seu próprio bem-estar monetário receberiam resultados que envolvem recompensas altamente

desiguais. Em nenhuma dessas sociedades estudadas por Henrich *et al.* (ou qualquer outra sociedade em que jogos desse tipo sejam realizados) são observados tais resultados. Os jogadores cujos papéis são tais que eles ficariam com todo o lucro monetário, com exceção de épsilon, caso eles e seus parceiros jogassem EPS, sempre ofereceram aos seus parceiros de maneira substancial mais do que épsilon, e mesmo assim os parceiros recusaram, por vezes, tais ofertas ao custo de não receberem nenhum dinheiro. Além disso, diferentemente dos sujeitos tradicionais da economia experimental - estudantes universitários em países industrializados -, os sujeitos de Henrich *et al.* nem mesmo jogaram estratégias de Equilíbrio de Nash no que diz respeito aos resultados monetários. (Isto é, jogadores estrategicamente em vantagem ofereceram maiores porções de lucro àqueles estrategicamente em desvantagem do que era necessário para induzir a anuência com suas ofertas.) Henrich *et al.* interpreta esses resultados sugerindo que todas as pessoas reais, diferentemente do "homem econômico racional", valorizam resultados igualitários em alguma medida. Contudo, seus experimentos também mostram que essa medida varia significativamente com a cultura, e está correlacionada com variações em duas variáveis culturais específicas: as recompensas comuns à cooperação (até que ponto a vida econômica na sociedade depende da cooperação com parentes não imediatos) e a integração de mercado agregado (um constructo edificado a partir de graus medidos independentemente de complexidade social, anonimato, privacidade e tamanhos de assentamento.) À medida que os valores dessas duas variáveis aumentam, o comportamento do jogo muda (de forma fraca) na direção da jogada do equilíbrio de Nash. Portanto, os pesquisadores concluem que as pessoas são dotadas geneticamente com preferências por igualitarismo, mas o peso relativo dessas preferências é programável via processos de aprendizagem social condicionadas por sinalizações culturais locais.

Ao avaliar a interpretação de Henrich *et al.* desses dados, devemos notar primeiramente que nenhum axioma de TPR, ou de vários modelos de decisão mencionados na**Seção 8.1**, que são aplicados conjuntamente com a modelagem jogo-teórica dos dados de escolhas humanas, especificam ou implicam a propriedade de egoísmo estrito. (*vide* ROSS, 2005a, cap. 4;BINMORE, 2005b e 2009); e qualquer texto de economia ou de teoria dos jogos que permita que os

matemáticos falem por si). Portanto, a teoria dos jogos ortodoxa não prevê que as pessoas jogarão estratégias de EPS ou de EN derivadas pelo tratamento de recompensas monetárias como equivalentes à utilidade. Binmore (2005b) está então justificado em criticar Henrich *et al.* por sugerirem retoricamente que seus trabalhos empíricos envergonham a teoria dos jogos ortodoxa.

Isso não significa sugerir que a interpretação antropológica dos resultados empíricos deve ser tomada como incontroversa. Binmore (1994, 1998, 2005a, 2005b) argumentou por muitos anos, com base numa ampla gama de dados comportamentais, que, quando as pessoas jogam jogos com não familiares, elas tendem a aprender a jogar equilíbrios de Nash no que diz respeito às funções de utilidade que correspondem aproximadamente às funções de renda. Como ele aponta em (BINMORE, 2005b), os dados de Henrich *et al.* não testam essa hipótese para as suas sociedades de pequena escala, porque seus sujeitos não foram expostos aos jogos de teste pelo período de aprendizagem (um tanto longo, no caso do jogo de Ultimato) que os modelos teóricos e computacionais sugerem que são requeridos para as pessoas convergirem para EN. Quando as pessoas jogam jogos não familiares a elas, elas tendem a modelá-los por referência aos jogos com os quais estão acostumadas na experiência do dia a dia. Em particular, elas tendem a jogar versões **de lance único** de jogos de laboratório, embora fossem familiarizadas com jogos **repetidos**, uma vez que jogos em versão de lance único são raros na vida social normal fora dos contextos institucionais especiais. Muitos dos comentários interpretativos realizados por Henrich *et al.* são consistentes com essa hipótese a respeito de seus sujeitos, embora eles rejeitem explicitamente a hipótese ela mesma. O que é controverso aqui - os problemas de deixar de lado a teoria "ortodoxa" - é menos sobre o que os sujeitos particulares nesse experimento estiveram fazendo do que sobre o que os seus comportamentos deveriam nos levar a inferir sobre a evolução humana.

Gintis (2004, 2009a) argumenta que dados do tipo que temos discutido apoiam a seguinte conjectura acerca da evolução humana. Nossos ancestrais aproximaram maximizadores de aptidões individuais. Em algum lugar ao longo da linha evolutiva, esses ancestrais chegaram em circunstâncias em que um número suficiente deles otimizaram suas aptidões individuais ao agir de modo a otimizar o

bem-estar de seu grupo (SOBER; WILSON, 1998), tal que uma modificação genética se tornou fixa na espécie: desenvolvemos preferências não apenas sobre o nosso próprio bem-estar individual, mas também sobre o bem-estar relativo de todos os membros de nossas comunidades, indexadas às normas sociais **programáveis** em cada indivíduo via aprendizagem cultural. Assim, o pesquisador contemporâneo, ao aplicar a teoria dos jogos para modelar uma situação social, é aconselhado a desenterrar as funções de utilidade de seus sujeitos pela (i) descoberta de qual comunidade (ou comunidades) de que são membros, e então (ii) inferir a(s) função(ões) de utilidade programada(s) em membros dessa comunidade (comunidades), estudando representantes de cada comunidade relevante em uma variedade de jogos e assumindo que os resultados são equilíbrios coordenados. Visto que as funções de utilidade são as variáveis dependentes aqui, os jogos devem ser determinados de maneira independente. Gintis supõe que podemos comumente manter fixos ao menos os formatos estratégicos dos jogos relevantes em virtude de (a) nossa confiança de que as pessoas preferem resultados igualitários, todo o resto permanecendo igual, aos não igualitários dentro dos "grupos internos" culturalmente envolvidos a que percebem a si mesmos como pertencentes e (b) uma exigência de que equilíbrios de jogos são retirados de atratores estáveis em modelos jogo-teóricos evolutivos plausíveis da dinâmica histórica da cultura.

