UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE FILOSOFIA A CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

SILVIO KAVETSKI

NATURALISMO ÉTICO E O ARGUMENTO DA TERRA GÊMEA MORAL

FLORIANÓPOLIS
2022

SILVIO KAVETSKI

NATURALISMO ÉTICO E O ARGUMENTO DA TERRA GÊMEA MORAL

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Filosofia do Centro de Filosofia e Ciências Humanas da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Doutor em Filosofia.
Orientador: Prof. Dr. Darlei Dall'Agnol

FLORIANÓPOLIS

2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Kavetski, Silvio
Naturalismo ético e o argumento da terra gêmea moral / Silvio Kavetski ; orientador, Darlei Dall'Agnol, 2022.
196 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Filosofia e Ciências Humanas, Programa de Pós-Graduação em Filosofia, Florianópolis, 2022.

Inclui referências.

1. Filosofia. 2. Metaética. 3. Realismo Moral. 4. Naturalismo Moral. 5. Argumento da Terra Gêmea Moral. I. Dall'Agnol, Darlei . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Filosofia. III. Título.

SILVIO KAVETSKI

NATURALISMO ÉTICO E O ARGUMENTO DA TERRA GÊMEA MORAL

O presente trabalho em nível de Doutorado foi avaliado e aprovado em 29 de julho de 2022 pela banca examinadora composta pelos seguintes membros:

Prof. Dr. Leonardo de Mello Ribeiro
Universidade Federal de Minas Gerais

Profª. Drª. Janyne Sattler
Universidade Federal de Santa Catarina

Prof. Dr. Jerzy Andre Brzozowski
Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para a obtenção do título de Doutor em Filosofia.

Documento assinado digitalmente
Vilmar Debona
Data: 09/08/2022 14:19:14-0300
CPF: 341.282.968-48
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Vilmar Debona
Coordenador do Programa de Pós-Graduação em Filosofia

Documento assinado digitalmente
DARLEI DALL AGNOL
Data: 09/08/2022 13:42:16-0300
CPF: 414.182.200-00
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Darlei Dall'Agnol (orientador)

FLORIANÓPOLIS
2022

Dedico este trabalho à Amélia Kavetski,
minha mãe, cuja força e positividade inspiram.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha família por todo incentivo para que eu pudesse concluir e Doutorado em Filosofia. Meus pais, Pedro e Amélia Kavetski, me apoiaram incondicionalmente sempre que precisei de algo. Os anos de 2020 e 2021 foram os mais difíceis devido às inúmeras consequências desagradáveis da Pandemia de Covid-19 e eles sempre respeitaram o meu espaço quando precisei me "isolar" para ler e escrever, suportaram meus piores dias de estresse e me deram todo o apoio necessário para continuar. Agradeço, também, a meu irmão, Porfírio Kavetski, que sempre esteve de prontidão para me ajudar com as inúmeras viagens e mudanças, principalmente nos últimos três anos. Sua ajuda para que eu pudesse retornar à Florianópolis nos meses finais para concluir essa tese foi fundamental.

Meus sinceros agradecimentos ao professor Darlei Dall'Agnol por ter me orientado durante o Mestrado e o Doutorado. Aprendi muito durante esses anos, especialmente através de seu exemplo como acadêmico competente, rigoroso com as ideias e sofisticações filosóficas e sempre motivado em desenvolver teses e argumentos originais. Agradeço pela sua amizade, atenção, encorajamento e, sobretudo, pela paciência em respeitar meu próprio ritmo de pesquisa.

À Janyne Sattler, Jerzy Andre Brzozowski e Leonardo de Mello Ribeiro por terem participado da Banca de Defesa. Agradeço por todas as sugestões para tornar a versão final desse trabalho melhor. Obviamente que quaisquer erros restantes são de minha inteira responsabilidade.

À todas as pessoas que participaram do Seminário de Aprofundamento em Pesquisas Éticas (SAPE) durante a minha estada no programa de pós-graduação. Ler e discutir obras de diferentes assuntos filosóficos, apresentar partes da própria pesquisa em andamento, discutir as pesquisas dos colegas e fazer novas amizades constituíram uma das partes mais importantes da minha formação durante esses anos.

Aos amigos com quem convivi quase que diariamente, aprendi muito e compartilhei tantas experiências: Kherian Gracher, Delvair Moreira, Bruno Aislã, André Luiz, André Cardozo, Fernando Paladine, Luizinho, Vinícius e Jaaziel. Vocês sempre terão um lugar especial em meu coração.

À Lucilene Gutelvil, a quem simplesmente não consigo expressar toda minha gratidão em algumas frases. Você contribuiu significativamente com as condições de possibilidade para a realização desta tese de doutorado.

Agradeço, também, a Geyson e Dalila, pelo acolhimento, amizade, risadas e pescarias nesses últimos meses de moradia em Florianópolis.

Por fim, agradeço à CAPES, por financiar essa pesquisa.

RESUMO

O *Realismo Moral Naturalista* (RMN) é a tese de que há fatos e propriedades morais que são independentes de nossas mentes (realismo moral), que tais fatos e propriedades morais são constituídos ou multiplamente realizados por fatos e propriedades naturais (naturalismo moral) e que as teorias morais substantivas rastreiam as propriedades naturais que constituem as propriedades morais (definicionismo de primeira ordem). O principal desafio a este tipo de teoria metaética é o *Argumento da Terra Gêmea Moral* (ATGM). Os proponentes do ATGM constroem um experimento mental cujo juízo intuitivo é de que falantes de comunidades distintas expressam desacordos morais genuínos. No entanto, as melhores propostas metassemânticas do RMN supostamente não acomodam tal intuição. Assim, o RMN deve ser recusado, uma vez que, argumentavelmente, não possui o diagnóstico correto para o conteúdo semântico dos predicados morais. Na literatura atual, existe ampla resistência por parte dos defensores do RMN em relação ao ATGM. Há quatro estratégias principais de réplica: (i) o ATGM não preserva a similaridade com o Argumento da Terra Gêmea de Hilary Putnam, em relação ao qual é dependente; (ii) é possível haver desacordos morais genuínos mesmo que os predicados dos falantes não compartilhem conteúdo extensional; (iii) ou o juízo intuitivo é enganoso ou há explicações alternativas; (iv) há metassemânticas alternativas para o RMN que não são vulneráveis ao ATGM. Diante dessa controvérsia, o objetivo da presente tese é fornecer uma defesa do ATGM contra essas quatro linhas de ataque. Argumenta-se que o ATGM resiste aos principais problemas apontados pelos seus adversários e, assim, constitui uma dificuldade relevante para a aceitabilidade do RMN. O primeiro capítulo revisa a literatura e introduz as ferramentas conceituais necessárias para a compreensão do debate. Os quatro capítulos seguintes lidam com cada uma das linhas de réplica ao ATGM. O segundo considera a suposta dependência para com o Argumento da Terra Gêmea, de Putnam, e argumenta em favor da autonomia do ATGM. O terceiro discute e defende o pressuposto do ATGM de que desacordos genuínos requerem identidade extensional. O quarto é uma defesa do juízo intuitivo do experimento mental em que o ATGM é baseado. Por fim, o quinto capítulo considera algumas teorias metassemânticas do RMN que, supostamente, não são vulneráveis ao desafio em questão.

Palavras-chave: Metaética. Realismo Moral. Naturalismo Moral. Argumento da Terra Gêmea Moral.

ABSTRACT

Naturalistic Moral Realism (NMR) is the thesis that there are moral properties and facts that are independent of our minds (moral realism), that these properties and facts are constituted or multiple realized by natural properties and facts (moral naturalism), and that substantive moral theories track the natural properties that moral properties are constituted by (first order definism). The main challenge to this kind of metaethical theory is the *Moral Twin Earth Argument* (MTEA). MTEA's proponents build a thought experiment whose intuitive judgement is that speakers from distinct communities express genuine moral disagreements. However, the best NMR's metasemantic proposals supposedly do not accommodate such intuition. Thus, the NMR must be rejected, since, arguably, it does not have the correct account concerning the semantic content of the moral predicates. In the current literature, there is widespread resistance from RMN's advocates to the MTEA. There are four main response strategies: (i) the first one suggests that the MTEA does not preserve the similarity with Hilary Putnam's Twin Earth Argument, on which it is supposedly dependent; (ii) the second holds that it is possible to have genuine moral disagreements even if the speaker's predicates do not share extensional content; (iii) the third claims that either the intuitive judgement is misleading or there are alternative explanations; (iv) and the fourth line of reply maintains that there are alternative metasemantics for NMR that are not vulnerable to the MTEA. Therefore, the aim of this thesis is to provide a defense of the MTEA against these four lines of attack. It is argued that the MTEA resist the main problems pointed out by its critics and, thus, constitute a relevant difficulty for the acceptability of NMR. The first chapter reviews the literature and introduces the conceptual tools needed to understand the debate. The next four chapters deal with the four lines of reply. The second considers the supposed dependence on Putnam's Twin Earth Argument and argues for the MTEA's autonomy. The third discusses and defends the MTEA assumption that genuine disagreements require extensional identity. The fourth is a defense of the intuitive judgment of the thought experiment on which the MTEA is based. Finally, the fifth chapter considers some NMR's metasemantic theories that supposedly are not vulnerable to the challenge at hand.

Keywords: Metaethics. Moral Realism. Moral Naturalism. Moral Twin Earth Argument.

LISTA DE ABREVIATURAS

AITGM: Argumento Invertido da Terra Gêmea Moral

AQA: Argumento da Questão Aberta

ATG: Argumento da Terra Gêmea

ATGM: Argumento da Terra Gêmea Moral

CIE: Condição da Identidade da Extensão

CSE: Condição do Sucesso Explanatório

CSF: Condições Semânticas Formais

CSS: Condições Semânticas Substantivas

DRCC: Desacordo Requer Conflito de Conteúdo

FN: Falácia Naturalística

H&T: T. Horgan e M. Timmons

IUS: Intuição da Univocidade Semântica

LM&D: S. Laurence, E. Margolis e A. Dawson

ME: Mundo Estranho

NSC: Naturalismo Semântico Causal

RE: Realismo Expressivista

RMN: Realismo Moral Naturalista

SMNE: Semântica Moral Normativamente Enriquecida

T: Terra

TG: Terra Gêmea

TGMAC: Terra Gêmea Moral Anti-Consequencialista

TGMD: Terra Gêmea Moral Discursiva

TGME: Terra Gêmea Moral Escravagista

TGMEG: Terra Gêmea Moral Egoísta

TGMS: Terra Gêmea Moral Subjetivista

TRC. Tese da Regulação Causal

TRCM: Tese da Regulação Causal Moral

TRS: Tese Referência/Significado

SUMÁRIO

INTRODUÇÃO

Em filosofia moral é comum distinguir entre *ética normativa*, *ética aplicada* e *metaética*. Na *ética normativa*, os filósofos estão interessados na formulação de teorias sobre quais ações são certas e erradas, boas e más e o que faz com que assim o sejam. Há um objetivo prático e um objetivo teórico, portanto. O objetivo prático de uma teoria é fornecer um procedimento de decisão completo e seguro para que possamos decidir se um curso de ação é correto, permissível ou injusto. O objetivo teórico consiste em fornecer uma justificação sobre o que faz com que ações certas sejam certas e ações erradas sejam erradas. No entanto, não é sempre óbvio como essas teorias se aplicam a casos particulares complexos. Por isso, há uma área especificamente dedicada à discussão sobre como as teorias e princípios gerais da ética normativa se aplicam a esses casos difíceis. Trata-se da *ética aplicada*. Nessas duas áreas da filosofia moral, os filósofos constantemente defendem juízos morais tais como 'o aborto deve ser permitido', 'é errado causar dor a animais não-humanos', 'intervenções biotecnológicas com fins não terapêuticos devem ser proibidas', e assim por diante. Mas, para além disso, podemos fazer um tipo de pergunta de ordem mais abstrata, como: *o que estamos fazendo* quando proferimos juízos morais e o que são tais juízos morais? Por exemplo, são meramente convenções sociais? Envolvem expressão de emoções particulares a respeito de certas ações? Tratam apenas de opiniões que podem variar amplamente ou há respostas corretas e objetivas às controvérsias morais? Se há respostas objetivas, o que faz com que assim o sejam? Comando divino? Razão? Essa área mais geral e abstrata de investigação é o domínio da *metaética*.

Mas, por que alguém deveria se ocupar com problemas metaéticos? Afinal, pode parecer que a preocupação central dos filósofos deva ser com os problemas da ética normativa e aplicada. As perguntas sobre como devemos conduzir nossas vidas e o que significa agir corretamente podem parecer mais importantes, já que poucas questões são tão inevitáveis e difíceis quanto essas. Além disso, a metaética parece ser um domínio exclusivamente teorético e totalmente distante da vida prática. Isso, todavia, não é o caso. Embora haja controvérsia acadêmica a respeito das implicações da metaética para a ética normativa (e vice-versa), há consenso sobre a fundamentalidade da metaética para uma compreensão mais robusta sobre a natureza da moralidade e sua relevância para um melhor entendimento das discussões nos domínios da ética normativa e aplicada. Em outras palavras, resolver problemas metaéticos pode servir para a consolidação de uma teoria normativa cujos alcances atingem problemas morais ordinários. Considere, por exemplo, o seguinte fato sobre a moralidade: a ocorrência de

desacordos morais. Desacordos morais ordinários podem revelar um desacordo fundamental num nível mais abstrato, o nível dos problemas metaéticos. Por exemplo, ao discutir algum problema moral, *x* pode pressupor que há uma relação estreita entre moralidade e religião enquanto *y* pode negar isso; *x* pode manter que a moralidade é o resultado de convenção social, enquanto *y* pode sustentar que há verdades morais independentes do que as pessoas pensam; *x* pode aceitar que não há fatos morais e que as opiniões morais são sempre relativas, enquanto *y* pode manter a objetividade da moralidade; *x* pode sustentar que os termos morais servem apenas para expressar emoções, enquanto *y* pode pressupor que não há lugar para emoções na moralidade. E, como a vida moral ordinária torna claro, pode ser que tais desacordos morais, extensamente compartilhados, sejam o resultado de desacordos num *segundo nível* da moralidade: o nível metaético. Entender essas questões é de fundamental importância para o domínio prático da moralidade. É por isso que a metaética expandiu-se amplamente nas últimas décadas devido à possível contribuição para um melhor entendimento da moralidade em geral.

O presente trabalho é sobre metaética. E, de modo a explicar adequadamente a natureza dessa investigação e os passos seguidos, proponho o seguinte nesta seção introdutória. Primeiro, apresento, grosso modo, o contexto da discussão e o problema que será analisado. Segundo, esclareço a tese que pretendo defender e os passos seguidos em cada um dos capítulos. Por fim, faço alguns comentários sobre a localização teórica desse trabalho.

Contexto e problema

O naturalismo moral sustenta que há fatos e propriedades morais e que esses fatos e propriedades morais são *naturais*. Como afirmam Billy Dunaway e Tristram McPherson, "pode-se pensar que, seja qual forem os desafios que os realistas naturalistas enfrentem, eles estão em solo firme no que diz respeito à semântica" (DUNAWAY & MCPHERSON, 2014, p. 1). A razão disso é que os naturalistas sustentam que os predicados morais funcionam exatamente como os predicados não morais, de modo que tudo o que precisam fazer é adotar a melhor teoria semântica disponível sobre os termos não morais e aplica-la aos termos morais. No entanto, esse é um dos maiores problemas para as teorias naturalistas e sua tradição remonta ao *Principia Ethica* (1903) de George Edward Moore.

Moore apresentou uma objeção ao naturalismo moral que ficou conhecida sob o nome de *Argumento da Questão Aberta* (doravante, AQA). Moore formulou um tipo de *teste semântico* que supostamente mostrava que qualquer relação de identidade entre propriedades

morais e propriedades naturais era falsa. Em termos muito simples, a ideia era a seguinte: faz sentido (para falantes competentes) perguntar sobre uma propriedade supostamente definidora de 'bom' se ela própria é boa? Se sim, a propriedade definida e a propriedade definidora não podem ser analiticamente equivalentes, uma vez que não compartilham o mesmo conjunto de características semânticas. Tal ferramenta argumentativa teve uma influência tremenda no desenvolvimento posterior da metaética. O naturalismo moral fora amplamente considerado como implausível e, dado que a teoria alternativa proposta pelo próprio Moore (o não-naturalismo moral) tinha grandes custos metafísicos e epistêmicos, houve um crescimento expressivo de abordagens não-cognitivistas por várias décadas.