O requerimento (b) como uma restrição à modelagem jogo-teórica das disposições estratégicas humanas gerais não é mais muito controverso - ou, ao menos, não é mais controverso do que a capacidade adaptativa genérica na antropologia evolutiva da qual ela é uma expressão. No entanto, alguns comentadores são céticos quanto à sugestão de Gintis de que haveria uma descontinuidade genética na evolução na sociabilidade humana. (Para uma antropologia cognitiva-evolutiva que nega explicitamente tal descontinuidade, *vide* STERELNY, 2003). Parcialmente com base em tal ceticismo (mas mais diretamente nos dados comportamentais), Binmore (2005a, 2005b) resiste a modelar pessoas como dotadas de preferências auto engendradas pelo igualitarismo. De acordo com o modelo de Binmore (1994, 1998, 2005a), a classe básica de problemas estratégicos enfrentados por animais sociais não-eusociais são jogos de coordenação. Comunidades humanas desenvolvem normas

culturais para selecionar equilíbrios nesses jogos, e muitos desses equilíbrios serão compatíveis com altos níveis de comportamentos aparentemente altruístas em alguns (mas não em todos os) jogos. Binmore argumenta que as pessoas adaptam suas concepções de justiça a quaisquer que sejam suas regras de seleção de equilíbrio localmente prevalentes. No entanto, ele sustenta que o desenvolvimento da **dinâmica** de tais normas deve ser compatível, a longo prazo, com a negociação de equilíbrio entre indivíduos egoístas. De fato, ele argumenta que, à medida que as sociedades desenvolvem instituições que encorajam o que Henrich *et al.* chama de integração de mercado interno (discutido acima), suas funções de utilidade e normas sociais tendem a convergir em uma racionalidade econômica egoísta no que diz respeito ao bem-estar. Isso não significa que Binmore é pessimista quanto às perspectivas para o igualitarismo: ele desenvolve um modelo que mostra que sociedades de negociantes de modo geral egoístas podem ser levadas ao longo de caminhos de equilíbrios dinamicamente estáveis em direção a normas de distribuição correspondentes à justiça rawlsiana (RAWLS, 1971). De acordo com Binmore, as principais barreiras para tal evolução são precisamente os tipos de preferências altruístas que conservadores valorizam como forma de desencorajar o exame de mais equilíbrios de negociação igualitária que estão ao alcance ao longo dos caminhos de equilíbrio das sociedades.

Felizmente, a resolução desse debate entre Gintis e Binmore não precisa esperar por descobertas acerca do profundo passado evolutivo humano que talvez nunca tivemos. Os modelos fazem previsões empíricas rivais de alguns fenômenos testáveis. Se Gintis está correto, então há limites, impostos pela descontinuidade na evolução hominídea, na medida em que as pessoas podem aprender a ser egoístas. Essa é a principal importância da controvérsia discutida acima sobre a interpretação de Henrich *et al.* de seus dados de campo. O modelo de seleção de equilíbrio social de Binmore também depende, diferentemente do de Gintis, de disposições difundidas dentre as pessoas para infligir punições de segunda ordem aos membros da sociedade violadores das normas sociais. Usando um modelo de teoria dos jogos, Gintis (2005) mostra que isso é implausível se os custos de punição forem significativos. Contudo, Ross (2008a) argumenta que a pressuposição difundida na literatura de que a punição de

violação de normas deve ser custosa resulta de uma falha em distinguir adequadamente entre modelos da evolução original da sociabilidade, por um lado, e modelos da manutenção e do desenvolvimento de normas e instituições assim que um conjunto inicial delas é estabilizado. Finalmente, Ross aponta também que os objetivos de Binmore são tanto normativos quanto descritivos: ele visa mostrar a igualitários como diagnosticar os erros em racionalizações conservadoras do *status quo* sem exigir revoluções que colocam a estabilidade dos caminhos de equilíbrio (e, portanto, o bem-estar social) em risco. É um princípio sólido na construção de propostas de reforma que elas devem ser "à prova de patifes" (como Hume colocou), isto é, devem ser compatíveis com menos altruísmo do que pode prevalecer nas pessoas. Portanto, apesar do fato de que a maioria dos pesquisadores que trabalham nas fundamentações jogo-teóricas da organização social parecerem, no momento, ficar do lado de Gintis e dos outros membros da equipe de Henrich *et al.*, o modelo alternativo de Binmore possui algumas considerações fortes a seu favor. Aqui, então, está outra questão ao longo da fronteira da aplicação da teoria dos jogos aguardando por resolução nos anos vindouros.

9. Olhando Adiante: Áreas de Inovação Atual

Em 2016, o **Journal of Economic Perspectives** publicou um simpósio em "O que está Acontecendo em Teoria dos Jogos?". Cada um dos participantes notou de forma independente que a teoria dos jogos se tornou tão firmemente entrelaçada com teorias microeconômicas em geral que seria difícil distinguir a questão da investigação na fronteira móvel de toda essa subdisciplina, que, por sua vez, é a maior parte da economia como um todo. Assim, o limite entre a **filosofia** da teoria dos jogos e a filosofia da microeconomia são agora similarmente indistintas. Obviamente, como tem sido salientado, as aplicações da teoria dos jogos se estendem para além do domínio tradicional da economia, em direção a todas as ciências comportamentais e sociais. Mas à medida que os métodos da teoria dos jogos se fundem com os métodos da microeconomia, um comentador poderia igualmente enxergar essas extensões como aplicações exportadas da microeconomia.

Seguindo-se décadas de desenvolvimento discutido (incompletamente) no presente artigo, os últimos anos têm sido relativamente quietos em relações a inovações fundamentais do tipo que convidam contribuições de filósofos. Contudo, algumas partes dos fundamentos originais estão sendo recentemente revisitados.

A introdução de von Neumann e Morgenstern (1944) da teoria dos jogos dividiu a investigação em duas partes. A teoria dos jogos **não-cooperativos** analisa casos construídos sob a pressuposição de que cada jogador maximiza sua própria função de utilidade enquanto trata as respostas estratégicas esperadas de outros jogadores como restrições. Como discutido acima, o jogo específico ao qual von Neumann e Morgenstern aplicaram a sua modelagem era o pôquer, que é um jogo de soma zero. A maior parte do presente artigo tem focado nos muitos desafios teóricos e intuições que surgiram da extensão da teoria dos jogos não-cooperativos para além do domínio de soma zero. Mas isso, na verdade, desenvolve apenas metade da teoria clássica de von Neumann e Morgenstern. A outra metade desenvolveu a teoria dos jogos **cooperativos**, sobre o qual nada tem sido falado aqui. A razão para esse silêncio é que, para a maioria dos teóricos dos jogos, a teoria dos jogos cooperativos é, na melhor das hipóteses, uma distração, e, na pior, uma tecnologia que **confunde** o ponto da teoria dos jogos por ignorar o aspecto de jogos que os torna potencialmente interessantes e iluminadores ao se aplicar a exigência de que o equilíbrio seja selecionado de modo endógeno sob as restrições impostas por Nash (1950a). Afinal de contas, é isso que torna o equilíbrio autoaplicável, da mesma maneira que preços em mercados competitivos o são, e assim os torna estáveis a menos que levem um choque por algo do lado de fora. Nash (1953) argumentou que soluções para jogos cooperativos devem sempre ser verificadas mostrando-se que elas também são soluções para jogos não-cooperativos formalmente equivalentes. A conquista de Nash no artigo foi a identificação analítica da equivalência relevante. Uma maneira de interpretar isso foi demonstrando a redundância da teoria dos jogos cooperativos.