No entanto, a metaética testemunhou uma espécie de renascimento do naturalismo moral na década de 80 devido, principalmente, a vários avanços em filosofia da linguagem. Com o desenvolvimento das teorias semânticas externalistas, por filósofos como Saul Kripke (1972) e Hilary Putnam (1975), podia-se, agora, responder de forma convincente ao AQA de Moore. E foi Richard Boyd (1988) quem propôs uma forma de externalismo semântico para os termos morais e desenvolveu uma das versões mais sofisticadas e completas de naturalismo moral. Grosso modo, tal teoria sustenta que a referência de um termo *t* é estabelecida por conexões causais que determinam que tais e tais objetos pertencem à extensão de *t*. Essa ideia têm sido chamada de *Tese da Regulação Causal* (TRC). Aplicando-a à ética, temos o seguinte: para cada termo moral *t*, há uma propriedade natural *N* tal que, somente *N* e nenhuma outra propriedade, regula causalmente o uso de *t*. Assim, um termo moral, como 'bom' refere uma propriedade – *bondade* – que tem uma constituição natural e é exatamente essa propriedade natural que determina a quais instâncias de bondade nós podemos aplicar o termo moral em questão. Para Boyd, a extensão de um termo moral não é constituída por uma propriedade natural apenas, mas por um conjunto de propriedades naturais e a determinação do que é tal conjunto é trabalho da investigação presente e futura (por isso, a especificação exata de qual propriedade *N* a propriedade moral refere, é aberta, não exaustiva). E, como naturalista, Boyd sustenta que o comportamento semântico dos termos morais é igual ao comportamento semântico dos termos não morais. O trabalho de Boyd faz parte de um programa naturalista mais amplo levado adiante por vários filósofos, dentre eles Nicholas Sturgeon e David Brink. Conjuntamente, esses filósofos desenvolveram o que ficou conhecido como *Realismo de Cornell*, que constitui uma das abordagens naturalistas mais difundidas.

No entanto, em uma série de trabalhos (1991, 1992a, 1992b), Terence Horgan e Mark Timmons (H&T) se dedicaram a refutar tal forma de naturalismo moral. Eles sugeriram que, se

a teoria de Boyd é verdadeira, então falantes competentes devem, ao menos tacitamente, reconhecer a sua verdade. Neste sentido, formularam um experimento de pensamento a partir do qual construíram um argumento que é, supostamente, devastador ao naturalismo moral, seja à teoria de Boyd seja a versões contemporâneas alternativas. Tal argumento ficou conhecido como *Argumento da Terra Gêmea Moral* (ATGM). H&T nos pedem para imaginar um planeta alternativo (Terra Gêmea Moral – TG) que, para *quase* tudo que há na nossa Terra (T), há uma réplica em TG. Os habitantes de TG usam termos morais como 'bom', 'correto', 'justo' e esses termos possuem as mesmas características *formais* que os nossos termos morais; isto é, eles usam os predicados morais para avaliar ações, pessoas e instituições, para deliberar sobre o seu bem-estar, estão dispostos a agir do modo que corresponda aos seus julgamentos sobre 'certo' e 'errado' etc (H&T, 1992, p. 164). A única diferença entre T e TG é que a propriedade natural que regula causalmente o uso de 'bom' nos dois planetas é diferente (digamos *N* e *N**). O ponto de H&T é o seguinte: dada uma suposta discussão entre membros de T e TG em que uns afirmassem, por exemplo, que 'mentir é errado' e outros que 'mentir não é errado', se o naturalismo moral fosse verdadeiro teríamos que concluir que não haveria desacordo moral genuíno, já que os habitantes dos dois planetas estariam predicando coisas diferentes sobre a ação de mentir (*N* e *N**). Em resumo, as afirmações não seriam inconsistentes. Todavia, sugerem H&T, falantes competentes tem a intuição de que os habitantes de T e TG estão em um desacordo moral genuíno. E tal intuição parece ser amplamente compartilhada, forte e persistente. H&T sustentam que temos boas razões para preservar as intuições linguísticas de falantes competentes. Portanto, temos evidência para recusar o naturalismo moral.

O argumento proposto por H&T gerou uma sofisticada controvérsia e têm sido um dos principais focos de discussão sobre a plausibilidade do naturalismo moral nas últimas três décadas. Defender uma posição naturalista em ética requer que se tenha algum tipo de resposta ao ATGM. E é precisamente isso que muitos naturalistas vem tentando fazer nos últimos anos. Desde a apresentação do ATGM por H&T, têm aparecido na literatura uma variedade de réplicas que visam livrar o naturalismo moral dessa objeção. Essas réplicas podem ser organizadas em quatro grupos.

a) Uma das inspirações do ATGM é o experimento de pensamento apresentado por Putnam (1975) na tentativa de recusar o descritivismo semântico, a saber o Experimento da Terra Gêmea, e, em vários momentos, H&T fazem estipulações bastante similares a tal experimento. Assim, alguns críticos do ATGM têm

argumentado que há uma espécie de dependência estrita entre esses dois experimentos mentais, de modo que, qualquer ponto de não analogia implicaria em problemas para o ATGM.

b) Um segundo grupo de réplicas busca questionar a tese de que desacordos morais genuínos requerem identidade conceitual sobre os predicados morais usados pelos falantes. Essa é uma das pressuposições do ATGM, uma vez que, da suposta intuição de que há desacordo genuíno entre habitantes de T e TG, conclui-se que os predicados morais devem possuir convergência semântica.
c) Um terceiro grupo de réplicas visa atacar a legitimidade da intuição de que há desacordo genuíno entre membros de T e TG. Tal como em todo experimento mental, os juízos intuitivos desempenham um papel central e os críticos do ATGM que aderem a essa estratégia argumentam, de formas diferentes, que tais juízos não são confiáveis.
d) O quarto grupo visa formular teorias semânticas alternativas à proposta de Boyd. Vários filósofos inclusive concordam que o ATGM refuta o naturalismo moral, mas apenas se se pressupor que a abordagem de Boyd é a única disponível. Assim, buscam elaborar teorias que não sejam vulneráveis ao ATGM.

A discussão sobre se essas críticas representam um *knock out* decisivo contra o ATGM ou se este argumento sobrevive a tais réplicas é de fundamental importância. Se o ATGM persiste, mesmo considerando o número expressivo de críticas, então o naturalismo moral encontra-se diante de um sério problema e isso conta muitos pontos em favor de teorias metaéticas concorrentes. Por outro lado, se os críticos de H&T estão corretos e o ATGM não é ameaça alguma ao naturalismo moral, então um dos mais sofisticados argumentos da metaética contemporânea deve ser abandonado e o naturalismo moral não precisa fazer concessões a seus adversários. Além disso, não há na literatura trabalhos que se dediquem a analisar de forma abrangente essas críticas ao ATGM, de modo que uma discussão detalhada desses argumentos pode resultar numa contribuição importante.

O que defenderei

Embora empreguem diferentes estratégias de ataque, é seguro dizer que grande parte da literatura converge no sentido de recusar o ATGM. No entanto, há quem tente argumentar que

esses ataques não fornecem razões suficientes para rejeitarmos o desafio semântico ao RMN. Os próprios H&T (2000, 2015) e Michael Rubin (2008, 2014a, 2014b, 2014c, 2015) são os dois exemplos mais expressivos. No presente trabalho, me incluo nesse segundo grupo e tento avançar alguns passos nessa direção questionando essa aparente convergência de que o ATGM não ameaça o RMN. Devo dizer que possuo grande débito para com o trabalho desenvolvido por Rubin a respeito do ATGM. Em muitas partes endosso e avanço vários de seus argumentos.

De forma mais específica, no decorrer desse trabalho buscarei defender que parte relevante dos ataques ao ATGM não nos fornecem razão suficiente para abandonarmos tal argumento. Assim, o que esse trabalho pretende mostrar é que o ATGM resiste aos principais problemas apontados pelos seus adversários e que, consequentemente, permanece sendo a razão de um importante *custo* para a aceitabilidade do naturalismo moral.

Neste sentido, irei percorrer o seguinte plano. No primeiro capítulo, tentarei introduzir de forma mais aprofundada o debate sobre o RMN, o ataque de H&T a esta teoria e a discussão subsequente. Com isso, pretendo munir o (a) leitor (a) com as ferramentas necessárias para a compreensão dos pontos que buscarei desenvolver nos capítulos seguintes. Portanto, trata-se de um capítulo de exposição da literatura. A partir disso, iniciarei, propriamente, a parte argumentativa.

Nos quatro capítulos seguintes, abordarei cada uma das quatro estratégias de ataque ao ATGM. No segundo capítulo, considerarei o primeiro grupo de críticas. Aqui os principais representantes são Stephen Laurence *et. al*. (1999) e Heimir Geirsson (2014). Essa primeira abordagem não visa atacar nenhuma das premissas específicas que constituem o ATGM, mas desenvolve uma espécie de crítica por analogia. Para esses filósofos, existe uma dependência num sentido relevante entre o ATGM e o Experimento da Terra Gêmea desenvolvido por Putnam de modo que há uma série de pressupostos que devem ser preservados. O ponto é que o ATGM e o experimento de Putnam apresentam algumas divergências, tal como argumentam os defensores dessa abordagem. E o resultado é que o ATGM não pode ser plausivelmente mantido. Desse modo, meu objetivo nesse segundo capítulo será argumentar em favor da autonomia do ATGM.

No terceiro capítulo, analisarei o segundo grupo de críticas ao ATGM. Aqui, como veremos os ataques são direcionados à segunda premissa desse argumento e são representados principalmente por David Copp (2000), David Merli (2002) e David Plunkett e Tim Sundell (2013). O desafio semântico desenvolvido por H&T depende da pressuposição de que desacordos morais genuínos requerem identidade na extensão dos predicados usados pelos

falantes. De diferentes formas, os críticos do ATGM tentam recusar esse pressuposto argumentando que dois falantes podem estar em desacordo genuíno mesmo que os termos relevantes através dos quais se expressam possuem referências distintas. Nesse capítulo, tentarei defender o pressuposto de H&T desses três ataques.

No quarto capítulo, lidarei com o terceiro grupo de críticas ao ATGM. Essas réplicas visam atacar a intuição central do experimento de pensamento por detrás do ATGM, que, como veremos, constitui a terceira premissa desse argumento. Há ampla variedade de ataques nesse ponto, mas os três trabalhos mais completos são de Neil Levi (2011), Andrea Viggiano (2008) e John Sonderholm (2013). Aqui, novamente, tentarei defender que os argumentos apresentados por esses filósofos não são definitivos para recusarmos o ATGM.

No quinto e último capítulo, considerarei duas réplicas que constituem o quarto grupo de ataques ao ATGM. Como veremos, a estratégia dos críticos aqui é desenvolver teorias semânticas alternativas à disponível para o RMN argumentando que o ATGM não atinge tais teorias. Há uma pluralidade de desenvolvimentos neste sentido. Irei me limitar às propostas de David Brink (2001) e David Copp (2001). Buscarei argumentar que a abordagem de Brink não representa grandes avanços para o RMN diante do ATGM e que a melhor alternativa seria a adoção do tipo de teoria defendida por Copp. No entanto, apesar de ter alguma resposta ao ATGM, o defensor do RMN teria que se distanciar significativamente do tipo de teoria que tradicionalmente endossa.

A localização da tese

Como afirma David Enoch (2011), seria ótimo se uma teoria metaética em específico tivesse todas as vantagens possíveis e, quando comparada com visões alternativas, apresentasse sempre superioridade teórica. No entanto, é um ideal muito rigoroso pensar que uma teoria é *decisivamente* melhor do que suas concorrentes. A metaética é a tentativa de explicar as características e os problemas que dizem respeito à linguagem e ao pensamento moral ordinário e as teorias metaéticas fornecem explicações que variam em grau de atratividade a respeito de pontos específicos. Uma teoria pode fornecer um bom argumento semântico a seu favor e, no entanto, enfrentar problemas para explicar seus compromissos epistêmicos. Ou, pode apresentar uma abordagem persuasiva a respeito da psicologia moral envolvida no ato de se fazer juízos morais, mas sofrer baixas ao tentar fornecer explicações ontológicas sobre as propriedades morais, por exemplo.

Assim, sugere Enoch, a atitude mais razoável é encarar o terreno das teorias metaéticas como um placar de soma e perda de pontos metaéticos. Uma vez que não existem teorias perfeitas possuidoras de pontos positivos apenas, a atitude mais razoável seria fazer uma espécie de pesagem entre os pontos positivos e negativos, entre os *custos* e *benefícios*, e escolher a teoria que seja a melhor no geral. Ou seja, buscar aquela teoria que, embora inevitavelmente perca pontos explanatórios a respeito de alguns problemas, ganhe pontos a respeito de outros desafios teóricos e que, no total, compensem os pontos perdidos. A teoria que somar mais pontos no placar metaético, por assim dizer, terá mais razões a seu favor.

O trabalho do metaeticista permeia esses pontos específicos a respeito dos custos e benefícios das teorias. E, como em qualquer área de especialização, há divisão do trabalho. Assim, não é incomum encontrar trabalhos que se propõe a fornecer razões a favor do realismo moral, por exemplo, mas cuja contribuição é a respeito de um aspecto muito específico que, no máximo, resulta na soma de alguns pontos a favor de tal teoria. Ou, ainda, trabalhos que objetivam apresentar razões robustas contra o realismo moral, enquanto teoria geral, e propõem um argumento sobre tal ou tal pressuposto particular da teoria. Esse tipo de prática não se trata de exagero retórico por parte dos filósofos, pois a metaética contemporânea é marcada por ampla variedade de teorias, de modo que é de extrema exigência esperar que um trabalho lide com a totalidade dos argumentos e objeções de uma área.

Neste sentido, há sempre uma forma *macro* e uma forma *micro* de se olhar para um trabalho em metaética. De acordo com a primeira, olhamos para o tipo de contribuição feita desde uma perspectiva mais geral. Dizer que tal ou tal argumento visa mostrar a verdade do realismo moral é uma forma macro de se encarar a proposta. De acordo com forma micro, olhamos para os pontos particulares que compõem o placar metaético de uma teoria. Dizer que o subjetivismo não fornece uma boa explicação para o fenômeno da objetividade moral, por exemplo, é manter um olhar micro. Provavelmente, esse custo do subjetivismo, sozinho, não constitui razão suficiente para rejeitarmos a teoria, mas subtrai alguns de seus pontos. Um trabalho em metaética sempre mantem uma perspectiva macro, mas, frequentemente, sua contribuição se dá no nível micro.

É com essa distinção em mente que gostaria que o presente trabalho fosse encarado. Há uma perspectiva macro sendo seguida. Mas a tentativa de fornecer uma contribuição se dá num nível muito mais específico e localizado. Em termos macros, irei defender que devemos rejeitar o naturalismo moral e optar por alguma forma de não-cognitivismo. Esses são os arredores gerais da discussão. Mas o trabalho real é a respeito de um problema específico que pode

determinar apenas a soma de alguns pontos a favor do não-cognitivismo e contra o naturalismo no placar metaético. Desse modo, se obtiver sucesso, certamente não estabeleço a falsidade do naturalismo moral, e nem a verdade do não-cognitivismo, mas, apenas, que há certos pontos contra o naturalismo moral e que podem ser relevantes na hora de se pesar entre os custos e benefícios de tal teoria.

CAPÍTULO 1 – NATURALISMO ÉTICO E O ARGUMENTO DA TERRA GÊMEA MORAL

1. Introdução

O naturalismo moral é a tese de que propriedades morais são propriedades naturais (seja tal relação de *identidade* ou *constituição*). Desde os princípios da metaética como uma área estabelecida da filosofia moral, essa teoria enfrenta grande número de ataques, fundamentalmente com George Edward Moore. Em seu *Principia Ethica* (1903), Moore apresentou o que ficou conhecido como *Argumento da Questão Aberta* (AQA) para sustentar que o naturalismo moral comete o erro de confundir propriedades categoricamente distintas (Cf. DALL'AGNOl, 2014). Dada a influência desse argumento, somada à insatisfação com o tipo de teoria positiva proposta pelo próprio Moore, muitos metaeticistas consideraram o naturalismo moral como uma teoria fadada ao fracasso e as décadas seguintes testemunharam um significativo florescimento do não-cognitivismo moral, bem como novas objeções ao naturalismo (por exemplo, o *Argumento da Tradução* proposto por Richard Hare (1952)). No final da década de oitenta, entretanto, um grupo de filósofos apresentou uma versão mais sofisticada e promissora do naturalismo moral fazendo com que essa teoria ganhasse mais espaço e respeito. Metaeticistas como David Brink (1984, 1989), Nicholas Sturgeon (1985) e Richard Boyd (1988), sustentaram que se poderia mostrar a plausibilidade do naturalismo seja aplicando estratégias argumentativas do realismo científico seja desenvolvendo uma metassemântica externalista inspirada em Hilary Putnam (1975) e Saul Kripke (1980). Além disso, Frank Jackson e Phillip Pettit (1995), inspirados nos novos desenvolvimentos em filosofia da mente, particularmente o Funcionalismo, também fizeram sólida defesa do naturalismo moral. Este naturalismo "renovado", além de – argumentativamente – dar conta de uma série de objeções usuais, evita o desafio colocado pelo AQA ou pelo argumento de Hare.