A teoria dos jogos cooperativos começa a partir da pressuposição de que os jogadores já concordaram, por meio de algum processo não especificado, com um vetor de estratégias, e, portanto, com um resultado. Então o analista

implementa a teoria para determinar o conjunto mínimo de condições sob as quais o acordo permanece estável. A ideia é comumente ilustrada pelo exemplo de uma coalizão parlamentar. Suponha que haja um partido dominante que deve ser membro de qualquer coalizão caso for para comandar a maioria dos votos parlamentares em legislação e confiança. Poderia haver então uma gama de possíveis agrupamentos alternativos de outros partidos que poderiam sustentá-la. Para tornar o exemplo mais estruturado e interessante, imagine que alguns partidos não atuarão em uma coalizão que inclua certos outros partidos específicos; então, o problema enfrentado pelos organizadores da coalização não é simplesmente uma questão de somar os votos potenciais. O teórico dos jogos cooperativos identifica o conjunto de possíveis coalizões. Em acréscimo ao partido dominante, pode haver alguns outros partidos que acabam por se tornar necessários em toda coalizão possível. Identificar esses partidos revelaria, nesse exemplo, o **núcleo** do jogo, os elementos compartilhados por todos os equilíbrios. O núcleo é o conceito de solução chave da teoria dos jogos cooperativos, pelo qual Shapley dividiu o prêmio Nobel. ((SHAPLEY, 1953) é um grande artigo.) Nash (1953) definiu o "programa Nash" como aquele que consiste em verificar um equilíbrio cooperativo particular mostrando que jogadores não-cooperativos **poderiam** chegar até ele por meio do processo de negociação sequencial, especificada em (NASH, 1950b), e que **todos** os resultados de tal negociação incluiriam o núcleo.

À luz do exemplo, não é nenhuma surpresa que os cientistas políticos foram os principais usuários da teoria cooperativa durante os anos em que que a teoria dos jogos não-cooperativos não estava ainda totalmente desenvolvida. Ela tem sido aplicada de maneira útil também por economistas do trabalho estudando negociações de acordo entre empresas e sindicatos, e por analistas de negociações comerciais internacionais. Poderíamos ilustrar o valor de tal aplicação por referência ao segundo exemplo. Suponha que, dado o peso dos ***lobbies*** domésticos na África do Sul, o governo sul-africano nunca concordará com nenhum acordo comercial que não lhe permita proteger seu setor de montagem automotiva (Esse tem sido o caso, de fato, até o momento.) Então, a permissão para tal proteção é parte do núcleo de qualquer tratado comercial que outro país ou bloco poderia realizar com a África do Sul. Saber disso pode ajudar

os partidos durante negociações a evitar a retórica ou os comprometimentos com outros *lobbies*, em qualquer um dos países em negociação, que colocaria o núcleo fora de alcance e, assim, garantiria a falha da negociação. Esse exemplo também nos ajuda a ilustrar as limitações da teoria dos jogos cooperativos. A África do Sul terá que equilibrar os interesses de alguns outros *lobbies* para proteger sua indústria automotiva. **Quais** outros serão escolhidos será uma função do jogo de formato extensivo de propostas sequenciais não-cooperativas e contrapropostas, e os negociantes sul-africanos, caso tenham feito sua devida diligência, devem estar atentos a quais caminhos na árvore aniquilam interesses domésticos específicos. Assim, realizar a análise cooperativa não os alivia da necessidade de conduzir também a análise não-cooperativa. Seus consultores de teoria dos jogos poderiam da mesma forma simplesmente codificar os parâmetros não-cooperativos em seus Software de Jogador, que produzirá o núcleo como *output* caso solicitado.

Mas a teoria dos jogos cooperativos não morreu ou se tornou confinada às aplicações da ciência política. Têm surgido uma gama de problemas políticos, que envolvem muitos jogadores com diferentes atributos com funções de utilidade ordinal simétricas; para esses problemas, a modelagem não-cooperativa, mesmo que possível em princípio, é absurdamente desajeitada e computacionalmente exigente, mas a modelagem cooperativa é perfeitamente adequada. É importante que estejamos lidando com funções de utilidade ordinal, pois frequentemente não há preços nos mercados relevantes. O exemplo clássico (GALE; SHAPLEY, 1962) é um mercado de casamentos. Abstraindo-se da escala de dramas e comédias românticas individuais, a sociedade apresenta, por assim dizer, um vasto conjunto de pessoas que querem formar pares mas se preocupam bastante com quem elas acabarão emparelhadas. Suponha que temos um conjunto finito de tais pessoas. Imagine que o casamenteiro, ou o aplicativo da moda, divida primeiramente o conjunto em dois subconjuntos próprios, e anuncia uma regra que todos no subconjunto *A* proporão em casamento alguém no subconjunto *B*. Cada uma daquelas pessoas em *B* que receber uma proposta de casamento sabe que ela é a primeira escolha de alguém em *A*. Ela seleciona sua primeira escolha das propostas que recebeu e retorna as restantes ao jogo. Cada uma daquelas pessoas em *A* cujas propostas iniciais não foram aceitas proporão agora em

casamento alguém que elas não propuseram anteriormente, inclusive, possivelmente, aquelas pessoas que estão com propostas da rodada anterior - Nkosi sabe que Barbara preferiu Amália na rodada 1, mas Nkosi não fazia parte daquele conjunto de escolhas de Bárbara e então poderia tomar o lugar de Amalia na rodada 2.) Provavelmente, existe uma rodada final após a qual nenhuma proposta adicional será feita, e o aplicativo casamenteiro terá encontrado o núcleo do jogo cooperativo, porque nenhuma pessoa *i* no conjunto *B* preferirá formar um par com alguém do conjunto *A* que prefere *i* a quem quer que esteja com um conjunto *A* de propostas dos sonhos. Todas as pessoas do conjunto *B* aceitarão agora a proposta que elas receberem, e, caso os dois conjuntos tenham a mesma cardinalidade e todos prefiram formar um par, então ninguém ficará sozinho.

Isso não é um modelo diretamente aplicável de um mercado de casamentos, portanto ninguém ficará rico vendendo o aplicativo casamenteiro descrito acima. O problema é que não temos garantia de que, no exemplo, Nkosi e Amalia não sejam feitas uma para a outra, mas não possam formar um par porque ambas começaram no subconjunto *A*. Em livros didáticos sobre teoria dos jogos, esse problema é frequentemente contornado assumindo-se que o conjunto *A* contém homens e que o conjunto *B* contém mulheres, e que todos estão tão comprometidos com a heterossexualidade que eles prefeririam formar um par com qualquer pessoa do sexo oposto do que com alguém do próprio sexo. Por outro lado, o modelo fornece algum *insight*, de uma maneira que modelos comumente o fazem, caso não insistamos em aplicá-lo muito literalmente. Depois de esmiuçar isso, vê-se a lógica dos fatos acerca da sociedade que alguém que esteja desenvolvendo um aplicativo casamenteiro real deveria entender melhor: que o aplicativo terá que registrar propostas sob consideração mas que ainda não foram aceitas, deixar no mercado pessoas com propostas sob consideração, e se lembrar de quem rejeitou anteriormente quem (sem criar uma catástrofe emocional generalizada ao postar publicamente essa informação). O aplicativo real não será capaz de encontrar o núcleo do jogo cooperativo de modo confiável a menos que o conjunto de pessoas no mercado seja pequeno, restrito, e auto-organizado em subconjuntos ao menos o suficiente para fornecer informações como “pessoa de tipo *X* procura pessoa de tipo *Y*” para propriedades *X* e *Y* que todo mundo prioriza. (Existem tais propriedades, ao menos como uma

aproximação?) Mas os aplicativos casamenteiros reais parecem funcionar bem o suficiente para transformar a maneira com que a maioria das pessoas jovens encontram companheiros em países com acesso à internet geralmente disponível. As relações entre mercados de casamentos teoricamente idealizados e reais são amplamente revisadas em CHIAPPORI, 2017.