No entanto, há alguns filósofos que estão revisitando as objeções clássicas e tentando adaptá-las a essas versões contemporâneas do naturalismo moral. Numa série de artigos, Terence Horgan e Mark Timmons (1990-1991, 1992a, 1992b) (doravante, H&T) desenvolveram um argumento, que ficou conhecido como o *Argumento da Terra Gêmea Moral* (ATGM), que, supostamente, é fatal ao "novo" naturalismo. Eles sustentam que o naturalismo moral está em sérios problemas, uma vez que suas versões iniciais são refutadas pelo OQA e

suas versões atuais pelo ATGM. Esse conflito entre o desafio colocado pelo ATGM e os defensores do naturalismo moral gerou uma frutífera e refinada controvérsia acadêmica nos últimos anos e a discussão continua em aberto.

É a plausibilidade e o alcance desse desafio ao naturalismo moral que pretendo analisar no presente trabalho. Assim, nesse primeiro capítulo gostaria de introduzir o (a) leitor (a) neste debate. Na seção 2 considerarei, brevemente, o que é o naturalismo moral. Em qualquer área da filosofia é sempre difícil estabelecer uma definição acabada sobre o que os filósofos entendem sobre determinado conceito. Por isso, apresentarei apenas algumas características constitutivas de uma teoria naturalista julgando ser suficiente para que o (a) leitor (a) acompanhe o andamento da discussão (apresentarei mais detalhadamente uma teoria naturalista em seção posterior). Na seção 3, irei expor o desafio clássico ao naturalismo moral, o AQA. Dado que muitas vezes há confusões entre AQA e Falácia Naturalística dedicarei um espaço para alguns esclarecimentos. O AQA desempenhou tremenda influência no desenvolvimento da metaética. Por isso, dedicarei a seção 4 para considerar um pouco dessa influência histórica. Dado que o AQA inspirou outros tipos de ataques ao naturalismo moral, irei considerar o *Argumento da Tradução* de R. Hare. Como veremos, tal argumento desempenha papel importante no contexto deste trabalho. O retorno do naturalismo moral em nova roupagem foi antecedido pela descoberta de um problema que impactou o tipo de teoria amplamente aceita até então, o *Problema Frege-Geach*. Assim, explicitarei este problema na seção 5. Na seção 6, apresentarei uma das variantes mais influentes do naturalismo moral contemporâneo e que servirá de pano de fundo no decorrer desse trabalho, a saber, o *Realismo de Cornell*. Na seção 7, irei expor ATGM. Este argumento representa um retorno aos desafios clássicos ao naturalismo, a saber, o AQA e o *Argumento da Tradução*. Há uma série de ataques ao ATGM na literatura contemporânea. Neste sentido, na seção 8, finalmente, agruparei esses ataques em quatro categorias argumentativas. Todo esse caminho é necessário para que se possa compreender o problema do presente trabalho e os rumos que tentarei desenvolver nos capítulos seguintes.

2. Naturalismo Moral

O naturalismo moral é uma forma de realismo moral. É extremamente difícil fornecer uma definição exaustivamente explanatória de realismo moral, dada a diferença dramática com que os filósofos usam este termo. Então, parece ser mais frutífero adotar uma estratégia de

sim/não elencando uma série de teses e dizendo quais são parte do realismo moral e quais não são. Seguindo Peter Railton (2018), podemos dizer que os realistas morais[1]:

(i) Sustentam que enunciados morais são *verdadeiros ou falsos*;
(ii) Sustentam que pelo menos *alguns* desses enunciados são de fato verdadeiros;
(iii) Fornecem uma *interpretação literal* para as sentenças morais;
(iv) Sustentam que a verdade, neste contexto moral, é *independente das opiniões* das pessoas;
(v) Sustentam que essas verdades morais independentes das opiniões são *substantivamente explanatórias*.

Emotivistas tais como Alfred Ayer (1936) e Charles Stevenson (1937) negavam (i). A inclusão de (ii) é estritamente necessária para classificar uma posição realista, pois um antirrealista pode aceitar (i) e negar (ii). Os teóricos do erro, por exemplo, defendem que os enunciados morais são passíveis de verdade e falsidade e, no entanto, são sistemática e uniformemente falsos (Cf. MACKIE, 1977). Além disso, realistas morais são *literalistas* e isso quer dizer o seguinte. Um não-literalista pode dizer que a sentença moral 'a tortura é errada' significa que 'a sociedade considera que a tortura é errada', 'eu desaprovo a tortura' ou 'boooooooo para a tortura'. Um literalista não considera que há um significado implícito nas afirmações morais. Portanto, em 'a tortura é errada' estamos atribuindo uma propriedade moral a um estado de coisas e a instanciação dessa propriedade moral pelo estado de coisas gera um fato moral. Ainda, para contar como realista, uma teoria deve sustentar (iv). Isso significa que o realista moral nega que um enunciado moral como 'a tortura é errada' é verdadeiro para alguns e falso para outros. A condição expressa em (iv) elimina o relativismo moral. Por último, dizer que essas verdades literais independentes das opiniões das pessoas são substantivamente explanatórias significa que nós temos tais e tais percepções e crenças morais porque o mundo *é realmente* tal como o percebemos.

Há duas formas de realismo moral. O *naturalismo moral* e o *não-naturalismo* moral. Ambos os tipos de teoria endossam as teses acima. A diferença central está na explicação ontológica que se atribui às propriedades morais. Não-naturalistas mantêm que fatos e propriedades morais são de um tipo inteiramente diferente dos fatos e propriedades naturais. Os

[1] Outro excelente trabalho que trata justamente do problema da definição do realismo moral é FINLAY, S. Four Faces of Moral Realism. *Philosophy Compass*, 2, (6), p. 820-849.

naturalistas argumentam que fatos e propriedades morais são naturais. Dizer que o naturalismo moral é a tese de que propriedades morais são naturais, obviamente, não é muito informativo até que se diga o que são propriedades naturais. Aqui, também, há desacordo. No entanto, podemos nos contentar com a definição geral de que propriedades naturais são aquelas propriedades que constituem o objeto de estudo das ciências naturais. Apesar das divergências, essa visão parece ser dominante[2],[3].

Apresentarei uma visão mais precisa do naturalismo moral adiante quando considerarmos o Realismo de Cornell. Mas, como é necessário ter alguma ideia geral para podermos prosseguir, podemos fornecer uma caracterização breve. Segundo Billy Dunaway e Tristran McPherson (2016), o naturalismo moral envolve *quatro* teses principais. (i) *Descritivismo*, isto é, a ideia de que sentenças morais descrevem ou representam o mundo. Esta visão contrasta com teorias não-cognitivistas, por exemplo, que sustentam que os termos morais têm um significado distintivo associado a estados psicológicos conativos. (ii) *Não-indexicalidade*. Considere as sentenças: 'Está frio *aqui*' e 'É *ilegal* consumir bebidas alcoólicas na rua'. Uma compreensão completa dessas sentenças requer mais informação sobre os seus contextos (onde é o 'aqui' e informação sobre qual sistema legal está em questão). Essas sentenças são relativas ao contexto a que se referem (certamente há lugares em que não está frio e sistemas legais em que consumir bebidas alcoólicas na rua não é ilegal). Termos como 'aqui' e 'ilegal' são indexicais. Teorias metaéticas relativistas, subjetivistas ou contextualistas mantêm que os termos morais são indexicais. O naturalismo moral nega isso. (iii) Propriedades morais são *metafisicamente explanatórias*. Esta perspectiva envolve um comprometimento ontológico com a existência de propriedades morais. Já que as propriedades morais alegadamente desempenham um papel explicativo, os naturalistas sustentam elas são parte da constituição natural do mundo. Esta ideia permite que se distinga o naturalismo moral da teoria

[2] Veja, por exemplo, Moore, 1903, p. 92; Sturgeon, 2003, p. 543; Brink, 1989, p. 22; Shafer-Landau, 2006, p. 212-213 e Timmons, 1999, p. 12.

[3] David Copp (2003) fez uma espécie de taxonomia sobre as principais definições de propriedades naturais. Ele diz que há quatro abordagens centrais na literatura: o modelo reducionista, o modelo das definições ostensivas, o modelo das definições metafísicas e o modelo das definições epistêmicas. O modelo reducionista consiste em identificar um domínio de propriedades indiscutivelmente naturais e empregar estratégias de identificação das propriedades morais com essas propriedades naturais. O modelo das definições ostensivas consiste em apontar supostas propriedades naturais no mundo. O modelo metafísico possui quatro variantes. De acordo com a primeira, diz-se que propriedades naturais são descritivas ou factuais. De acordo com a segunda, propriedades morais são propriedades que possuem eficácia causal. De acordo com a terceira, é natural tudo aquilo que faz parte da dimensão espaço-temporal. E de acordo com a quarta variante, o mundo natural é o mundo materialista ou fisicalista. Por último, o modelo das definições epistêmicas é a ideia de que o mundo natural é tudo aquilo que é estudado pelas ciências naturais. Por ser a visão dominante, estou sugerindo que adotemos para compreendermos o tipo de teoria metaética que estamos considerando neste trabalho.

do erro, ficcionalismo ou quase-realismo, por exemplo. (iv) Propriedades morais são redutíveis a (ou metafisicamente compatíveis com) *propriedades naturais*.

O naturalismo moral é uma visão bastante atrativa. Dado que é uma forma de realismo, ela permite que o naturalista acomode a ideia de que os juízos morais são objetivos e explique como e porque os juízos morais são passíveis de verdade e falsidade. Também promete dar sentido à ideia de conhecimento moral e, por conseguinte, estabelecer que há respostas corretas na moralidade. Como uma forma de realismo é, argumentativamente, preferível a visões realistas alternativas, tais como o não-naturalismo, pois não se compromete com a existência de propriedades morais *sui generis*. Portanto, não há nenhum mistério no que diz respeito aos seus compromissos ontológicos e epistêmicos. Ainda, o naturalismo moral parece estar apto a acomodar a ideia de que o certo e o errado não são uma questão de opinião subjetiva, pois os fatos morais são independentes das convicções que as pessoas, em geral, mantêm. Assim, se se puder mostrar a plausibilidade dessa teoria, há muitos ganhos teóricos.

Todavia, o naturalismo moral é alvo de vários ataques. De longe, o mais famoso desses ataques é o *Argumento da Questão Aberta* (AQA), que fora proposto por Moore em 1903, mas cujos alcances são atuais. O *Principia Ethica* de Moore fundou a metaética como disciplina filosófica. E o AQA, proposto nesse livro, definiu seus rumos. Passemos, então, a uma exposição de tal argumento contra o naturalismo moral.

3. Argumento da Questão Aberta e Falácia Naturalística

Por que hoje ainda insistimos na importância do AQA se os metaeticistas já mostraram que – da forma como Moore o apresentou – o argumento não nos convence de falácia ou erro conceitual algum? Esta questão é retoricamente colocada por Stephen Darwall, Allan Gibbard e Peter Railton (1992) à fim de chamar atenção para a influência e persistência de tal argumento. A resposta é terreno comum na metaética: o que Moore descobriu não foi a prova de uma falácia,

> [...] mas de um dispositivo argumentativo que traz à tona efetivamente, embora implicitamente, certas características de 'bom' ... que parecem ser obstáculo para a aceitação de que qualquer definição naturalista ou metafísica [de 'bom'] seja correta" (DARWAL, GIBBARD & RAILTON, 1992, p. 115-116).

Esse argumento teve um impacto impressionante no desenvolvimento da metaética. Quando Moore o apresentou, os filósofos ou o endossaram completamente ou foram pelo menos inspirados a pensar que deveria haver algo errado com o naturalismo moral.

Muitas vezes, esse erro que Moore acusa os naturalistas de cometerem, é referido pela expressão 'falácia naturalística' (FN). Na verdade, esta é a expressão que encontramos no *Principia Ethica*. 'Argumento da questão aberta' se tornou usual devido ao argumento que Moore apresenta no intuito de provar a suposta falácia. Há muita discussão sobre o que é exatamente a FN e o AQA. Neste trabalho, não pretendo entrar nos pormenores dessa controvérsia. O objetivo é apresentar uma versão suficientemente bem difundida dessas noções para que se possa notar a sua influência na metaética e, principalmente, no argumento que irei discutir nesse trabalho[4].

Neil Sinclair (2018) elenca quatro modos pelos quais a FN é frequentemente referida. Comete a FN quem:

(i) define ou analisa os *conceitos* morais em termos de conceitos não morais, naturais ou metafísicos;
(ii) toma os *termos* morais como sendo sinônimos com termos não morais, naturais ou metafísicos;
(iii) *identifica* propriedades morais com propriedades não morais, naturais, metafísicas ou alguma outra propriedade complexa;
(iv) *deriva* conclusões morais (ou 'dever') a partir de premissas não morais (ou 'ser').

Como podemos notar, (i) e (ii) são versões semânticas da suposta falácia, (iii) é uma versão ontológica e (iv) é uma versão inferencial. Neste trabalho, seguirei a interpretação de Fred Feldman (2018). Tal interpretação concilia a versão semântica e a versão ontológica da FN. Eis o que ele considera ser a tese fundamental de Moore:

> FN: uma pessoa comete a falácia naturalística se, e somente se, ela *identifica* alguma propriedade avaliativa com alguma propriedade natural ou metafísica (FELDMAN, 2018, p. 37).

[4] Ao (à) leitor (a) interessado (a) nas discussões sobre como melhor interpretar a FN e o AQA, recomendo *The Nataralistic Fallacy*, de Neil Sinclair (2018). Trata-se de uma coletânea de artigos de diferentes autores sobre problemas específicos a respeito da FN e do AQA.

E como algumas passagens do *Principia* sugerem que Moore está colocando a FN em termos semânticos, e não ontológicos, é possível fazer isso sem grandes alterações. Basta assumir duas coisas: (a) que predicados expressam propriedades e (b) que se uma propriedade expressa por um predicado é avaliativa, então o predicado também o é (o mesmo vale para propriedades naturais) (FELDMAN, 2018. p. 40). Assim, temos a formulação semântica da FN:

> FN(s): uma pessoa comete a falácia naturalística se, e somente se, ela sustenta que uma *expressão* avaliativa é sinônima com uma expressão naturalista ou metafísica.

Essa variante faz com que compreendamos melhor porque Moore as vezes parece colocar a FN ora em termos de *identificação* ora em termos de *sinonímia*. Quando coloca em termos de identificação está se referindo à versão ontológica. Quando coloca em termos de sinonímia está se referindo à versão semântica. Ambas as versões são instâncias do mesmo suposto erro.

Agora que já sabemos qual é este erro denunciado por Moore, que tipo de argumento ele apresenta em favor de tal acusação? Ele formulou um *teste semântico* cujo objetivo era tornar claro que toda identificação entre propriedades morais e propriedades naturais deve ser falsa. Foi o que ficou conhecido como AQA (a reconstrução de tal argumento é baseada no parágrafo 13 do *Principia Ethica*. A passagem pode ser lida de diferentes formas e mais de um modo de entender o argumento ali contido. Não pretendo entrar em controvérsias exegéticas e, por isso, sigo a interpretação de Alexander Miller (2011), que é suficientemente difundida na literatura. Com o perdão do (a) leitor (a), citarei a passagem no original, já que é bastante árduo compor uma tradução suficientemente precisa.

> The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition may be offered, it may always, be asked, with significance, of the complex so defined, whether it is itself good. To take, for instance, one of the more plausible, because one of the more complicated of such proposed definitions, it may easily be thought, at first sight, that to be good may mean to be that which we desire to desire. Thus if we apply this definition to a particular instance and say "When we think that A is good, we are thinking that A is one of the things which we desire to desire," our proposition may seem quite plausible. But, if we carry the investigation further, and ask ourselves "Is it good to desire to desire A?" it is apparent, on a little reflection, that this question is itself as intelligible, as the original question, "Is A good?"—that we are, in fact, now asking for exactly the same information about the desire to desire A, for which we formerly asked with regard to A itself. But it is also apparent that the meaning of this second question cannot be correctly analyzed into "Is the desire to desire A one of the things which we desire to desire?": we have not before our minds anything so complicated as the question "Do we desire to desire to desire to desire A?" Moreover

> any one can easily convince himself by inspection that the predicate of this proposition—"good"—is positively different from notion of "desiring to desire" which enters into its subject: "That we should desire to desire A is good" is not merely equivalent to "That A should be good is good." It may indeed be true that what we desire to desire is always good; perhaps, even the converse may be true: but it is very doubtful whether this is the case, and the mere fact that we understand very well what is meant by doubting it, shews clearly that we have two different notions before our mind. (MOORE, 1903, p. 67-8).

De acordo com a interpretação que pretendo seguir, Moore considera que qualquer reivindicação de identidade entre propriedades deveria ser analítica. Assim, ele supõe que se o naturalista sustenta que propriedades morais são propriedades naturais, ele está dizendo que elas são analiticamente equivalentes. Isso significa o seguinte. Considere a definição naturalista para a propriedade moral *bondade* – por exemplo, *aquilo que maximiza o agregado de felicidade*[5]. Dada a suposta identidade entre o *definiens* e o *definiendum*, dever-se-ia poder substituir *bondade* por *aquilo que maximiza o agregado de felicidade*, e vice-versa, em qualquer discurso e manter intacto o significado daquilo que é dito. Como é possível notar na passagem acima, o AQA busca atacar justamente este ponto. A ideia é a seguinte: faz sentido (para falantes competentes) perguntar sobre qualquer propriedade supostamente definidora de *bondade*, se é bom? Por exemplo, faz sentido perguntar 'Este *x* maximiza o agregado de felicidade, mas é bom?'. Se o naturalismo é verdadeiro, então certamente não faz, pois, dado que há uma relação de analiticidade entre a propriedade moral e a propriedade natural, se estaria perguntando algo como 'Seria o *agregado de felicidade*, *agregado de felicidade*?' ou 'Seria bom, bom?'. Na terminologia de Moore, este é um tipo de *questão fechada*. Uma questão é fechada se a pergunta sincera implica que você não domina ou não entende o significado das palavras ou conceitos envolvidos na formulação ou que você comete alguma confusão conceitual (tal como perguntar, 'Este *x* é solteiro, mas é não-casado?'). No entanto, como sugeriu Moore, o problema para o naturalista é que essas questões não são fechadas, mas *questões abertas*. Quer dizer, para falantes competentes, sempre faz sentido perguntar sobre a propriedade definidora de *bondade* se é ou não bom. Disso, Moore concluiu que não há equivalência analítica entre a propriedade moral e a propriedade natural, pois o fato de a questão ser aberta revela que há algo a mais no significado da expressão (ou propriedade) moral que a expressão (ou propriedade) natural não capta. Então, deve haver diferenças entre propriedades morais e propriedades naturais. Moore sustentou que esse raciocínio poderia ser generalizado

[5] Estou atualizando o exemplo de Moore segundo o qual a propriedade de ser bom, isto é, a bondade, é aquilo que desejamos.

para toda e qualquer definição de propriedades morais em termos de propriedades naturais. Assim, concluiu, o naturalismo moral é falso.

Podemos tentar organizar seu argumento da forma canônica do seguinte modo[6]:

P1. Suponha-se que o predicado '*bom*' é sinônimo, ou analiticamente equivalente, ao predicado natural '*N*'.

C1. Então, é parte do significado da afirmação de que '*x* é *N*' que '*x* é *bom*'.

P2. Mas, então, alguém que se pergunte seriamente sobre um *x*, que é *N*, se é *bom*, estaria cometendo uma confusão conceitual.

P3. Dada qualquer propriedade natural *N*, é sempre uma *questão aberta* se um *x*, que é *N*, é ou não *bom*. Quer dizer, é sempre uma questão *significante* perguntar de um *x*, que é *N*, se é ou não *bom*.

C2. Assim, não pode ser o caso que '*bom*' seja sinônimo ou analiticamente equivalente com o predicado natural '*N*'.

C3. Portanto, a propriedade de *ser bom* não pode, como uma questão de necessidade conceitual, ser idêntica à propriedade de *ser N*.

Note que se trata de um teste semântico para extrair uma conclusão ontológica. Do fato de definições dos termos morais serem questões abertas, conclui-se que os predicados morais e suas definições naturalistas não são analiticamente equivalentes e que, portanto, as propriedades expressas por tais predicados morais são diferentes das propriedades naturais.

4. Influência

Qual a influência desse argumento na história da metaética? Não é incomum ver o desenvolvimento subsequente da metaética como algum tipo de reação ao argumento apresentado por Moore. Considere esta afirmação de Thomas Baldwin:

> [A] teoria ética inglesa do século XX é ininteligível sem referência ao *Principia Ethica*; em resumo, sua história até a década de 1960, mais ou menos, é de que,

[6] Embora faça leves alterações, adoto o modelo de A. Miller de apresentar o AQA. Cf. MILLER, A. *An Introduction to Contemporary Metaethics*. Cambridge: Polity Press, 2011, p. 13-14.

> embora Moore tenha sido tomado como tendo refutado o 'naturalismo ético', pensava-se que sua própria teoria não-naturalista fazia demandas metafísicas e epistemológicas inaceitáveis; assim, o único recurso era abandonar a crença numa realidade moral objetiva e aceitar uma abordagem emotivista, prescritivista ou, de outro modo, antirrealista dos valores éticos (BALDWIN, 1990, p. 66).

Podemos notar três grandes momentos da metaética depois do AQA, como sustenta Baldwin. O primeiro é a crença generalizada de que Moore realmente tinha refutado o naturalismo moral. O segundo é a insatisfação com o resultado positivo do AQA, isto é, o não-naturalismo metaético de Moore. E o terceiro, resultado desta insatisfação, é a ascensão do não-cognitivismo (ou expressivismo)[7].

No *Principia Ethica*, Moore defende uma forma de não-naturalismo metaético e sua principal evidência a favor desta teoria é negativa. Ou seja, ele argumenta a favor do não-naturalismo atacando o naturalismo com o AQA. No entanto, tal resultado positivo do AQA parecia implicar numa série de problemas. Grosso modo, o não-naturalismo é uma forma de realismo moral que sustenta que há fatos e propriedades morais e tais fatos e propriedades não são idênticos ou redutíveis a fatos e propriedades naturais. Propriedades morais são *autônomas*, ou *sui generis*. Elas não são expressão de atitudes emocionais dos sujeitos, mas são instanciadas pelos estados de coisas (pessoas, ações etc). Portanto, para o não-naturalista, os juízos morais são verdadeiros ou falsos e seu valor de verdade independe dos sujeitos[8]. Como Moore argumentou, 'bom' denota uma propriedade não-natural simples, não analisável e que não é parte da ordem causal (1993). No entanto, o problema é estabelecer claramente o que são exatamente essas propriedades *sui generis* e como temos acesso epistêmico a elas. O não-naturalismo de Moore tinha custos epistemológicos e metafísicos altos.

Um exemplo é o seguinte. Um fato moral que consiste na instanciação de uma propriedade moral e não faz parte da ordem causal não pode ser detectado pela percepção dos sentidos. Assim, que tipo de faculdade cognitiva nos dá acesso ao fato de que 'a tortura é errada'? Como captamos este tipo de fato moral e como esta capacidade intelectual funciona? Muitas vezes o não-naturalismo é referido como 'intuicionismo', pois os seus adeptos tentam

---

[7] No decorrer desse trabalho, usarei os termos 'não-cognitivismo' e 'expressivismo' de forma lata. Embora a corrente de teorias metaéticas expressivistas tenha origem e compartilhe teses centrais das teorias não-cognitivistas, há filósofos que preferem ser classificados unicamente como expressivistas, pois suas abordagens visam se afastar das formas clássicas de não-cognitivismo. Estou ciente desses usos mais estritos de 'não-cognitivismo' e 'expressivismo', mas, para os propósitos mais gerais do presente trabalho, usarei tais termos de forma intercambiável.

[8] Essa é uma caracterização bastante rudimentar da teoria. Há muitas variantes. Para um olhar mais aprofundado Cf. RIDGE, M. Moral Non-Naturalism. *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed). URL: https://plato.stanford.edu/archives/fall2019/entries/moral-non-naturalism. Acesso em 19/05/2021.

fornecer explicações a esses problemas postulando essa habilidade intelectual de captar a realidade moral como uma espécie de "intuição". Mas o problema persiste: o que é essa capacidade cognitiva e como ela funciona? Além disso, como explicar a natureza ontológica dessas propriedades não-naturais? Que tipo de "coisa" elas são? Essas eram algumas das insatisfações com a teoria de Moore. Mas uma preocupação em especial marcou o início de um novo período na metaética.

Qual é o ponto de se engajar numa controvérsia moral? Ao menos parte relevante parece ser a tentativa de que o resultado faça algum tipo de diferença no modo como as pessoas agem. Esperamos que as pessoas sintam pelo menos algum tipo de *motivação* para fazer o que elas pensam ser o correto a se fazer. Isso é de fundamental importância porque sugere que crenças morais são diferentes de outros tipos de crença. Crenças não morais não parecem ter esse aspecto motivacional inerente. Por exemplo, se alguém lhe informa de que o pub tal produz o melhor chopp da região, pode-se esperar que você vá até lá, mas isso depende de algo anterior: se você gosta ou não de chopp. Assim, quando alguém lhe convence sobre algum assunto de natureza não moral não se espera que você esteja motivado a agir de certo modo; isso depende de seus gostos, desejos e preferências. Por outro lado, se você está convencido de que o correto é doar para instituições de caridade, então espera-se que você doe para a caridade mesmo que não tenha nenhum desejo anterior. Além disso, se você não doar para a caridade é razoável pensar que não estava sendo totalmente sincero ao proclamar que é correto ou que temos o dever de doar para a caridade. Essa ideia de que as crenças morais têm algum tipo de conexão com a motivação é chamada de *internalismo motivacional* ou *internalismo moral*. Como afirmam Darwall, Gibbard e Railton (1992), o internalismo motivacional coloca um problema para a própria teoria de Moore, que se compromete com a tese de que juízos morais são crenças que representam o mundo.

Assim, já que o naturalismo moral supostamente é falso, dado o AQA, e já que a proposta concorrente de Moore parece ter mais custos do que benefícios, como explicar a natureza dos juízos morais? O caminho está no tipo de resposta que se tentou dar para ao porquê de o internalismo motivacional ser verdadeiro. Isto é, parece haver uma conexão entre endossar juízos morais e agir de acordo com esses juízos morais, mas *por que* isso é o caso? M. Schroeder nos diz que a resposta mais influente para esta questão

> [...] têm sido que crenças morais tem uma conexão especial com a motivação que as crenças não-morais não têm porque crenças morais são um *tipo de estado mental diferente* das crenças não-moais. Os proponentes desta resposta costumam combiná-la com uma representação de *como* esses dois tipos de estados mentais diferem. De

> acordo com essa abordagem, crenças não-morais são *sobre* algo. Elas são como mapas que nos informam sobre a disposição da terra, em que 'a terra' é uma metáfora para sobre o que elas são. Enquanto que crenças morais, de acordo com essa abordagem, não são mapas sobre nada. Elas se parecem mais com *objetivos* que temos, objetivos sobre qual destino do mapa queremos alcançar (SCHROEDER, 2010, p. 11).

A metáfora do mapa sugerida por Schroeder traz uma nova forma de explicar a natureza dos juízos morais. De acordo com essa abordagem, enunciados morais não descrevem ou representam uma realidade, tal como os naturalistas e os não-naturalistas sustentam. Não há fatos morais, pois os predicados morais não expressam propriedades. Enquanto juízos descritivos tal como 'há um ônibus em frente à universidade' representam fatos, juízos morais tem uma *função semântica* distinta; eles expressam algum tipo de *atitude* do indivíduo sobre o mundo. Dado que a tese central dessa visão é a ideia de que a função semântica dos juízos morais é expressar um estado mental conativo do indivíduo, e não descrever a realidade, temos o que ficou conhecido como o *não-descritivismo*[9].

Mas o que esse tipo de teoria metaética tem a ver com o internalismo motivacional? Ora, enquanto o internalismo motivacional é um custo para o não-naturalismo, o não-descritivismo, fornece uma explicação bem plausível para tal tese. Os não-descritivistas são adeptos da *Teoria Humeana da Motivação*. De acordo com essa teoria, as crenças, por si só, são insuficientes para motivar um indivíduo à ação; é necessário um componente a mais, *desejos*. A ideia é bastante intuitiva. A aceitação de que algo é o caso (por exemplo, que há uma torta de morango na geladeira), por si só, não é suficiente para levar um indivíduo a algum tipo de ação a respeito deste algo; é necessário um elemento a mais para explicar a ação, um desejo (por exemplo, querer comer a torta). Esta conjunção entre o desejo de comer a torta e a crença de que há uma torta na geladeira fornecem uma explicação para a motivação do indivíduo. Sendo assim, ao equiparar os juízos morais com estados mentais conativos (desejos, preferências, atitudes etc), os não-descritivistas estão numa boa posição para acomodar o internalismo motivacional.

As teorias não-descritivistas tiveram uma grande influência depois da publicação do *Principia*. Os primeiros autores a apresentar uma versão, ainda que rudimentar, de tal teoria são Charles Kay Ogden e Ivor Armstrong Richards (1923). Eles desempenharam um papel importante já que vieram a influenciar o que é provavelmente a defesa mais conhecida de uma forma de não-descritivismo, o emotivismo moral por Ayer (1936). Numa das passagens mais citadas nos manuais de metaética, Ayer apresenta sua teoria emotivista do seguinte modo:

[9] Essas teorias também são denominadas 'não-cognitivismo' ou 'expressivismo'.

> A presença de um símbolo ético numa proposição não adiciona nada ao seu conteúdo factual. Portanto, se digo a alguém 'Você agiu incorretamente ao roubar aquele dinheiro', não estou afirmando nada além do que se simplesmente dissesse 'Você roubou aquele dinheiro'. Ao adicionar que esse ato é errado, não estou fazendo nenhuma afirmação para além disso. Estou simplesmente demonstrando minha *desaprovação moral*. É como se tivesse dito 'Você roubou aquele dinheiro' *num tom de horror peculiar* ou escrito com a adição de alguma marca especial de exclamação. O tom ou a marca de exclamação não adicionam nada ao significado literal da sentença. Meramente serve para mostrar que a expressão da sentença está acompanhada de *certos sentimentos no falante* (AYER, 1936, p. 107. Itálico meu).

Em passagens como esta, Ayer está contrastando o significado de termos morais com o significado de termos não morais e atribuindo um papel distintivo aos primeiros. Ao contrário dos naturalistas e não-naturalistas, um termo moral não possui conteúdo factual ou descritivo, mas um significado expressivo. Assim, não é necessário fazer alusão ao conteúdo descritivo para sabermos o significado dos termos morais. O conteúdo emotivo é central. Ayer defendeu que tal conteúdo emotivo é uma atitude de aprovação ou desaprovação que o sujeito expressa sobre ações ou pessoas[10].

Logo depois, temos uma defesa mais robusta do não-descritivismo emotivista com Stevenson (1937). Ao contrastar sua teoria com o que chama de 'teorias tradicionais do interesse', Stevenson apresenta sua visão do seguinte modo.

> As teorias tradicionais do interesse mantem que os juízos morais são descrições de estados e interesses existentes – que eles simplesmente dão informação sobre interesses. [...] Sem dúvida, *há sempre algum elemento de descrição nos juízos morais*, mas isso não é tudo. Seu *principal uso* não é indicar fatos, mas criar uma *influência*. Ao invés de apenas descrever os interesses das pessoas, eles *mudam* ou os *intensificam*. Os juízos morais *recomendam* um interesse num objeto ao invés de afirmar um interesse já existente (STEVENSON, 1937, p. 16. Itálico meu).

Como podemos observar, ao contrário de Ayer, Stevenson não acreditava que o conteúdo descritivo é dispensável para entendermos o significado dos termos morais. Como ele diz, "há sempre algum elemento de descrição nos juízos éticos". O ponto central de sua tese é que entender apenas o significado descritivo não é suficiente para explicar o significado dos juízos morais, pois descrever não é seu "principal uso". Stevenson se diferencia de Ayer, também, na explicação fornecida para o estado mental não-cognitivo envolvido na realização de um juízo moral. Ele diz que uma sentença tal como '*x* é bom' significa algo como 'eu aprovo *x*, faça-o

[10] É interessante notar que Ayer empregava o AQA como evidência a favor do não-descritivismo, e não do não-naturalismo, como Moore (Cf. AYER, 1971, p. 136-159).

também'. Além de fornecer descrição sobre o estado psicológico de quem enuncia tal sentença, '*x* é bom' é usada para "recomendar", encorajar, "criar uma influência" no ouvinte a aprovar *x* também. Assim, o juízo moral tem um componente *descritivo* que informa sobre a emoção do indivíduo que o expressa e um componente *imperativo*, que objetiva evocar no ouvinte o mesmo sentimento ou emoção.

Mas, se juízos morais tratam apenas de expressão de atitudes, sentimentos de aprovação e desaprovação e tentativa de evocar o mesmo sentimento no ouvinte, parece que não há uma forma objetiva de resolver controvérsias morais. Se S_1 sustenta que o assassinato é errado e S_2 sustenta que não é, então parece que a única coisa que a ética tem a nos dizer sobre isso é que S_1 tem o sentimento de desaprovação em relação ao assassinado e pretende evocar tal sentimento em S_2 e vice-versa. Mas a moralidade não parece ser meramente uma área em que as pessoas expressam seus sentimentos. Elas pensam que é realmente verdadeiro o que defendem. Além disso, na época em que Ayer e Stevenson apresentavam suas teorias metaéticas, o Positivismo Lógico e o seu *Princípio de Verificabilidade* – que dizia que uma sentença tem significado se, e somente se, é ou analítica ou empiricamente verificável (AYER, 1971, p. 7) – eram amplamente aceitos. Portanto, como as sentenças morais não são nem analíticas e nem empiricamente verificáveis, eram excluídas do domínio das sentenças com sentido, algo que, também, parece inaceitável.