O renascimento da teoria dos jogos cooperativos como um local de interesse renovado tem ocorrido porque têm sido encontrados problemas políticos que satisfazem as pressuposições cruciais do modelo, diferentemente da ilustração diminuta original usando mercados de casamentos totalmente heterossexuais. Os exemplos principais combinam candidatos universitários e universidades, e combinam pessoas que precisam de transplantes de órgão com doadores (*vide* ROTH, 2015). Nesses mercados, não há ambivalência na partição dos conjuntos a serem combinados. As preferências ordinais são as relevantes: as universidades não vendem classificações para o lance mais alto (ao menos não geralmente), e órgãos não estão à venda (ao menos não legalmente). Os modelos são realmente aplicados, e eles possuem demonstrativamente eficiência aprimorada e salvam vidas.

É comum na ciência que modelos que são encaixes praticamente desajeitados para seus problemas originais acabem por fornecer soluções altamente eficientes para novos problemas gerados pela mudança tecnológica. A internet tem criado um ambiente para aplicações de algoritmos de combinação - viajantes e locatários de *flats*, pessoas querendo jantar e restaurantes, estudantes e tutores, e (lamentavelmente) pessoas socialmente alienadas e fornecedores de propaganda e fanatismo - que poderiam ser projetados por um teórico a qualquer momento desde as inovações originais de Shapley, mas que teriam sido praticamente impossíveis de implementar anteriormente. Essas aplicações da teoria dos jogos cooperativos têm sido aplicadas em conjunto com a teoria dos jogos não cooperativos de leilões (KLEMPERER, 2004) para impulsionar projetos de mercado de bens e serviços de modo tão eficiente a ponto de aniquilar o outrora poderoso shopping center até mesmo nos subúrbios dos EUA. Por que hotéis são muito mais lucrativos e facilmente disponíveis do que havia sido o caso em todas menos nas maiores cidades antes de 2007? A resposta é que os algoritmos de preços dinâmicos (GERSHKOV; MOLDOVANU, 2014) misturaram a

teoria do pareamento e a teoria dos leilões para permitir que hotéis, combinados com agregadores de serviços de viagens online, encontrem clientes dispostos a pagar taxas *premium* por seus locais e datas ideais, e então preencher os quartos remanescentes com caçadores de ofertas cujas preferências são mais flexíveis. Companhias aéreas operam tecnologia similar. Assim, a teoria dos jogos continua a ser uma das invenções do século XX que está conduzindo revoluções no século XXI, e **Samuelson (2016)** prevê uma onda de interesses renovados na matemática mais profunda dos jogos cooperativos e suas relações com jogos não-cooperativos.

Uma gama de aplicações adicionais tanto da teoria dos jogos clássica quanto da evolutiva tem sido desenvolvida, mas esperamos ter fornecido agora o suficiente para convencer o leitor da tremenda, e em constante expansão, utilidade dessa ferramenta analítica. Se o seu apetite para investigar a teoria dos jogos mais a fundo foi despertado por este artigo, saiba que você tem agora uma compreensão suficiente dos fundamentos para ser capaz de trabalhar com a grande literatura, da qual alguns destaques serão listados abaixo.

Bibliografia

Anotações

Na seção seguinte, os livros e artigos que ninguém seriamente interessado na teoria dos jogos pode se dar ao luxo de perder estão marcados com (**).

O livro didático mais acessível que abrange todos os principais ramos da teoria dos jogos é (DIXIT; SKEATH; REILEY, 2014). Um estudante novo no campo deve trabalhar com esse material antes de passar para qualquer outra coisa.

A teoria dos jogos tem inúmeras aplicações, das quais esse artigo sugeriu apenas algumas. Os leitores que procuram por mais, mas não desejam mergulhar na matemática, podem encontrar várias boas fontes. (DIXIT; NALEBUFF, 1991 e 2008) são especialmente fortes em exemplos políticos e sociais. (MCMILAN, 1991) enfatiza as aplicações aos negócios.

O grande avanço histórico que lançou oficialmente a teoria dos jogos é (VON NEUMANN; MORGENSTERN, 1944), e aqueles com interesse acadêmico

em teoria dos jogos devem ler com os artigos clássicos (NASH, 1950a, 1950b, 1951). Uma coleção muito útil de artigos chave fundamentais, todos clássicos, é (KUHN, 1997). Para um tratamento matemático contemporâneo extraordinariamente sofisticado filosoficamente, (BINMORE, 2005c) (**) é um clássico. A segunda metade de (KREPS, 1990) (**) é o melhor ponto de início disponível para um ***tour*** sobre as preocupações filosóficas em torno da seleção de equilíbrio para normativistas. (KOON, 1992) leva esses problemas mais adiante. (FUDENBERG; TIROLE, 1999) é o texto matemático mais exaustivo e completo disponível. (GINTIS, 2009b)(**) fornece um texto repleto de excelentes exercícios, e também é único por tratar a teoria evolutiva dos jogos como fornecendo a base fundamental para a teoria dos jogos em geral. Desenvolvimentos recentes na teoria fundamental estão bem representados em (BINMORE; KIRMAN; TANI, 1993). Qualquer pessoa que queira aplicar a teoria dos jogos às escolhas humanas reais, geralmente relacionadas de modo aleatório ao invés de determinístico com os axiomas de otimização, precisa entender a teoria da resposta discreta (TRD) como um conceito de solução. O desenvolvimento original disso é encontrado em (MCKELVEY; PALFREY, 1995) e (MCKELVEY; PALFREY, 1998). (GOEREE; HOLT; PALFREY,2016) fornece uma revisão abrangente e atualizada da TRD e de suas principais aplicações.

Os fundamentos filosóficos dos conceitos básicos da teoria dos jogos como os economistas os entendem são apresentadas em (LACASSE; ROSS,1994). (ROSS; LACASSE, 1995) esboça as relações entre jogos e as pressuposições axiomáticas da microeconomia e da macroeconomia. Enigmas filosóficos nesse nível fundamental são criticamente discutidos em (BICCHIERI, 1993). (LEWIS, 1969) faz uma aplicação mais ampla na filosofia dos conceitos de equilíbrio jogo-teóricos, embora faça algumas pressuposições fundamentais que os economistas geralmente não partilham. Seu programa é levado muito mais adiante, e sem as pressuposições contestadas, em (SKYRMS, 1996(**) e 2004). (*vide* também NOZICK, 1998.) GAUTHIER, 1986 apresenta uma literatura não levantada nesse artigo, em que é investigada a possibilidade de fundamentações jogo-teóricas para a ética contratualista. Esse trabalho é criticamente discutido em (VALLENTYNE, 1991), e estendido para um cenário dinâmico em (DANIELSON, 1992). Contudo, (BINMORE, 1994, 1998) (**) critica fortemente

esse projeto como inconsistente com a psicologia natural. Filósofos também acharão interessante (HOLLIS, 1998).