Assim, o próximo grande passo do não-descritivismo foi dado por Richard Hare (1952). Sua estratégia consistia em mostrar que a moralidade é uma atividade racional e não meramente um conjunto de sentenças sem sentido. Sua teoria preserva os traços da tradição não-descritivista, mas se diferencia de seus antecessores, também, no que diz respeito à caracterização positiva do estado mental envolvido num juízo moral. Hare sustenta que os juízos morais são uma espécie de juízo *prescritivo* e que as sentenças morais são melhor entendidas como expressando *imperativos*. Ele tentou mostrar que as sentenças imperativas, assim como as descritivas, são regidas por regras lógicas (tal como, não contradição, disjunção, conjunção, negação), ao contrário do que se pensava até então (HARE, 1952, p. 1-16). Além disso, Hare argumentou que as prescrições morais são um tipo distintivo de prescrição, isto é, elas têm uma característica que ele chamou de *universalizabilidade* (HARE, 1952, p.137-150). Isso significa que se um indivíduo aceita um juízo moral sobre uma situação particular, então ele deve aceitar, sob pena de inconsistência, que o mesmo juízo se aplica a situações relevantemente similares. De acordo com Hare, se aceitássemos que os imperativos, ao contrário do que sustentavam os verificacionistas, tem sentido e são regidos por regras lógicas

e que juízos morais expressam prescrições que são universais, seria possível conferir racionalidade à atividade moral.

Assim como Ayer e Stevenson, Hare foi amplamente influenciado pelo AQA de Moore e, inclusive, apresentou uma reformulação do AQA (HARE, 1952, p. 79-93). Ao defender o Prescritivismo Universal, Hare apresentou outro tipo de argumento, hoje é bastante famoso na metaética, contra as teorias descritivistas. Irei chamar tal argumento de *Argumento do Desacordo*. É um tipo de raciocínio que parte premissas sobre nossas intuições sobre o desacordo para extrair conclusões sobre semântica. Considerar este argumento aqui é tão importante quanto o AQA, pois o tipo de argumento que pretendo analisar no decorrer deste trabalho possui grande débito para com Hare.

O argumento é baseado no seguinte experimento mental. Hare nos pede para imaginar uma situação na qual um missionário desembarca numa ilha habitada por canibais e percebe que, a despeito de possuírem línguas totalmente diferentes, assim como ele, os canibais usam o termo 'bom' como um termo geral de *aprovação* e *recomendação* para pessoas e ações. No entanto, ele percebe que os canibais aplicam o termo 'bom' a classes diferentes de ações e, portanto, se surpreendem ao descobrir como ele, o missionário, usa o termo, pois

> Eles sabem que quando o missionário usa a palavra ele está *recomendando* a pessoa ou objeto ao qual ele a aplica. A única coisa que eles acharão estranha é que ele a aplica a pessoas inesperadas, pessoas que são dóceis e gentis e que não colecionam grandes quantidades de escalpos, ao passo que eles [canibais] estão acostumados a recomendar pessoas ousadas e belicosas e que colecionam mais escalpos do que a média (HARE, 1952, p. 148).

Além disso, Hare supõe que o missionário e os canibais se engajam numa discussão moral. Podemos imaginar que o missionário sustenta algo como:

(i) Coletar escalpos não é bom.

Por outro lado, os canibais sustentam que:

(i) Coletar escalpos é bom.

A maioria das pessoas diria que há um desacordo genuíno aqui. Assim, a partir desta intuição, Hare defende que, embora o missionário e os canibais apliquem o termo 'bom' a coisas diferentes, há algo no significado deste termo que é convergente.

Mas que tipo de implicação isso tem para o naturalismo moral? Hare nos pede para considerarmos uma teoria metaética cujo significado dos termos morais é captado exaustivamente por seu conteúdo descritivo. Segundo tal teoria, quando dizemos que algo é bom, não estamos fazendo nada além do que afirmando que tal coisa tem tal e tal propriedade descritiva. Tal teoria metaética seria algum tipo de naturalismo moral, tal como Hare o compreende (HARE, 1952, p. 80-93). Agora, se o naturalismo é verdadeiro,

> Então quando o missionário dissesse que pessoas que não colecionavam escalpos eram boas (em sua língua), e os canibais dissessem que pessoas que colecionavam muitos escalpos eram boas (em sua língua), eles não estariam discordando[...]. (HARE, 1952, p. 149).

Isso porque 'bom', tal como usado pelo missionário, predica, entre outras coisas, 'não cometer assassinatos' e 'bom', tal como usado pelos canibais, predica, entre outras coisas, 'coletar o maior número de escalpos possível'. Sendo assim, quando o missionário enuncia (i) e os canibais enunciam (ii) eles significam coisas bem diferentes com o uso de 'bom' e, se este é o caso, então as afirmações não são incompatíveis. Isto é, pode ser verdadeiro, de acordo com o conteúdo descritivo de 'bom' para o missionário, que coletar escalpos não é bom. E pode, também, ser verdadeiro, de acordo com o conteúdo descritivo de 'bom' para os canibais, que coletar escalpos é bom. Portanto, o naturalismo moral nos constrange à conclusão de que não há desacordo genuíno entre o missionário e os canibais.

Mas, certamente, a maioria de nós tem a intuição de que há um desacordo moral genuíno aqui, sustenta Hare. Quer dizer, não parece que o missionário e os canibais estão falando sobre coisas totalmente diferentes. Pelo contrário, é muito plausível assumir que o conteúdo do que é afirmado pelo missionário nega o conteúdo daquilo que é afirmado pelos canibais, e vice-versa. Segundo Hare, a lição que devemos tirar disso é que, embora as pessoas possam atribuir significado descritivo diferente aos termos morais, o fato de haver desacordos morais genuínos nesses cenários hipotéticos mostra que há um significado *primário* que é central aos termos morais e que é preservado, mesmo em contextos em que o significado descritivo varia. Ele argumenta que isso é evidência a favor do não-cognitivismo.

Temos um forte argumento contra o naturalismo moral aqui. O que Hare está dizendo é que se aceitarmos o descritivismo (naturalismo) teríamos que aceitar algo que não parece ser verdadeiro, quer dizer, que não haveria desacordo moral num cenário em que duas comunidades atribuem extensões diferentes aos termos morais, pois se estaria predicando coisas diferentes com termos ortograficamente similares. A nossa intuição é de que o desacordo é preservado

mesmo que os falantes usem termos morais com significado extensional distinto. Aqui é importante notar que Hare faz uso de uma condição implícita que é condição necessária para haver desacordo entre falantes: a *Condição da Analiticidade da Extensão*. Veremos esta condição de forma mais rigorosa adiante, mas, grosso modo, ela diz que, para haver desacordo genuíno entre dois sujeitos, a extensão dos termos empregados por eles em suas afirmações deve possuir a mesma extensão. Disso se segue que, se há desacordo então essas duas comunidades, mesmo que atribuam significado descritivo diferente aos termos morais, devem preservar algo no significado que torna o desacordo possível. Se isso é o caso, então há algo a mais no significado dos termos morais para além de meramente descrever. E como o naturalismo moral sustenta que a função dos juízos morais é apenas descrever estados de coisas, deve ser falso.

Podemos ver, neste sentido a ampla influência do argumento de Moore no desenvolvimento da metaética nas décadas seguintes. A aceitação do AQA bem como a insatisfação com o não-naturalismo, fizeram surgir algumas variantes do não-cognitivismo. No entanto, a descoberta de um problema na década de 60 trouxe sérios impactos este tipo de teoria, que era dominante até então, e começou a preparar o terreno para o ressurgimento do naturalismo moral. Trata-se do *Problema Frege-Geach*.

5. O Problema Frege-Geach

Na década de 60, Peter Geach (1965) apresentou um problema ao não-descritivismo que veio a representar um ponto de virada na metaética. É o que ficou conhecido como *Problema Frege-Geach*. Consideremos, brevemente, tal problema[11].

É amplamente aceito na filosofia da linguagem o que se costuma chamar de *princípio da composicionalidade*. Este princípio diz que o significado de uma sentença complexa é o resultado do significado de suas partes. O princípio é muito plausível, pois explica porque temos a capacidade de entender um número potencialmente infinito de novas sentenças apenas com o domínio de um número finito de palavras e princípios gramaticais de combinação dessas palavras.

Suponha que o não-descritivismo é verdadeiro e que as seguintes sentenças têm significado não-cognitivo:

[11] Há várias formas de apresentar o problema. Para uma forma mais rigorosa e completa, sugiro SCHROEDER, 2010, p. 41-47.

(i) Torturar o gato é errado.
(ii) Influenciar o irmão pequeno a torturar o gato é errado.

Agora, considere uma sentença que combina as partes das sentenças (i) e (ii):

(iii) Se torturar o gato é errado, então influenciar o irmão pequeno a torturar o gato é errado.

Esta sentença, obviamente, é diferente de (i) e (ii), pois trata-se de um *condicional*. E, dado o *princípio da composicionalidade*, "torturar o gato é errado" deve ter o mesmo conteúdo semântico tanto em (i) quando em (iii). Tanto é assim, que o seguinte argumento é válido e expressa um modo bastante comum de raciocínio moral.

(i) Torturar o gato é errado.
(ii) Se torturar o gato é errado, então influenciar o irmão pequeno a torturar o gato é errado.
(iii) Portanto, influenciar o irmão pequeno a torturar o gato é errado.

No entanto, o não-cognitivista não pode aceitar isto. Não parece que o conteúdo da sentença (i) (conteúdo emotivo) é o mesmo que o conteúdo que essa sentença contribui para o significado da sentença complexa (ii). Pois, não é necessário assentir ao antecedente de um condicional para afirmar o seu consequente. Isto é, não é necessário que S_1 tenha um sentimento de desaprovação a respeito de torturar gatos para poder endossar a sentença condicional (ii). Portanto, o significado de "torturar o gato é errado" em (i) e (ii) deve ser diferente, se assumirmos a verdade do não-cognitivismo. E isso implica que alguém que faça o raciocínio acima e extraia a conclusão (iii) comete um erro no percurso do argumento (falácia do equívoco). Mas claramente não é o que acontece. O raciocínio acima parece estar em perfeita ordem. Assim, o problema para o não-descritivismo é o de explicar como sentenças morais funcionam em contextos complexos (tais como condicionais) pressupondo que tais sentenças tem um significado não-cognitivo.

O resultado do *Problema Frege-Geach*, como afirma Sinclair (SINCLAIR, 2019, p. 27), foi que "as décadas de 1970, 80 e 90 viram novas formas de realismo moral naturalista sendo desenvolvidas".

6. A (Re) Ascenção do Naturalismo Moral

Podemos dizer que o mais importante movimento de reestruturação e de uma nova defesa do naturalismo moral veio de três filósofos da Universidade de Cornell (Sturgeon (1985), Boyd (1988) e Brink (1986, 1989)). Aplicando os recentes desenvolvimentos de outras áreas da filosofia, como epistemologia, filosofia da linguagem, metafísica e filosofia da mente, esses filósofos conferiram ao naturalismo moral um status de respeitabilidade teórica que fez ressurgir o interesse por esta teoria. Esses filósofos desenvolveram o tipo de naturalismo mais influentes da metaética contemporânea: o que ficou conhecido como *Realismo de Cornell*.

Essa teoria marca uma distinção importante entre *naturalismo não-reducionista* e *naturalismo reducionista*. De acordo com Darwall, Gibbard e Railton (1992), podemos compreender a distinção do seguinte modo. Os reducionistas sustentam ou que o vocabulário moral é analisável em termos do vocabulário não moral ou que as propriedades morais são idênticas às propriedades naturais como uma questão de fato sintético (ou ambos). Os não-reducionistas sustentam que não há nenhum tipo de relação analítica entre termos ou propriedades morais e naturais, mas de *constituição* ou *realização múltipla*. Assim, o não-reducionismo conta como uma visão naturalista mesmo que não haja uma redução óbvia, direta e imediata do moral ao natural (Miller, 2013). Por vezes, essa distinção também é colocada em termos de *naturalismo analítico* (reducionista) e *naturalismo não-analítico* ou *sintético* (não-reducionista).

Nesta seção, buscarei apresentar uma versão resumida do Realismo de Cornell no intuito de mostrar em maiores detalhes como o naturalismo moral visa explicar a natureza das propriedades morais. Além disso, dado que qualquer teoria naturalista deve fornecer um diagnóstico a respeito do AQA, irei mostrar como essa nova forma de naturalismo lidam com tal problema.

6.1. Realismo de Cornell

Obviamente, há uma unidade teórica que constitui núcleo do Realismo de Cornell. No entanto, alguns de seus defensores investiram mais tempo em algumas ideias do que em outras. Por isso, a melhor forma de acessarmos essa teoria é organizando-a em dois momentos: a tese de que as propriedades morais possuem eficácia explanatória (Sturgeon e Brink) e a teoria metassemântica de Boyd. Não devemos ver tal separação de ideias como desenvolvimentos distintos, mas como complementares. Além disso, irei denominar o tipo de teoria defendido por esses filósofos de *Realismo Moral Naturalista* (doravante, RMN). Comecemos, então, com a discussão sobre a eficácia explanatória das propriedades morais.

6.1.1. Explicações Morais

Como apontam Darwall, Gibbard e Railton (1992), os realistas de Cornell recorrem a um modelo argumentativo de inferência à melhor explicação fazendo uso de analogias com exemplos das ciências. Há vários tipos naturais – químicos ou biológicos – que não são obviamente redutíveis aos tipos naturais da física, mas, dado que desempenham um papel explanatório relevante, são considerados tipos naturais. Assim, a estratégia é argumentar que, embora as propriedades morais não tenham uma redução óbvia às propriedades naturais, como elas possuem eficácia explanatória, elas devem fazer parte da constituição do mundo natural. A ideia é que estamos justificados em postular uma entidade (fato, propriedade) desde que tal entidade seja requerida para a melhor explicação de algum fenômeno da realidade. Há uma discussão clássica entre Gilbert Harman (1977, 1986), Nicholas Sturgeon (1986a, 1986b, 1988) e David Brink (1989) a respeito dessa estratégia.

Em *The Nature of Morality* (1977), Harman apresentou a seguinte objeção ao naturalismo moral. Ele pergunta:

> Você pode observar alguém fazer algo, mas você pode alguma vez perceber a *correção* ou *incorreção* do que ele faz? Se você virar uma esquina e observar um grupo de baderneiros colocarem gasolina em um gato e incendiá-lo, você não precisa *concluir* que o que eles estão fazendo é errado; você não precisa *descobrir* nada exterior; você pode *ver* que isso é errado. Mas é a sua reação evidência à atual incorreção do que você vê ou é simplesmente a reflexão do seu "senso" moral, um "senso" que você adquiriu talvez como resultado da sua educação moral? (HARMAN, 1977, p. 4).

E sua resposta é a seguinte:

> [a] observação desempenha um papel na ciência que não parece desempenhar na ética. A diferença é que você precisa fazer suposições sobre certos fatos físicos para explicar a ocorrência das observações que suportam uma teoria científica, mas você não parece precisar fazer suposições sobre nenhum fato moral para explicar a ocorrência das

> chamadas observações morais que tenho falado. No caso moral, parece que você precisa fazer suposições apenas sobre a *psicologia* ou *sensibilidade moral* da pessoa que faz a observação moral. No caso científico, a teoria é testada contra o mundo (HARMAN, 1977, p. 6).

Em outras palavras, Harman está sustentando que não estamos justificados em postular a existência de fatos e propriedades morais como constituição natural do mundo porque tais fatos e propriedades não possuem a eficácia explanatória que os fatos e propriedades da física, por exemplo. Estamos justificados a postular a existência de coisas tais como prótons, mesmo que essas entidades não sejam visíveis, pois sua existência fornece explicações relevantes para a experiência. No entanto, para explicarmos porque é errado incendiar um gato cruelmente não precisamos recorrer à propriedades morais, como a *incorreção*; nossa sensibilidade moral parece desempenhar esse papel. Assim argumenta Harman.

Em *Moral Explanations* (1985), Sturgeon responde ao desafio de Harman. Ele propõe o seguinte *teste contrafactual* para determinar a relevância explanatória de determinado fato: se uma suposição particular é completamente irrelevante para a explicação de certo fato, então o fato seria obtido, e o poderíamos ter explicado do mesmo modo, mesmo se tal suposição particular estivesse ausente (STURGEON, 1985, p. 65). Harman sustenta que a *incorreção* é "completamente irrelevante" para a explicação da crença de que aquilo que os jovens estão fazendo é errado. Mas se esta ação não tivesse a propriedade da *incorreção*, você, ainda assim, teria a crença de que é errada? O teste contrafactual nos pede para imaginar um mundo possível em que os jovens estão colocando gasolina em um gato e incendiando-o, mas que essa ação não tem a propriedade da *incorreção*. A estratégia de Sturgeon consiste em argumentar que é impossível imaginarmos tal mundo. Ele apela para a tese da superveniência que, grosso modo, diz o seguinte. Não pode haver diferenças morais entre duas ações possíveis sem que haja alguma diferença natural entre elas. A ideia é que se a avaliação moral entre a ação A e B é diferente, então, necessariamente, deve haver diferença entre as características naturais que constituem tais ações. Disso se segue que há algum tipo de relação importante entre propriedades morais e propriedades naturais. Considere, por exemplo, uma ação incorreta. Algumas características relevantes do que fazem uma ação ser correta podem ser causar dor, ser cruel com alguém, infringir algum tipo de sofrimento psicológico etc. A ideia é que a *incorreção* sobrevém a essas características naturais. Dado que a *incorreção* é superveniente às propriedades naturais, então imaginar um mundo possível que os jovens estão fazendo algo que não tenha a propriedade da *incorreção* é imaginar um mundo possível em que eles não estão causando dor deliberadamente ao animal, não estão fazendo-o sofrer etc. Mas imaginar tal

mundo não é imaginar um mundo em que eles estão colocando gasolina no animal e incendiando-o. Portanto, argumenta Sturgeon, num mundo possível em que os jovens estão colocando gasolina num gato e incendiando-o, eles estão fazendo algo que tem a propriedade da *incorreção*. Por fim, Sturgeon sustenta que propriedade moral, ao menos em parte, desempenha um papel explanatório nas crenças morais e que estas não são meramente fruto da "sensibilidade moral"[12].

Harman ainda responde Sturgeon em *Moral Explanations and Natural Facts* (1986) e essa controvérsia dá início a uma sofisticada discussão sobre a eficácia explanatória das propriedades morais. Há várias alternativas à proposta de Sturgeon. A ideia de *programa explanatório*, desenvolvida por Jackson e Pettit (1990), por exemplo, é uma refinada teoria a favor do poder explanatório das propriedades morais. A estratégia argumentativa é similar: se as propriedades morais explicam algum fenômeno da realidade (crença moral, por exemplo), então estamos justificados em postular sua existência como parte da constituição do mundo natural. E se isso for o caso, temos razões para aceitar a plausibilidade do naturalismo moral.

Esse modelo argumentativo, embora seja bastante sofisticado e tenha colocado o naturalismo moral numa posição de respeitabilidade como teoria metaética, deixa uma série de perguntas sem respostas. Por exemplo, dado que, depois de Moore, qualquer teoria naturalista deve fornecer um diagnóstico do AQA, por que e como esse tipo de naturalismo evita o AQA? Ou, por que não é vulnerável ao *Argumento do Desacordo* de Hare? Como *sabemos* quais ações possuem quais propriedades morais? Como a *referência* de termos morais como "bom" e "correto" é determinada? Sendo assim, ainda há muito trabalho a ser feito. Por isso, sempre devemos considerar a teoria metassemântica de Boyd em conjunto com as ideias de Sturgeon e Brink. Em *How to be a Moral Realist* (1988), Boyd desenvolveu uma teoria metassemântica externalista para os predicados morais e tentou responder detalhadamente aos principais

[12] De acordo com os Realistas de Cornell, esse tipo de raciocínio se aplica a inúmeras instâncias. Considere esta passagem de D. Brink:

> […] explicamos boa parte do comportamento de Calígula, tal como a execução de pessoas inocentes, através da sua *crueldade*, e o fato de ele ter este vício de caráter explica, em grande parte, nossa opinião sobre ele e suas ações. De acordo com relatos populares, é o fato de que Lincoln era um homem *equitativo* e *justo* que explica a sua oposição à escravidão […]. Além disso, frequentemente explicamos fatos e eventos sociais, assim como ações individuais, apelando aos fatos morais sobre instituições legais e sociais. Acreditamos que vícios políticos (e. g., *injustiça social*) às vezes causam, e assim ajudam a explicar, instabilidade, movimentos de protesto e revoluções; e achamos que virtudes políticas das instituições e leis de uma sociedade (e. g., *a justiça social*) podem ajudar a explicar a sua estabilidade. Neste sentido, podemos citar a *injustiça* da escravidão como parte da explicação do seu desaparecimento ou a *injustiça* do sistema político e legal da África do Sul como causa da sua instabilidade e do protesto contra o apartheid (BRINK, 1989, p. 87).

problemas que uma teoria naturalista deve enfrentar. O resultado foi que o Realismo de Cornell se tornou uma das mais completas e influentes defesas do naturalismo moral não-reducionista.

6.1.2. A Teoria Metassemântica de Boyd

Como afirmei acima, um dos primeiros problemas que deve ocupar o proponente de uma teoria naturalista é fornecer um diagnóstico sobre o AQA. Para Boyd esse diagnóstico se fez possível através de uma nova forma de pensar sobre a *referência* linguística. Nas décadas de 1960 e 1970, filósofos como Kripke (1980) e Putnam (1975) deram início ao desenvolvimento das teorias externalistas da referência. De acordo com essas teorias, os termos referem o objeto que representam *diretamente*, quer dizer, sem a necessidade de o referente ter de satisfazer um conjunto de descrições, tal como as teorias descritivistas amplamente aceitas até então. Boyd aplicou essa forma de pensar sobre a referência para os termos morais.

Segundo a teoria descritivista da referência, um termo fixa o referente *x* se este referente satisfaz todos os sentidos, ou conjunto de descrições, possíveis que são associados ao termo. Por outro lado, de acordo com a teoria externalista da referência, os termos referem o que referem *diretamente*, quer dizer, de forma não mediada por qualquer coisa a mais representada como parte da sentença ou conteúdo do pensamento. A teoria da referência direta exclui a ideia de que há um sentido descritivo como parte do significado dos termos. O termo capta apenas o referente, e nada além disso tal como um conjunto de sentidos diferentes.

Há uma série de argumentos que os externalistas forneceram contra o descritivismo[13]. Um desses argumentos, dada a sua importância para o contexto deste trabalho, vale a pena considerar brevemente aqui. Trata-se do *Argumento da Terra Gêmea*, de Putnam. Em *The Meaning of 'Meaning'* (1975), Putnam ofereceu um experimento de pensamento com o objetivo de argumentar que o significado dos termos não é puramente psicológico ou determinado exclusivamente pelo conjunto de descrições a ele associadsas, mas depende do mundo externo. Ele nos pede para imaginar um mundo possível (Terra Gêmea (TG)) que é muito similar ao nosso mundo atual (Terra (T)). Exceto por uma peculiaridade, TG é exatamente similar a T. Tal peculiaridade é que a composição química do líquido que os habitantes de TG chamam 'água' não é H_2O, como em T, mas uma composição completamente diferente, denominada XYZ. O estado superficial da água é o mesmo em T e TG. Em TG as pessoas falam português e até mesmo o termo usado para denominar esse líquido é 'água'. A questão agora é: habitantes de

[13] Para um resumo desses argumentos, veja Lycan (2000), capítulos 3 e 4.

T e TG se referem ao mesmo objeto quando usam o termo ‘água’? Putnam argumenta que não. Embora associem o mesmo conjunto de descrições a termo ‘água’ (incolor, que serve para matar a sede...etc), a extensão de objetos que o termo capta é totalmente diferente.

O descritivista teria que aceitar a conclusão de que ‘água-*t*’ e ‘água-*tg*’ possuem a mesma extensão, pois tanto os habitantes de T quanto os habitantes de TG possuem as mesmas intensões sobre ‘água’ (isto é, aplicam o mesmo conjunto de descrições definidas). Mas isso é falso, como argumenta Putnam, pois um indivíduo A e um indivíduo B podem ter um mesmo conjunto de descrições definidas sobre um termo F e A se referir a um objeto e B a outro objeto distinto. Portanto, conclui Putnam, “o significado não está na cabeça” (PUTNAM, 1975, p. 277). Quer dizer, não depende totalmente do modo como pensamos sobre determinada entidade e das descrições que associamos a ela. Este argumento mostra que algo externo ao falante é necessário para a fixação do significado, algo que não depende da mente do falante e do conjunto de propriedades que ele associa ao predicado. O resultado do *Argumento da Terra Gêmea* é que o descritivismo semântico deve ser falso e que, obviamente, somente uma teoria externalista do significado é adequada.

O AQA e o *Argumento do Desacordo* de Hare tinham como pressuposto a teoria descritivista da referência. O AQA sustenta que propriedades morais não podem ser propriedades naturais porque, para falantes competentes, é sempre uma questão aberta se N (propriedade natural) é M (propriedade moral). Se é uma questão aberta, isso significa que os falantes associam diferentes descrições ou pensamentos a N e M, caso contrário essas propriedades teriam a mesma extensão. No *Argumento do Desacordo*, embora Hare conclua que há comunalidade semântica entre os termos morais dos canibais e do missionário (senão o desacordo genuíno não seria possível), ele estipula que os termos morais de tais comunidades têm significado descritivo diferente porque seus falantes aplicam esses termos a estados de coisas distintos. Isto é, ele supõe que a extensão dos termos morais depende do que as pessoas acreditam sobre sua extensão.

Com o externalismo semântico em mãos, o naturalista tem agora uma ferramenta importante para se desviar dos ataques de Moore e Hare. Ele pode simplesmente argumentar que a diferença no modo como os falantes aplicam determinado predicado não reflete em diferenças na extensão desses predicados. Lembre que uma das premissas fundamentais do AQA é que qualquer reivindicação de identidade entre propriedades deve ser analítica. Mas com o desenvolvimento dessas teorias externalistas, os filósofos reconheceram a possibilidade de relações sintéticas de identidade. Considere o exemplo clássico sobre “água” e ‘H_2O’. A

despeito desses termos terem significado diferente na linguagem ordinária, hoje sabemos que 'água' é realmente 'H_2O'. Note que se o AQA é cogente, teríamos que aceitar que 'água' não é 'H_2O' porque a questão '*x* é água, mas é x H_2O?' é aberta, mesmo para falantes competentes (suponha, antes da descoberta de que água é H_2O). O que essa nova teoria semântica mostrou é que dois termos ou propriedades podem ter uma relação de identidade mesmo que não sejam analiticamente equivalentes. Neste sentido, o primeiro grande passo para uma reabilitação do naturalismo moral estava dado: o AQA não era mais um obstáculo. Faltava ainda uma teoria positiva. Esse segundo grande passo foi dado por Boyd.

Boyd aplicou a teoria externalista da referência dos nomes próprios e tipos naturais para a moralidade e desenvolveu a teoria externalista da *regulação causal*[14]. Ele argumenta que

> […] alguns termos têm definições que são determinadas por uma coleção de propriedades tal que a posse de um número adequado dessas propriedades é suficiente para recair sob a extensão do termo. […] nossos conceitos de tais tipos são "abertos" e há uma indeterminação na extensão *legitimamente* associada com as definições sobre o agregado de propriedades ou atribuição de critérios. A "imprecisão" ou "vagueza" de tais definições é vista como uma característica perfeitamente apropriada do uso linguístico ordinário […] (BOYD, 1988, p. 322).

Neste sentido, um termo depende, não de *uma* propriedade que é a sua extensão, mas de um *conjunto* de propriedades que *constituem* a sua extensão e, por conseguinte, formam a sua definição. Essas propriedades são unificadas por relações nomológicas que determinam que tais e tais indivíduos fazem parte do agregado. As relações nomológicas são descobertas através de investigação científica. Portanto, não temos uma definição precisa de quais propriedades constituem o agregado. Através das descobertas empíricas o agregado vai se constituindo e assim formamos a nossa definição natural. Esse tipo de definição não apresenta condições suficientes para a extensão de um termo, pois sempre podemos descobrir novas propriedades que são unificadas ao agregado. Boyd denomina essas definições naturais de *agregado de propriedades homeostáticas*.

Considere o seguinte exemplo. Usamos o termo 'metal' para expressar uma variedade de propriedades. Portanto, há um agregado de propriedades que são características dos indivíduos que compõe a extensão deste termo, tais como condutibilidade, ductilidade, maleabilidade, brilho, etc. Frequentemente, dizemos que algo é metal quando apresenta essas propriedades. Além disso, essas propriedades são co-instanciadas por diferentes indivíduos. A

[14] Várias passagens do restante desta seção são amplamente baseadas em *Realismo, Naturalismo e Semãntica Moral* (Cf. KAVETSKI, 2017, p. 96-112).

prata, o ouro, o alumínio, o ferro, o chumbo, etc., são instâncias de ‘metal’. Neste sentido, a definição de ‘metal’ não é precisa ou *exaustiva*, pois sempre pode acontecer de descobrirmos um novo material que apresente a propriedade compartilhada pelos indivíduos de caráter metálico e, portanto, fazer parte do aglomerado de indivíduos que constituem a definição de ‘metal’.

Assim, Boyd argumenta sobre a definição de um tipo natural, que há um grupo de propriedades F que são agregados de forma contingente e que co-ocorrem em vários casos (BOYD, 1988, p. 323). Ele utiliza o termo ‘homeostase’ para se referir a esse grupo de indivíduos que compõem a definição de um termo. A ideia é de um conjunto de propriedades que compartilham suas qualidades constitutivas com um todo maior e por isso essas propriedades são unificadas em grupos.

Há duas características relevantes sobre agregados de propriedades homenostáticas. Quando temos um conjunto de propriedades homeostáticas F,

> Ou a presença de algumas das propriedades em F tendem (sob condições apropriadas) a favorecer a presença de outras, ou há mecanismos e processos subjacentes que tendem a manter a presença das propriedades em F, ou ambos (BOYD, 1988, p. 323).

Podemos dizer que essas são condições para termos um agregado de propriedades homeostáticas. É possível tornar essas condições mais explícitas dizendo que F é um agregado de propriedades homeostáticas se, e somente se:

(i) a presença de algumas das propriedades em F (sob condições apropriadas) tende a favorecer a presença de outras;
(ii) há mecanismos ou processos subjacentes que tendem a manter a presença das propriedades em F;
(iii) ambos, (i) e (ii)[15].

Como vimos, há um termo *t* que é aplicado ao agregado homeostático (‘metal’). Ora, *t* não pode ter uma definição analítica (BOYD, 1988, p. 323). A extensão do agregado de propriedades homeostáticas é sempre determinada *a posteriori*, portanto é sempre uma *questão em aberto*. Na medida em que descobrimos novos indivíduos que compartilham as propriedades e os mecanismos que unificam o agregado de propriedades, expandimos a extensão do termo

[15] Apesar de manter algumas diferenças, a ideia de transformar a citação de Boyd em três princípios mais explícitos é baseada em Michael Rubin (Cf. RUBIN, 2009, p. 125).

para esses indivíduos, mas não há a exigência de que todos os membros do agregado apresentem uma uniformidade no sentido estrito. É por isso que agregados homeostáticos imperfeitos são possíveis, pois um indivíduo pode fazer parte do agregado F e mesmo assim não apresentar todas as propriedades de F (*x* pode ser um metal, por exemplo, apresentar maleabilidade e não apresentar brilho).

Uma vez que o agregado é constituído *a posteriori*, as propriedades unificadas pelo mecanismo homeostático são tipos naturais. Como sugere Boyd, "a importância causal do agregado de propriedades homeostáticas *F* associado ao mecanismo homeostático subjacente relevante é tal que o tipo ou propriedade denotada por *t* é um tipo natural" (BOYD, 1998, p. 323). Além disso, dado que o agregado homeostático é constituído por propriedades naturais, sua definição é sempre passível de revisão.

Como podemos notar com essas definições em termos de agregados de propriedades homeostáticas, há um abandono da teoria semântica descritivista. O que fixa a referência não são os estados mentais ou as descrições que os indivíduos associam ao termo, mas as propriedades naturais que co-instanciam os mecanismos que unificam o agregado. Portanto, a referência de um termo *t* é estabelecida por conexões causais que determinam que tais e tais objetos fazem parte da extensão de *t*. Que o predicado *t* possa ser aplicado ao objeto *x* é determinado por uma cadeia causal independente do indivíduo que usa o termo (Cf. BOYD, 1988, p. 321). Considere a formulação de Boyd:

> *Grosso modo*, e para casos não degenerados, um termo *t* refere um tipo (propriedade, relação, etc) *k* quando há mecanismos causais cuja tendência é acarretar, através do tempo, que o que é predicado do termo *t* será aproximadamente verdadeiro de *k* […] (BOYD, 1988, p. 321).

Seguindo T. Horgan e M. Timmons (1992b), podemos reformular essa tese central do seguinte modo:

> T*ese da Regulação Causal (TRC)*: *k* regula causalmente o uso de *t*, em que *k* é a propriedade natural e *t* é o predicado usado para se referir a *k*.

Portanto, o que regula o nosso uso de 'metal', por exemplo, nos autorizando a aplicar este termo a tais e tais objetos são as propriedades naturais que, por apresentarem os mecanismos causais que os agregam num todo homeostático, determinam o que é e o que não é extensão de tal termo. Para citar um exemplo mais simples, o que regula o uso de 'água' para aplicá-lo ao

objeto *x* não é a minha intenção de aplicar o termo a *x*, mas é o fato de *x* apresentar a fórmula H_2O. Mas para agregados de propriedades homeostáticas, como afirma Boyd, aquilo que é predicado por *t* é apenas aproximadamente verdadeiro de *k*. Isso porque a definição do agregado é sempre passível de revisão e não apresenta condições suficientes para que um indivíduo faça parte do agregado. Podemos descobrir, por exemplo, que o material *x*, que considerávamos como uma co-extensão de 'metal', a despeito de ser formado por uma propriedade que é muito similar à propriedade que unifica os diferentes metais sobre o agregado de 'metal', na verdade não é extensão de 'metal', pois é formado por uma propriedade bastante peculiar até então desconhecida. Por isso, temos de admitir a possibilidade de que as propriedades que são predicadas por *t* são *aproximadas* de *k*.