Em uma classe separada, em decorrência de *insight*, originalidade, legibilidade, e importância transdisciplinar, estão os trabalhos do vencedor do prêmio Nobel, Thomas Schelling. Ele é a fonte da enorme literatura que aplica a teoria dos jogos às questões sociais e políticas de relevância imediata, e mostra com que leveza é possível usar a matemática se a lógica for suficientemente segura. Há quatro volumes, todos essenciais: (SCHELLING, 1960) (**), (SCHELLING, 1978/2006) (**),(SCHELLING, 1984) (**), (SCHELLING, 2006) (**).

(HARDIN, 1995) é um dos muitos exemplos da aplicação da teoria dos jogos a problemas em teoria política aplicada. (BAIRD; GERTNER; PICKER, 1994) revê os usos da teoria dos jogos em teoria legal e jurisprudência. (MUELLER, 1997) pesquisa a aplicaçõ em escolhas públicas. (GHEMAWAT, 1997) fornece estudos de caso destinados a servir como um modelo metodológico para aplicações práticas da teoria dos jogos a problemas de estratégia de negócios. (POUNDSTONE, 1992) fornece uma história vívida do Dilema do Prisioneiro e seu uso por estrategistas da Guerra Fria. (AMADAE, 2016) conta a mesma história, baseada em investigações acadêmicas originais, com menos complacência quanto às suas implicações. A auto-biografia de Ellsberg (ELLSBERG, 2017) confirma enormemente a perspectiva de Amadae. (DURLAUF; YOUNG, 2001) é uma útil coleção de aplicações à estruturas sociais e mudança social.

A teoria evolutiva dos jogos deve sua gênesis explícita a Manynard Smith (MAYNARD SMITH, 1982) (**). Para um texto que integra a teoria dos jogos diretamente com a biologia, *vide* HOFBAUER; SIGMUND,1998 (**). SIGMUND, 1993 apresenta esse material em um formato menos técnico e mais acessível. Algumas aplicações excitantes da teoria evolutiva dos jogos a uma gama de questões filosóficas, em que esse artigo se baseou fortemente, estão em (SKYRMS, 1996) (**). Essas questões e outras são discutidas criticamente de vários ângulos em (DANIELSON, 1998). Fundamentos matemáticos para jogos evolutivos são apresentados em (WEIBULL, 1995), e levados mais adiante em (SAMUELSON, 1997). Como notado acima, (GINTIS, 2009b) (**) é um livro didático introdutório que toma a modelagem evolutiva como fundamental para

toda a teoria dos jogos. (H. P. YOUNG, 1998) fornece modelos sofisticados da dinâmica evolutiva das normas culturais através de interações jogo-teóricas de agentes com capacidades cognitivas limitadas mas disposições para imitar uns aos outros. (FUDENBERG; LEVINE, 1998) fornece os fundamentos técnicos para modelagens desse tipo.

Muitos filósofos também ficarão interessados em (BINMORE, 1994, 1998, 2005a) (**), que mostram que a aplicação da análise jogo-teórica pode subscrever uma concepção rawlsiana de justiça que não requer recurso a pressuposições kantianas sobre o que agentes racionais desejariam por trás de um véu da ignorância no que diz respeito às suas identidades e papeis sociais . (Em acréscimo, Binmore oferece excursões por uma gama de outras questões tanto centrais quanto periféricas para ambos os fundamentos e as fronteiras da teoria dos jogos; esses livros são particularmente ricos em problemas que interessam aos filósofos.) Quase todo mundo ficará interessado em (FRANK,1988) (**), em que a teoria evolutiva dos jogos é utilizada para iluminar aspectos básicos da natureza e da emoção humanas; embora os leitores desse material possam encontrar críticas ao modelo de Frank em (ROSS; DUMOUCHEL, 2004).

As aplicações comportamentais e experimentais da teoria dos jogos são pesquisadas em (KAGEL; ROTH, 1995). (CAMERER, 2003) (**) é um estudo abrangente e mais recente dessa literatura, e não pode ser perdido por ninguém com interesse por essas questões. Uma pesquisa mais curta que enfatiza críticas filosóficas e metodológicas é (SAMUELSON, 2005). Os fundamentos filosóficos são também cuidadosamente examinados em (GUALA,2005).

Dois volumes dos principais teóricos que oferecem visões abrangentes dos fundamentos filosóficos da teoria dos jogos foram publicados em 2009. Eles são (BINMORE, 2009) (**) e (GINTIS, 2009a)(**). Ambos são indispensáveis para filósofos que visam participar em discussões críticas de questões fundacionais.

Um volume com dezenove principais teóricos dos jogos, expressando suas visões em tópicos motivacionais e fundacionais, é (HENDRICKS; HANSEN, 2007).

Um maravilhoso desenvolvimento recente nos fundamentos da teoria dos jogos é a invenção da teoria dos jogos condicionais por Stirling. (STIRLING, 2012)

se restringe à matemática, com algumas possibilidades principais de aplicação, juntamente com extensões técnicas que fornecem pontes com a economia, sendo encontrada no seguinte, (STIRLING, 2016). A importância filosófica desse trabalho é melhor compreendida à luz de considerações introduzidas em (BACHARACH, 2006).

A dinâmica jogo-teórica subpessoal recebe uma profunda porém acessível reflexão em (AINSLIE, 2001). Textos seminais em neuroeconomia, com uso extensivo e implicações para a teoria comportamental dos jogos, são (MONTAGUE; BERNS, 2002), (GLIMCHER, 2003) (**), e (CAMERER; LOEWENSTEIN; PRELEC, 2005). (ROSS, 2005a) estuda os fundamentos jogo-teóricos da microeconomia em geral, mas especialmente a economia comportamental e a neuroeconomia, da perspectiva da ciência cognitiva e em próximo alinhamento com Ainslie.

A teoria dos jogos cooperativos é consolidada em (CHAKRAVARTY; MITRA; SARKAR, 2015). Uma aplicação acessível e não técnica de aplicações do pareamento pelo economista cujos trabalhos nisso recebeu um prêmio Nobel, é (ROTH, 2015).

Bibliografia

AINSLIE, G. **Picoeconomics.** Cambridge: Cambridge University Press, 1992.

AINSLIE, G. **Breakdown of Will.** Cambridge: Cambridge University Press, 2001.

AMADAE, S. **Prisoners of Reason.** Cambridge: Cambridge University Press, 2016.

ANDERSEN, S *et al.* **Eliciting risk and time preferences**. Econometrica, 2008, 76: p. 583–618.

ANDERSEN, S. *et al.* **Dual criteria decisions**. Journal of Economic Psychology„ v. 41, p. 101-113, abr. 2014.