Boyd defende que 'bom'[16] é um agregado de propriedades homeostáticas (Cf. BOYD, 1988, p. 331). Se isso é verdadeiro, então não temos uma definição rigorosa que estabeleça precisamente a extensão desse termo. Estamos a descobrir quais propriedades são instâncias de 'bom'. Boyd sugere que 'bom' moral corresponde às "coisas que satisfazem necessidades humanas importantes" (BOYD, 1988, p. 329). Obviamente, há um número indeterminado e crescente de coisas que satisfazem as necessidades humanas. Boyd nos fornece uma pequena lista:

> […] algumas dessas necessidades são físicas ou médicas. Outras são psicológicas ou sociais; essas (provavelmente) incluem a necessidade por amor e amizade, a necessidade de envolver-se em esforços cooperativos, a necessidade de exercer o controle sobre a própria vida, a necessidade para apreciação e expressão intelectual e artística, a necessidade de lazer, etc (BOYD, 1988, p. 329).

Podemos imaginar inúmeras instâncias que satisfazem esses grupos de necessidades humanas, tais como ser autônomo, ser fisicamente saudável, ser mentalmente saudável, ter amigos, criar e apreciar algum tipo de arte e assim por diante. O fato é que existem muitas instâncias em que o 'bom' é *multiplamente realizado* e a questão sobre quais são essas instâncias é uma questão empírica difícil (Cf. BOYD, 1988, p. 329). Tal como a definição de tipos naturais é imprecisa, não exaustiva, passível de revisão etc., a definição do agregado de propriedades que constituem o 'bom' é aperfeiçoada ou complementada à medida que descobrimos novos 'bens'.

Note que 'bom' satisfaz as duas condições para ser um agregado de propriedades homeostáticas. Dizer que o agregado de propriedades homeostáticas se suporta mutuamente, significa que a presença de algumas instâncias do 'bom' tende a favorecer a presença de outras.

[16] Estou usando 'bom' para se referir aos termos e propriedades morais em geral.

Além disso, a realização de alguns ‘bens’, como acesso fácil ao cuidado médico, por exemplo, suporta vários outros, tais como ser física ou mentalmente saudável, ter autonomia sobre a própria vida e assim por diante. O mesmo é válido para qualquer outra instância de ‘bom’.

No exemplo sobre ‘metal’ havia mecanismos bem definidos para unificar os indivíduos em agregados homeostáticos. No que diz respeito aos predicados morais, Boyd reconhece há diversos mecanismos homeostáticos que unificam grupos de instâncias do ‘bom’. Quer dizer, diferentemente de um mecanismo unificador para todo o agregado, o ‘bom’ possui vários agregados internos. Ele cita mecanismos como *cultivo de atitudes de respeito*, *relações sociais igualitárias*, *regras de cortesia*, etc. Para cada um desses mecanismos, podemos imaginar inúmeras ações que os instanciam. É importante salientar que esses mecanismos unificadores não agregam regras ou normas de conduta, mas sim instâncias de ‘bom’, quer dizer, ações em que ‘bom’ é *realizado*.

Considerando que o agregado para ‘bom’ aparentemente não apresenta apenas *um* mecanismo unificador de todas as suas instâncias, mas vários, é bastante razoável se perguntar o que faz com que estes diferentes mecanismos sejam agrupados sob o agregado ‘bom’. A sugestão de Brink é importante aqui. Brink concorda com Boyd que há um agregado de propriedades que constituem o ‘bom’. Por isso, diz que o ‘bom’ é superveniente a este agregado de propriedades. Ele diz o seguinte.

> Diferentes teorias morais determinam quais propriedades morais sobrevêm sobre quais propriedades naturais […]. O utilitarismo hedonista, por exemplo, pode ser tomado como uma reivindicação naturalista; ele alega que *correção*[17] = a maximização do prazer. […] A determinação de quais propriedades e fatos naturais constituem quais propriedades e fatos morais é uma questão de teoria moral substantiva (BRINK, 1989, p. 175-178).

Isso significa que a determinação do mecanismo que agrega todas as propriedades que constituem ‘bom’ é dependente da teoria normativa de primeira ordem em questão. A definição da propriedade moral nos é dada sinteticamente e o papel da teoria normativa é similar ao papel desempenhado pelas teorias da química ou biologia, por exemplo, em fornecer definições a partir das propriedades reais dos objetos. Os químicos foram responsáveis pelo descobrimento de que água é H_2O. As melhores teorias normativas tentam fornecer as melhores definições para o agregado de propriedades que constituem ‘bom’. Assim, podemos dizer que o

[17] ‘Correção’ não se refere ao ato de corrigir ou àquilo que é correto, mas à propriedade moral. Em inglês o termo é ‘*rightness’*.

mecanismo que unifica as propriedades que constituem o bom pode ser, por exemplo, *a maximização do prazer*, se assumirmos um tipo de utilitarismo hedonista.

Se poderia perguntar: por que as teorias normativas muitas vezes diferem sobre quais objetos instanciam o agregado de 'bom'? A resposta é que é uma questão difícil saber quais propriedades pertencem ao agregado e quais não pertencem. É por isso que a nossa definição é constituída *a posteriori*. O agregado de propriedades homeostáticas é "aberto", por assim dizer, pois sempre é possível incluir novos objetos que fazem parte dos mecanismos que unificam o agregado. É por isso que o AQA não é um problema para esse tipo de teoria naturalista.

Boyd defende que 'bom' é *definido* pelo agregado de propriedades e pelos mecanismos que unificam estas propriedades (Cf. BOYD, 1988, p. 329). Se é definido pelo agregado, então é bom tudo aquilo que pertence ao agregado.

> *Bom*: *x* é bom se, e somente se, *x* instancia o agregado de propriedades homeostáticas de '*bom*'.

'Bom', então, abarca um aglomerado inumerável de propriedades que são a sua instância. O que garante que a escravidão não seja um ato '*bom*', por exemplo, é que esta propriedade não pertence ao agregado homeostático de '*bom*', e isso nos é informado pelas melhores teorias normativas.

Desse modo, podemos adaptar a *Tese da Regulação Causal* para os termos morais.

> *Tese da Regulação Causal Moral (TRCM)*: para cada termo moral *t*, há uma propriedade natural *N* tal que *N*, e somente *N*, regula causalmente o uso de *t* pelos humanos.

Assim, um predicado moral que denota, por exemplo, a propriedade da *correção*, tem uma extensão natural e é a própria extensão que determina a quais instâncias podemos aplicar o termo moral que denota tal propriedade.

Como vimos, a teoria de Boyd lida muito bem com o AQA e o *Argumento do Desacordo* de Hare não se aplica devido ao pressuposto semântico descritivista. Além disso, essa teoria promete lidar com uma série de outros problemas normalmente vinculados ao Realismo Moral e a fornecer as melhores explicações para as características fenomenológicas

da moralidade. Juntamente com a defesa de que as propriedades morais são explanatoriamente eficazes, temos aqui uma das defesas mais sofisticadas e atrativas RMN.

7. A (Re) Ascensão do Desafio Clássico ao Naturalismo: O Argumento da Terra Gêmea Moral

Na seção 4, vimos como o AQA influenciou a rejeição do naturalismo moral e que, dada a insatisfação com o tipo de não-naturalismo proposto por Moore e a necessidade de responder ao *problema da motivação*, os filósofos acabaram desenvolvendo formas de antirrealismo. As teorias de Ayer, Stevenson e Hare foram as pioneiras do influente movimento não-cognitivista na metaética. Como vimos, também, a descoberta do *Problema Frege-Geach* colocou tais teorias em séria ameaça e, juntamente com o desenvolvimento das teorias externalistas da referência que recusavam premissas centrais do AQA e do *Argumento do Desacordo*, formou-se um terreno fértil para o renascimento de novas formas de naturalismo moral. Talvez o grande marco da década de 80 para a metaética seja o retorno a essas novas formas de RMN, como vimos, por exemplo, com o Realismo de Cornell.

O próximo grande passo na metaética, que têm sido levado adiante até os dias atuais, é o trabalho realizado pelos não-cognitivistas na tentativa de responder ao *Problema Frege-Geach* e propor novas formas de não-cognitivismo. Como acontece com qualquer teoria metaética, há sempre razões positivas (desenvolvimentos internos à própria teoria) e negativas (ataques a teorias concorrentes) a seu favor. Dentre as razões positivas a favor do não-cognitivismo, destacam-se as propostas de Allan Gibbard (1990), Simon Blackburn (1984) e, mais recentemente, Michael Ridge (2014) e Mark Schroeder (2008). Dentre as razões negativas, há uma proposta que tem direcionado muita atenção nos últimos anos. Na década de 90, dois filósofos americanos propuseram um influente argumento contra as novas formas de naturalismo moral. Eles pretendiam dar nova roupagem para o AQA e o *Argumento do Desacordo* adaptando-os para as teorias naturalistas atuais.

Numa série de artigos (1991, 1992a, 1992b), Horgan e Timmons (doravante H&T) apresentaram um argumento supostamente devastador às novas formas de naturalismo moral. Tal ataque ao naturalismo moral ficou conhecido como *Argumento da Terra Gêmea Moral* (doravante, ATGM), já que faz referência ao conhecido experimento mental de Putnam que vimos acima. Originalmente, o ATGM foi proposto como desafio a uma versão particular de naturalismo moral, a saber, a teoria de Boyd. No entanto H&T sustentam que o argumento pode

ser generalizado para outras formas de RMN, ou que pode ser, até mesmo, considerado como uma fórmula de argumentação para desafiar qualquer versão de RMN. O argumento proposto por H&T, como veremos, é amplamente baseado no *Argumento do Desacordo*, de Hare, e no AQA, de Moore. Por isso, pode-se falar em um retorno aos desafios clássicos ao RMN. No que segue, no intuito de facilitar a discussão, irei apresentar o ATGM como desafio especificamente direcionado à teoria de Boyd.

Como vimos acima, o núcleo da teoria semântica de Boyd pode ser resumido no que denominei de *Tese da Regulação Causal Moral* (TRCM):

> *TRCM*: para cada termo moral *t*, há uma propriedade natural *N* tal que *N*, e somente *N*, regula causalmente o uso de *t* pelos humanos.

O ponto do ATGM é fornecer evidência para a falsidade da TRCM. H&T desenvolvem o argumento a partir da ponderação sobre o seguinte experimento mental. Suponha que a TRCM seja verdadeira, isto é, que o uso de termos morais como 'correto' e 'bom' seja causalmente regulado por uma propriedade natural *N*. Como vimos, o próprio naturalista aceita que é uma questão em aberto exatamente qual propriedade N regula causalmente o uso dos termos morais, pois há ampla divergência entre os filósofos a respeito de qual teoria de primeira ordem se deveria adotar. Mas, para tornar o experimento mais vívido, suponha que, depois de ampla investigação moral, se chegou à conclusão de que uma teoria de primeira ordem consequencialista é a correta e que a propriedade *N* que regula causalmente o uso de 'correto' ('correto-*t*'), por exemplo, é a propriedade de *maximizar o agregado de felicidade*.

Agora, suponha que há um planeta chamado Terra Gêmea Moral (TG) que é *quase* uma duplicação perfeita do nosso planeta (Terra (T)). Não fosse por uma característica que mencionarei logo abaixo, T e TG seriam *exatamente* similares. TG possui a mesma composição natural e geográfica de T, possui um país chamado Brasil onde os Terráqueos Gêmeos falam português-gêmeo, numa cidade chamada Florianópolis há uma instituição chamada Universidade Federal de Santa Catarina, e assim por diante. É importante notar que as características do uso ordinário da linguagem moral são as mesmas em T e TG. Assim, tanto Terráqueos quanto Terráqueos Gêmeos

> [...] usam os termos 'bom' e 'mau', 'certo' e 'errado' para *avaliar* ações, pessoas instituições e assim por diante. [...] o uso desses termos na Terra Gêmea Moral suporta todas as marcas "formais" que tomamos como características do vocabulário moral e da prática moral. Em particular, os termos são usados para raciocinar sobre

> considerações relacionadas ao *bem-estar* da Terra Gêmea Moral; os habitantes da Terra Gêmea Moral estão normalmente *dispostos a agir* de certos modos em correspondência com os julgamentos sobre o que é 'bom' e 'correto'; eles normalmente tomam as considerações sobre o que é 'bom' e 'correto' como sendo especialmente importantes [...] para decidir *sobre o que fazer*, e assim por diante. (H&T, 1992b, p. 164).

Poderíamos, ainda, adicionar mais algumas dessas características "formais", tais como: tanto em T quanto em TG os falantes tendem a evitar ações a que normalmente aplicam o termo 'errado' e a perseguir e apoiar ações a que aplicam o termo 'correto'; os habitantes dos dois mundos tendem a sentir culpa ou vergonha quando realizam alguma ação a que normalmente se aplica o termo 'errado' e a sentir ressentimento quando outros realizam este tipo de ação.

A única diferença é que em TG a propriedade natural que regula causalmente o uso dos termos morais dos Terráqueos Gêmeos não é *N*, mas *N**, digamos. Aqui, novamente, para tornar o experimento mais vívido, podemos supor que os habitantes de TG, depois de muita teorização moral, chegaram à conclusão de que a melhor teoria moral de primeira ordem é um tipo de deontologismo e a propriedade *N** que regula causalmente o uso de 'correto' ('correto-*tg*'), por exemplo, é a propriedade de *tratar os outros como fins em si mesmos*.

Agora, considere o seguinte caso.

> *O Cirurgião*. Depois de ter feito uma cirurgia nas costas do paciente, o médico remove um dos rins de tal paciente e este tem uma vida normal e nunca vem a saber a respeito da prática do cirurgião. Ao remover o rim, o médico implanta e salva a vida de outro paciente que necessitava urgentemente de um transplante[18].

Suponha que este caso fosse apresentado para os habitantes de T bem como para os habitantes de TG. Suponha que os habitantes de T dissessem: 'a ação do cirurgião é correta-*t*'. E os habitantes de TG contestassem: 'a ação do cirurgião não é correta-*tg*'. Aqui, também com o objetivo de tornar o experimento mais vívido, podemos dizer que os habitantes de T consideram que a ação do médico maximiza o agregado de felicidade e, portanto, é correto, enquanto os habitantes de TG consideram que a ação do médico não trata o paciente como um fim em si mesmo e, portanto, não é correta[19].

---

[18] Estou emprestando este exemplo de Michael Rubin (Cf. RUBIN, 2014, p. 289).

[19] Ressalto que esses detalhes sobre as teorias morais de primeira ordem (consequencialismo e deontologismo) e o caso sobre o cirurgião são meramente pedagógicos. Não é óbvio que um consequencialista ou um deontologista necessariamente teria o veredito moral que estou sugerindo no experimento. Irei discutir alguns problemas relacionados a esses detalhes nos próximos capítulos. Se se quiser ignorar esses detalhes, basta supor que a propriedade causalmente reguladora dos termos morais em T é *N* e em TG é *N** e que ambos apresentam visões

Agora, o ponto é: habitantes de T e TG estão em um desacordo moral genuíno? Note que se a TRCM for o diagnóstico verdadeiro a respeito da semântica dos termos morais, então Terráqueos e Terráqueos Gêmeos não podem estar em desacordo genuíno. Pois, como a propriedade natural causalmente reguladora do uso dos termos morais nos dois mundos é diferente, Terráqueos e Terráqueos Gêmeos não estariam predicando a mesma propriedade a respeito do ato em questão. Com 'a ação do cirurgião é correta-*t*', o habitante de T estaria dizendo algo como 'a ação do cirurgião maximiza o agregado de felicidade' ou 'a ação do cirurgião é *N*' enquanto, com 'a ação do cirurgião não é correta-*tg*', o habitante de TG estaria dizendo algo como 'a ação do cirurgião não trata o paciente como um fim em si mesmo' ou 'a ação do cirurgião não é *N**'. Quer dizer, ambas as proposições poderiam ser verdadeiras ao mesmo tempo, pois uma não nega o valor de verdade da outra. Isso porque num caso em que a ação do cirurgião maximizaria o agregado de felicidade, a afirmação dos habitantes de T seria verdadeira. E num caso em que a ação do cirurgião não tratasse o paciente como um fim em si mesmo a afirmação dos habitantes de TG seria verdadeira. Sendo assim, se a teoria semântica de Boyd é correta, então se deveria ter a intuição de que *não* há desacordo algum entre habitantes de T e TG.