BACHARACH, M. **Beyond Individual Choice:** Teams and Frames in Game Theory. Princeton: Princeton University Press, 2006.

BAIRD, D.; GERTNER, R.; PICKER, R. **Game Theory and the Law.** Cambridge, MA: Harvard University Press, 1994.

BELL, W. **Searching Behaviour.** London: Chapman and Hall, 1991.

BICCHIERI, C. **Rationality and Coordination.** Cambridge: Cambridge University Press, 1993.

BICCHIERI, C. **The Grammar of Society.** Cambridge: Cambridge University Press, 2006.

BICKHARD, M. **Social ontology as convention**. Topoi, 2008, 27: 139–149.

BINMORE, K. **Modeling Rational Players I.** Economics and Philosophy, 1987, 3: 179–214

BINMORE, K. **Game Theory and the Social Contract (v. 1):** Playing Fair. Cambridge, MA: MIT Press, 1994.

BINMORE, K. **Game Theory and the Social Contract (v. 2):** Just Playing. Cambridge, MA: MIT Press, 1998.

BINMORE, K. **Natural Justice**. Oxford: Oxford University Press, 2005a.

BINMORE, K. **Economic Man—or Straw Man?** Behavioral and Brain Sciences, 2005b, 28: 817–818.

BINMORE, K. **Playing For Real.** Oxford: Oxford University Press, 2005c.

BINMORE, K. **Does Game Theory Work? The Bargaining Challenge.** Cambridge, MA: MIT Press, 2007.

BINMORE, K. **Do conventions need to be common knowledge?** Topoi, 2008, 27: 17–27.

BINMORE, K. **Rational Decisions.** Princeton: Princeton University Press, 2009.

BINMORE, K.; KIRMAN, A.; TANI, P. (eds.). **Frontiers of Game Theory.** Cambridge, MA: MIT Press, 1993.

BINMORE, K.; KLEMPERER, P. **The Biggest Auction Ever:** The Sale of British 3G Telcom Licenses**.** Economic Journal, 2002, 112: C74–C96.

BINMORE, K. **Individual Decision Making.** *In:* J. Kagel and A. Roth (eds). Handbook of Experimental Economics. Princeton: Princeton University Press, 1995, p. 587–703.

BINMORE, K. **Behavioral Game Theory:** Experiments in Strategic Interaction. Princeton: Princeton University Press, 2003.

CAMERER, C.; LOEWENSTEIN, G.; PRELEC, D. **Neuroeconomics:** How Neuroscience Can Inform Economics. Journal of Economic Literature, 2005, 40: 9–64.

CHAKRAVARTY, S.; MITRA, M.; SARKAR, P. **A Course on Cooperative Game**

Theory. Cambridge: Cambridge University Press, 2015.

CHEW, S.; e MACCRIMMON, K. **Alpha-nu Choice Theory:** A Generalization of Expected Utility Theory. Working Paper No. 686, University of Columbia Faculty of Commerce and Business Administration, 1979.

CHIAPPORI, P.-A. **Matching With Transfers:** The Economics of Love and Marriage. Princeton: Princeton University Press, 2017.

CLARK, A. **Being There.** Cambridge, MA: MIT Press, 1997.

DANIELSON, P. **Artificial Morality.** London: Routledge,1992.

DANIELSON, P. (ed.) **Modelling Rationality.** Morality and Evolution. Oxford: Oxford University Press, 1998.

DENNETT, D. **Darwin's Dangerous Idea.** Nova York: Simon and Schuster, 1995.

DIXIT, A.; NALEBUFF, B. **Thinking Strategically.** Nova York: Norton, 1991.

DIXIT, A.; NALEBUFF, B. **The Art of Strategy.** Nova York: Norton, 2008.

DIXIT, A.; SKEATH, S.; REILEY, D. **Games of Strategy.** Nova York: W. W. Norton and Company, 2014.

DUGATKIN, L.; REEVE, H. (eds.) **Game Theory and Animal Behavior.** Oxford: Oxford University Press, 1998.

DUKAS, R. (ed.) **Cognitive Ecology.** Chicago: University of Chicago Press, 1998.

DURLAUF, S.; e YOUNG, H.P. (eds.) **Social Dynamics.** Cambridge, MA: MIT Press, 2001.

ELLSBERG, D. **The Doomsday Machine.** Nova York: Bloomsbury, 2017.

ERICKSON, P. **The World the Game Theorists Made.** Chicago: University of Chicago Press, 2015.

FRANK, R. **Passions Within Reason.** Nova York: Norton, 1988.

FUDENBERG, D.; e LEVINE, D. **The Theory of Learning in Games.** Cambridge, MA: MIT Press, 1998.

FUDENBERG, D.; e LEVINE, D. **Whither Game Theory? Towards a Theory of Learning in Games.** Journal of Economic Perspectives, 2016, 30(4): 151–170.

FUDENBERG, D.; TIROLE, J. **Game Theory.** Cambridge, MA: MIT Press, 1991.

GALE, D.; SHAPLEY, L. **College Admissions and the Stability of Marriage.** American Mathematical Monthly, 1962, 69 :9–15.

GAUTHIER, D. **Morals By Agreement.** Oxford: Oxford University Press, 1986.

GERSHKOV, A.; MOLDOVANU, B. **Dynamic Allocation and Pricing:** A Mechanism Design Approach. Cambridge, MA: MIT Press, 2014.

GHEMAWAT, P. **Games Businesses Play.** Cambridge, MA: MIT Press, 1997.

GILBERT, M. **On Social Facts.** Princeton: Princeton University Press, 1989.

GINTIS, G. **Towards the Unity of the Human Behavioral Sciences.** Philosophy, Politics and Economics, 2004, 31: 37–57.

GINTIS, G. **Behavioral Ethics Meets Natural Justice.** Politics, Philosophy and Economics, 2005, 5: 5–32.

GINTIS, G. **The Bounds of Reason.** Princeton: Princeton University Press, 2009a.

GINTIS, G. **Game Theory Evolving.** 2ª ed. Princeton: Princeton University Press, 2009b.

GLIMCHER, P. **Decisions, Uncertainty and the Brain.** Cambridge, MA: MIT Press, 2009b.

GLIMCHER, P.; KABLE, J.; LOUIE, K. **Neuroeconomic studies of impulsivity:** Now or just as soon as possible? American Economic Review (Papers and Proceedings), 2007, 97: 142–147.

GOEREE, J.; HOLT, C.; PALFREY, T. **Quantal Response Equilibrium.** Princeton: Princeton University Press, 2016.

GUALA, F. **The Methodology of Experimental Economics.** Cambridge: Cambridge University Press, 2005.

GUALA, F. **Understanding Institutions.** Princeton: Princeton University Press, 2016.

HAMMERSTEIN, P. **Why is reciprocity so rare in social animals? A protestant appeal.** *In:* P. Hammerstein (ed.), Genetic and Cultural Evolution of Cooperation. Cambridge, MA: MIT Press, 2003, p. 83–93.

HARDIN, R. **One For All.** Princeton: Princeton University Press, 1995.

HARRISON, G.W. **Neuroeconomics:** A critical reconsideration. Economics and Philosophy, 2008, 24: 303–344.

HARRISON, G.W.; RUTSTROM, E. **Risk aversion in the laboratory.** *In:* Risk Aversion in Experiments, J. Cox e G. Harrison (eds.). Bingley, UK: Emerald, 2008, p. 41–196.

HARRISON, G.W.; e ROSS, D. **The methodologies of neuroeconomics.**

Journal of Economic Methodology, 2010, 17: 185–196.

HARSANYI, J. **Games With Incomplete Information Played by 'Bayesian' Players.** Parts I-III. Management Science, 1967, 14: 159–182.

HENRICH, J. *et al.* (eds.) **Foundations of Human Sociality:** Economic Experiments and Ethnographic Evidence From 15 Small-Scale Societies. Oxford: Oxford University Press, 2004.

HENRICH, J. *et al.* **'Economic Man' in Cross-Cultural Perspective.** Behavioral and Brain Sciences, 2005, 28: 795–815.

HENDRICKS, V.; e HANSEN, P. (eds.) **Game Theory:** 5 Questions. Copenhagen: Automatic Press, 2007.

HOFBAUER, J.; SIGMUND, K. **Evolutionary Games and Population Dynamics.** Cambridge: Cambridge University Press, 1998.

HOFMEYR„ A.; ROSS, D. **Team Agency and Conditional Games.** *In:* M. Nagatsu (ed.). Philosophy and Social Science: An Interdisciplinary Dialogue. London: Bloomsbury, 2019.

HOLLIS, M. **Trust Within Reason.** Cambridge: Cambridge University Press, 1998.

HOLLIS, M.; SUGDEN, R. **Rationality in Action.** Mind, 1993, 102: 1–35.

HURWICZ, L.; REITER, S. **Designing Economic Mechanisms.** Cambridge: Cambridge University Press, 2006.

KAGEL, J.; ROTH, A. (eds.) **Handbook of Experimental Economics.** Princeton: Princeton University Press, 1995.

KEENEY, R.; e RAIFFA, H. **Decisions With Multiple Objectives.** Nova York: Wiley, 1976.

KING-CASAS, B. *et al.* **Getting to Know You:** Reputation and Trust in a Two-Person Economic Exchange. Science, 2005, 308: 78–83.

KLEMPERER, P. **Auctions:** Theory and Practice. Princeton: Princeton University Press, 2004.

KOONS, R. **Paradoxes of Belief and Strategic Rationality.** Cambridge: Cambridge University Press, 1992.

KREBS, J.; e DAVIES, N. **Behavioral Ecology:** An Evolutionary Approach. 2ª ed. Sunderland: Sinauer, 1984.

KREPS, D. **A Course in Microeconomic Theory.** Princeton: Princeton

University Press, 1990.

KUHN, H. (ed.) **Classics in Game Theory.** Princeton: Princeton University Press, 1997.

LACASSE, C.; ROSS, D. **'The Microeconomic Interpretation of Games'.** PSA 1994, v. 1, D. Hull, S. Forbes and R. Burien (eds.). East Lansing. MI: Philosophy of Science Association, 1994, p. 479–387.

LEDYARD, J. **Public Goods:** A Survey of Experimental Research. *In:* J. Kagel and A. Roth (eds.). Handbook of Experimental Economics. Princeton: Princeton University Press, 1995.

LEWIS, D. **Convention.** Cambridge, MA: Harvard University Press, 1969.

MAYNARD SMITH, J. **Evolution and the Theory of Games.** Cambridge: Cambridge University Press, 1982.

MCCLURE, S. *et al.* **Separate neural systems value immediate and delayed monetary rewards.** Science, 2004, 306: 503–507.

MCKELVEY, R.; e PALFREY, T. **Quantal response equilibria for normal form games.** Games and Economic Behavior, 1995, 10: 6–38.

MCKELVEY, R.; e PALFREY, T. **Quantal response equilibria for extensive form games.** Experimental Economics, 1998, 1: 9–41.

MCMILLAN, J. **Games, Strategies and Managers.** Oxford: Oxford University Press, 1991.

MILLIKAN, R. **Language, Thought and Other Biological Categories.** Cambridge, MA: MIT Press, 1984.

MONTAGUE, P. R.; e BERNS, G. **Neural Economics and the Biological Substrates of Valuation.** Neuron, 2002, 36: 265–284.

MUELLER, D. **Perspectives on Public Choice.** Cambridge: Cambridge University Press, 1997.

NASH, J. **'Equilibrium Points in n-Person Games.'** Proceedings of the National Academy of Science, 1950a, 36: 48–49.

NASH, J. **'The Bargaining Problem.'** Econometrica, 1950b, 18: 155–162.

NASH, J. **'Non-cooperative Games.'** Annals of Mathematics Journal, 1951, 54: 286–295

NASH, J. **Two-Person Cooperative Games.** Econometrica, 1953, 21: 128–140.

NOE, R.; VAN HOOF, J.; e HAMMERSTEIN, P. (eds.) **Economics in Nature.**

Cambridge: Cambridge University Press, 2001.

NOZICK, R. **Socratic Puzzles.** Cambridge, MA: Harvard University Press,1998.

ORMEROD, P. **The Death of Economics.** Nova York: Wiley, 1994.

PETTIT, P.; e SUGDEN, R. **The Backward Induction Paradox.** Journal of Philosophy, 1989, 86: 169–182.

PLATT, M.; e GLIMCHER, P. **Neural Correlates of Decision Variables in Parietal Cortex.** Nature, 1999, 400: 233–238.

PLOTT, C.; e SMITH, V. **An Experimental Examination of Two Exchange Institutions.** Review of Economic Studies, 1978,45: 133–153.

POUNDSTONE, W. **Prisoner's Dilemma.** Nova York: Doubleday, 1992.

PRELEC, D. **The Probability Weighting Function.** Econometrica, 1998, 66: 497–527.

QUIGGIN, J. **A Theory of Anticipated Utility.** Journal of Economic Behavior and Organization, 1982, 3: 323–343.

RAWLS, J. **A Theory of Justice.** Cambridge, MA: Harvard University Press, 1971.

ROBBINS, L. **An Essay on the Nature and Significance of Economic Science.** London: Macmillan, 1931.

ROSS, D. **Economic Theory and Cognitive Science:** Microexplanation. Cambridge, MA: MIT Press, 2005a.

ROSS, D. **Evolutionary Game Theory and the Normative Theory of Institutional Design:** Binmore and Behavioral Economics. Politics, Philosophy and Economics, v. 5, Issue 1, 2005b.

ROSS, D. **Classical game theory, socialization and the rationalization of conventions.** Topoi, 2008a, 27: 57–72.

ROSS, D. **Two styles of neuroeconomics.** Economics and Philosophy, 2008b, 24: 473–483.

ROSS, D. **Philosophy of Economics.** Houndmills, Basingstoke: Palgrave Macmillan, 2014.

ROSS, D.; DUMOUCHEL, P. **Emotions as Strategic Signals.** Rationality and Society, 2004, 16: 251–286.

ROSS, D.; LACASSE, C. **'Towards a New Philosophy of Positive Economics'.** Dialogue, 1995, 34: 467–493.

ROTH, A. **Who Gets What and Why?.** Nova York: Houghton Mifflin Harcourt, 2015.

SALLY, J. **Conversation and Cooperation in Social Dilemmas:** A Meta-analysis of Experiments From 1958 to 1992. Rationality and Society, 1995, 7: 58–92.

SAMUELSON, L. **Evolutionary Games and Equilibrium Selection.** Cambridge, MA: MIT Press, 1997.

SAMUELSON, L. **Economic Theory and Experimental Economics.** Journal of Economic Literature, 2005, 43: 65–107.

SAMUELSON, L. **Game Theory in Economics and Beyond.** Journal of Economic Perspectives, 2016, 30(4): 107–130.

SAMUELSON, P. **'A Note on the Pure Theory of Consumers' Behaviour.'** Economica, 1938, 5: 61–71.

SAVAGE, L. **The Foundations of Statistics.** Nova York: Wiley, 1954.

SCHELLING, T. **Strategy of Conflict.** Cambridge, MA: Harvard University Press, (1960>. Schelling, T (1960).

SCHELLING, T. **Micromotives and Macrobehavior.** Nova York: Norton. 2ª ed. 2006, 1978.

SCHELLING, T. **The Intimate Contest for Self-Command.** Public Interest, 1980, 60: 94–118.

SCHELLING, T. **Choice and Consequence.** Cambridge, MA: Harvard University Press, 1984.

SCHELLING, T. **Strategies of Commitment.** Cambridge, MA: Harvard University Press, 2006.

SELTEN, R. **'Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games.'** International Journal of Game Theory, 1975, 4: 22–55.

SIGMUND, K. **Games of Life.** Oxford: Oxford University Press, 1993.

SHAPLEY, L. **A Value of n-Person Games.** *In:* H, Kuhn and A. Tucker (eds.). Contributions to the Theory of Games II. Princeton: Princeton University Press, 1953, p. 307–317.

SKYRMS, B. **Evolution of the Social Contract.** Cambridge: Cambridge University Press, 1996.

SKYRMS, B. **The Stag Hunt and the Evolution of Social Structure.** Cambridge:

Cambridge University Press, 2004.

SMITH, V. **An Experimental Study of Competitive Market Behavior.** Journal of Political Economy, 1962, 70: 111–137.

SMITH, V. **Effect of Market Organization on Competitive Equilibrium.** Quarterly Journal of Economics, 1964, 78: 181–201.

SMITH, V. **Experimental Auction Markets and the Walrasian Hypothesis.** Journal of Political Economy, 1965, 73: 387–393.

SMITH, V. **Bidding and Auctioning Institutions:** Experimental Results. *In:* Y. Amihud (ed.). Bidding and Auctioning for Procurement and Allocation. Nova York: Nova York University Press, 1976, p. 43–64.

SMITH, V. **Microeconomic Systems as an Experimental Science.** American Economic Review, 1982, 72: 923–955.

SMITH, V. **Rationality in Economics.** Cambridge: Cambridge University Press. Sober, E., and Wilson, D.S. (1998). Unto Others, Cambridge, MA: Harvard University Press, 2008.

STERELNY, K. **Thought in a Hostile World.** Oxford: Blackwell, 2003.

STIRLING, W. **Theory of Conditional Games.** Cambridge: Cambridge University Press, 2012.

STIRLING, W. **Theory of Social Choice on Networks.** Cambridge: Cambridge University Press, 2016.

SRATMANN, T. **Logrolling.** *In:* D. Mueller (ed.). Perspectives on Public Choice. Cambridge: Cambridge University Press, 1997, p. 322–341.

STROTZ, R. **Myopia and Inconsistency in Dynamic Utility Maximization.** The Review of Economic Studies, 1956, 23: 165–180.

SUGDEN, R. **Thinking as a Team:** Towards an Explanation of Nonselfish Behavior. Social Philosophy and Policy, 1993, 10: 69–89.

SUGDEN, R. **Team Preferences.** Economics and Philosophy, 2000, 16: 175–204.

SUGDEN, R. **The Logic of Team Reasoning.** Philosophical Explorations, 2003, 6: 165–181.

SUGDEN, R. **The Community of Advantage.** Oxford: Oxford University Press, 2018.

THURSTONE, L. **The Indifference Function.** Journal of Social Psychology,

1931, 2: 139–167.

TOMASELLO, M. *et al.* **Understanding and Sharing Intentions:** The Origins of Cultural Cognition. Behavioral and Brain Sciences, 2004, 28: 675–691.

TVERSKY, A.; KAHNEMAN, D. **Advances in Prospect Theory:** Cumulative Representation of Uncertainty. Journal of Risk and Uncertainty, 1992, 5: 297–323.

VALLENTYNE, P. (ed.). **Contractarianism and Rational Choice.** Cambridge: Cambridge University Press, 1991.

VON NEUMANN, J.; MORGENSTERN, O. **The Theory of Games and Economic Behavior.** Princeton: Princeton University Press, 1944.

VON NEUMANN, J.; e MORGENSTERN, O. **The Theory of Games and Economic Behavior.** 2ª ed. Princeton: Princeton University Press, 1947.

WEIBULL, J. **Evolutionary Game Theory.** Cambridge, MA: MIT Press, 1995.

WILCOX, N. **Stochastic Models for Binary Discrete Choice Under Risk:** A Critical Primer and Econometric Comparison. *In:* J. Cox and G. Harrison (eds.). Risk Aversion and Experiments. Bingley, UK: Emerald, 2008.

YAARI, M. **The Dual Theory of Choice Under Risk.** Econometrica, 1987, 55: 95–115.

YOUNG, H.P. **Individual Strategy and Social Structure.** Princeton: Princeton University Press, 1998.

Sobre os editores, tradutores e revisores

Arthur de Castro Machado (tradutor): Mestrando em Filosofia na UFMG na linha de pesquisa em Lógica, Ciência, Mente e Linguagem. Graduado em Filosofia (licenciatura) pela UFOP.

Débora de Oliveira Silva (tradutora): Mestranda em Filosofia na Unicamp na linha de pesquisa em Teoria do Conhecimento e Filosofia da Ciência e da Linguagem. Graduada em Filosofia pela UFOP.

Guilherme A. Cardoso (editor, tradutor e revisor): doutor e mestre em filosofia pela UFMG com período sanduíche na Brown University sob supervisão de Richard Kimberly Heck. Pós-doutorado no CLE-Unicamp. Atualmente, é professor de filosofia da UFOP.

Hulian Ferreira de Araujo (tradutor): Graduando em Filosofia pela UFMG. Possui interesse nas áreas de Lógica, Filosofia da Lógica, Filosofia da Linguagem e Metafísica.

Sérgio R. N. Miranda (editor, tradutor e revisor): mestre em filosofia pela UFMG e doutor em filosofia pela Universität Bielefeld. Atualmente, é professor de filosofia da UFOP.

Editora
UFPel

DISSERTATIO
FILOSOFIA