No entanto, como sugerem H&T, nossa intuição é de que há um desacordo moral genuíno entre Terráqueos e Terráqueos Gêmeos. Parece que a resposta mais apropriada é dizer que os habitantes dos dois mundos não estão apenas trocando informação trivial e compatível, mas estão expressando um desacordo moral substantivo a respeito da prática do cirurgião. Mas, se a nossa intuição for de fato amplamente compartilhada, forte e persistente, como H&T argumentam, então temos evidência para rejeitar a TRCM. Por conseguinte, temos evidência para recusar o RMN.

Podemos organizar esse argumento contra o naturalismo moral de forma canônica do seguinte modo:

> P1. Se o RMN (Realismo Moral Naturalista) é verdadeiro, então 'correto-*t*' expressa um conteúdo semântico que é diferente do conteúdo semântico expresso por 'correto-*tg*' (e esses dois predicados não são intersubstituíveis).
>
> P2. Se 'correto-*t*' expressa um conteúdo que é diferente do conteúdo expresso por 'correto-*tg*' (e estes dois predicados não são intersubstituíveis), então Terráqueos e

contrastantes a respeito de um ato A; habitantes de T sustentando que 'A é correto-*t*' e habitantes de TG sustentando 'A não é correto-*tg*'.

> Terráqueos Gêmeos não expressam um desacordo substantivo genuíno quando um diz '*x* é correto-*t*' e o outro diz '*x* não é correto-*tg*', em que ambos utilizam '*x*' para se referir à mesma ação.
>
> P3. Terráqueos e Terráqueos Gêmeos, na realidade, expressam um desacordo substantivo genuíno quando um diz '*x* é correto-*t*' e o outro diz '*x* não é correto-*tg*', em que ambos utilizam '*x*' para se referir à mesma ação. (*Intuição da Univocidade Semântica*).
>
> C. Portanto, o RMN não é verdadeiro[20].

Façamos três observações importantes. Em primeiro lugar, é clara a *semelhança* deste argumento com o experimento mental de H. Putnam a respeito do externalismo semântico. No experimento de Putnam conclui-se que o significado do termo 'água' na Terra é diferente do significado do termo "água" na Terra Gêmea, pois o primeiro se refere a H_2O e o segundo a XYZ. No *Argumento da Terra Gêmea Moral* (ATGM) espera-se um resultado similar se a teoria moral externalista de Boyd for verdadeira. Isto é, espera-se que 'correto-*t*' e 'correto-*tg*' tenham significados diferentes, pois referem-se a propriedades naturais diferentes. Mas, com a *Intuição da Univocidade Semântica*, H&T acreditam que temos evidência para o fato de que os termos morais não se comportam da mesma forma que os termos não morais. Em segundo lugar, nota-se de imediato o débito que H&T tem para com o *Argumento do Desacordo* de R. Hare. No ATGM se extrai conclusões semânticas a respeito dos termos morais a partir de intuições sobre o desacordo num cenário hipotético; tal como no argumento de Hare. Em terceiro lugar, como o ATGM busca mostrar que há incompatibilidade semântica entre termos morais e termos não morais, podemos notar outra dívida teórica de H&T, a saber, para com o AQA de Moore. Na verdade, H&T inclusive apresentam uma reformulação do AQA contra o naturalismo moral a partir do ATGM (H&T, 1992b, p. 162-168). Neste sentido, com o ATGM é possível notar uma tentativa de retorno ao tipo de ataque ao naturalismo moral que outrora fora dominante.

8. O que temos agora?

[20] Esta forma de organizar o argumento se baseia amplamente em M. Rubin (Cf. RUBIN, 2014, p. 290).

O argumento proposto por H&T têm sido peça central na discussão sobre a plausibilidade do RMN nas últimas décadas. Assim como, depois que Moore apresentou o AQA, manter uma posição naturalista em metaética requeria um diagnóstico sobre tal argumento, depois que H&T apresentaram o ATGM, manter uma posição naturalista em metaética requer que se tenha algum tipo de resposta ao ATGM. E é precisamente isso que os adversários ao ATGM têm tentado fazer nos últimos anos. Há muitas linhas de ataque ao argumento proposto por H&T e pode-se dizer que a ampla maioria da literatura busca recusar este argumento. Agora, irei mencionar brevemente algumas das principais estratégias que têm sido desenvolvidas por aqueles filósofos que pretendem rejeitar o ATGM.

Como vimos, o ATGM tem três premissas importantes. Na literatura atual, ataques a cada uma dessas três premissas além de um quarto tipo de ataque direcionado à forma do ATGM. Portanto, podemos dizer que há quatro estratégias principais que são perseguidas por quem rejeita o ATGM.

Estratégia 1. A não-similaridade entre o Argumento da Terra Gêmea e o Argumento da Terra Gêmea Moral

Como mencionei acima, H&T constroem um experimento de pensamento muito similar ao cenário hipotético desenvolvido por H. Putnam quando este propõe o Argumento da Terra Gêmea (ATG) a favor do externalismo semântico. Em várias passagens H&T apresentam o seu ATGM de uma forma muito similar ao ATG e isso pode dar a entender que eles estão propondo o seguinte raciocínio: se a teoria semântica de Boyd é verdadeira, então a ponderação sobre o cenário hipotético entre os Terráqueos e os Terráqueos Gêmeos Morais deveria gerar o mesmo tipo de intuição que temos quando consideramos o cenário hipotético do ATG; no ATGM não temos o mesmo tipo de intuição semântica que temos no ATG; portanto, a teoria semântica de Boyd é falsa (e, como a teoria de Boyd é referência para o novo naturalismo moral, o naturalismo moral é falso). Sendo assim, alguns críticos de H&T acreditam que, mostrando a suposta não-similaridade entre o ATG e o ATGM, eles estão a mostrar que o desafio proposto por H&T falha em fornecer evidência contra o naturalismo moral. Este caminho é percorrido por Stephen Laurence, Eric Margolis & Angus Dawson (1999) e Heimir Geirsson (2003, 2014).

Estratégia 2. Recusar P2 – Atacando a tese de que desacordo genuíno requer identidade conceitual

P2 do ATGM nos diz que Terráqueos e Terráqueos Gêmeos não expressam desacordo genuíno se seus predicados morais expressam conteúdo diferente. Vários metaeticistas têm questionado se desacordos morais genuínos requerem identidade conceitual sobre os termos morais. Eles argumentam que, mesmo que os predicados morais tenham conteúdo semântico diferente, ainda assim é possível que falantes expressem desacordo moral genuíno. Há três principais ataques ao ATGM que empregam esta estratégia. David Copp (2000) sustenta que, mesmo que os termos morais usados em T e TG expressem *propriedades* diferentes e, portanto, tenham referências distintas, se olharmos atentamente para a noção de *tradução* veremos que 'correto-*t*' é a melhor tradução para 'correto-*tg*' (e vice-versa), e isso seria suficiente para garantir que os habitantes de T e TG discordam genuinamente. David Merli (2002) argumenta que, embora Terráqueos e Terráqueos Gêmeos não expressem desacordo *moral* genuíno, eles expressam um outro tipo de desacordo, a saber, desacordo *prático* sobre *o que deve ser feito*. Isso, sustenta Merli, seria razão suficiente para concluirmos que P2 é falsa. Além disso, David Plunkett e Tim Sundell (2013) apresentam uma teoria muito mais sofisticada do que as hipóteses de Copp e Merli. Eles desenvolvem o que chamam de negociação metalinguística para argumentar que, mesmo que concordemos com H&T e admitamos que os predicados morais dos habitantes de T e TG tenham significado diferente, tal variação semântica ocorre apenas no nível literal. Há um outro nível discursivo que está em prática no cenário descrito por H&T, argumentam Plunkett e Sundell, a saber, o nível metalinguístico. Aqui, habitantes de T e TG estão em desacordo genuíno. Portanto, supostamente há desacordo genuíno mesmo que os falantes não compartilhem o conteúdo semântico dos termos morais.

Estratégia 3. Recusar P3 – Atacando a Intuição da Univocidade Semântica

P3 é o resultado da nossa intuição quando ponderamos sobre o cenário hipotético de H&T. Como afirma Mark van Roojen (VAN ROOJEN, 2006, p. 168) a maioria das pessoas que tem lido os trabalhos de H&T parece concordar com a intuição que tais filósofos pretendem extrair com seu exemplo, quer dizer, de que realmente há desacordo genuíno entre habitantes de T e TG. Sendo assim, na tentativa de recusar P3, vários metaeticistas tem insistido que a intuição não é boa evidência a favor de H&T. Esta é uma estratégia de réplica bastante recorrente contra o ATGM. Há uma série de objeções neste sentido, mas três trabalhos, em especial se destacam (porque abordam as objeções já apresentadas com maior profundidade e

apresentam teorias mais sofisticadas). Neil Levy (2011) argumenta que uma descrição *correta* da Terra Gêmea Moral iria gerar intuições diferentes, isto é, que não há desacordo genuíno. Aandrea Viggiano (2008) sustenta que a *Intuição da Univocidade Semântica* é amplamente compartilhada apenas em virtude de nossas limitações epistêmicas atuais. Num cenário, tal como ele descreve, em que nossas intuições não são enviesadas pela ausência de certos fatos, estaríamos em condição de perceber que a intuição gerada pelo ATGM não é a de que há desacordo genuíno entre Terráqueos e Terráqueos Gêmeos. Ainda, John Sonderholm (2012) defende que não devemos confiar nas intuições geradas pelo ATGM porque o cenário descrito em tal argumento e a nossa condição epistêmica atual é "radicalmente diferente". De acordo com esses filósofos, a *Intuição da Univocidade Semântica* não é suficientemente confiável para recusarmos o naturalismo moral.

Estratégia 4. Recusar P1 – Desenvolvendo novas teorias semânticas para o naturalismo moral

Esta é a estratégia mais promissora. Como vimos acima, a teoria semântica de Boyd foi uma das reformulações mais sofisticadas do naturalismo moral e possibilitou que seus adeptos lidassem bem com algumas objeções clássicas, tais como o AQA e o *Argumento do Desacordo*. A ideia mais geral de H&T é que, ao refutar a teoria de Boyd, se estaria refutando o RMN (como salientamos, o ATGM foi primeiro formulado para atacar o Realismo de Cornell, mas pode ser generalizado para outras teorias naturalistas). Há vários metaeticistas que aceitam que o ATGM refuta a teoria semântica de Boyd, mas argumentam que isso não implica que o naturalismo moral é falso. Neste sentido, tais filósofos buscam construir teorias semânticas alternativas que escapem ao desafio de H&T. Essa proposta é bastante promissora porque muda o foco da discussão sobre o Realismo de Cornell e abre possibilidades para novos tipos de RMN. É claro que seus defensores tem o ônus de mostrar porque a sua teoria naturalista não é vulnerável ao ATGM.

Há várias propostas neste sentido. Destaco algumas. B. Dunaway e T. McPherson (2014) argumentam que o naturalista evita o ATGM apelando para a ideia de *magnetismo da referência* (*reference magnetism*). Eles argumentam que esta ideia pode nos ajudar a explicar a justaposição extensional entre os predicados morais dos habitantes de T e TG e, consequentemente, explicar a razão desses falantes estarem engajados num desacordo genuíno, mas sem comprometer a ideia de que as propriedades morais são propriedades naturais. Filósofos como Copp (2000), e Tim Henning (2011) argumentam que se adotarmos teorias

semânticas híbridas podemos escapar ao desafio de H&T. Além disso, Brink, (2001) propõe um tipo de semântica moral normativamente enriquecida, diferente da teoria de Boyd, e sustenta que é suficiente para que o defensor do RMN possa evitar o ATGM.

Essas são as quatro estratégias centrais desenvolvidas por quem deseja rejeitar o desafio proposto por H&T ao RMN. Obviamente, essas abordagens são assunto de disputa nas discussões atuais em metaética. Ainda não há na literatura um trabalho amplo que se dedique a analisar conjuntamente essas controvérsias a respeito do ATGM. Portanto, temos aqui um tópico relevante da metaética contemporânea com vários aspectos potenciais a serem explorados. É o que tentarei fazer nos capítulos seguintes.

9. Conclusão

Nesse capítulo introdutório tentei apresentar o contexto da discussão em que se enquadra o presente trabalho. Apresentei uma versão rudimentar do naturalismo moral bem como os dois desafios semânticos clássicos a esse tipo de teoria, a saber, o AQA e o *Argumento do Desacordo*. Depois disso, explicitei um pouco da influência principalmente do AQA na história da metaética com o desenvolvimento das teorias não-cognitivistas. Como vimos, a descoberta do *Problema Frege-Geach*, bem como um contexto filosófico em que as teorias externalistas da referência ganharam força, prepararam o terreno para um retorno do naturalismo moral. Apresentei algumas ideias do *Realismo de Cornell* e dei ênfase especificamente à teoria metassemântica de Boyd à fim de caracterizar o RMN. Por fim, reconstruí o ATGM, que é o desafio semântico central às formas de RMN e esbocei, rapidamente, algumas linhas de réplica adotadas na literatura atual. Temos, assim, o pano de fundo da discussão que pretendo me inserir a partir de agora.

CAPÍTULO 2 – A AUTONOMIA DO ARGUMENTO DA TERRA GÊMEA MORAL

1. Introdução

Em *The Language of Morals* (1952), Richard Hare forneceu uma das defesas mais sofisticadas do Não-Cognitivismo Moral à época. Dentre a série de argumentos originais ali apresentados, temos o *Argumento do Desacordo*. Como vimos no capítulo anterior, com o emprego de tal argumento, Hare visava recusar o Naturalismo Moral e estabelecer razões a favor do que chamava de Não-Descritivismo. Mas, independentemente de sua solidez, assim como o AQA de Moore, o *Argumento do Desacordo* teve alcances limitados. Pode funcionar muito bem como ataque às formas analíticas de Naturalismo Moral, mas não se aplica às formas sintéticas contemporâneas. Isso dá margem a afirmações tais como a seguinte a respeito do argumento de Hare: "podemos ser tentados a ver o desafio de Hare como não mais do que uma relíquia exótica tornada obsoleta pelo estado da arte atual" (MERLI, 2002, p. 209). No entanto, isso seria um erro, pois, como H&T mostraram, é possível fornecer uma nova roupagem ao *Argumento do Desacordo* de Hare. Daí, temos o *Argumento da Terra Gêmea Moral* (ATGM).

Na tentativa de estender o desafio de Hare às formas atuais do Naturalismo Moral, cujas pressuposições semânticas são externalistas (como vimos no capítulo anterior), H&T se inspiram no cenário hipotético desenvolvido por Hare e o adaptam à descrição fornecida por H. Putnam em seu *Argumento da Terra Gêmea* (doravante, ATG). Essa *aparente* dependência para com o ATG, chamou a atenção de alguns críticos do ATGM. Na interpretação de Stephen Laurence, Eric Margolis e Angus Dawson (doravante, LM&D) (1999), por exemplo, H&T simplesmente adotam o experimento mental de Putnam e substituem os termos não-morais por termos morais argumentando que, se o naturalismo moral é verdadeiro, se deveria ter a mesma intuição que se tem no argumento original de Putnam, isto é, se deveria ter a intuição de que os termos não são semanticamente unívocos; mas, como não se tem tal intuição, o naturalismo moral é falso. Como os autores afirmam, "todo o ponto do argumento direto de H&T é de que há uma assimetria entre as intuições geradas pela Terra Gêmea [de Putnam] e pela Terra Gêmea Moral" (LM&D, 1999, p. 155). Note que, se for assim que H&T constroem o argumento, há uma dependência da perfeita analogia entre o experimento mental construído com termos não-morais e do experimento mental construído com termos morais.

Sendo assim, embora incomum, uma das estratégias de réplica ao desafio de H&T têm sido atacar a suposta similaridade entre o ATGM e o ATG. Considere, por exemplo, as seguintes afirmações:

> Para que o argumento [ATGM] funcione, no entanto, os dois experimentos mentais [ATG e ATGM] devem ser construídos analogamente. O problema com o argumento [de H&T] é que eles não são construídos analogamente. Há uma série de desanalogias cruciais entre os dois experimentos mentais e são essas desanalogias que fazem grande parte do trabalho em gerar as intuições de que os argumentos de H&T dependem (LM&D, 1999, p. 155).

> Irei argumentar que as dissimilaridades [do ATGM] com o experimento mental de Putnam são significantes, de modo que a dependência em intuições do tipo-Putnam são questionáveis (GEIRSSON, 2014, p. 92).

Os trabalhos de LM&D (1999) e Heimir Geirsson (2014) são os dois representantes dessa linha de réplica ao ATGM. Eles pressupõem que há uma interdependência relevante entre o argumento desenvolvido por H&T e o experimento mental de Putnam. A partir disso, buscam evidenciar certas dissimilaridades entre os dois experimentos mentais argumentando que isso inviabiliza o alcance do ATGM.

Neste capítulo tentarei fornecer razões para se recusar esses ataques ao ATGM. Irei considerar, primeiramente, os três argumentos apresentados por LM&D. As réplicas que irei apresentar e defender são amplamente baseadas no trabalho de Michael Rubin (2008). Em segundo lugar, discutirei o trabalho de Geirsson. Em geral, tentarei defender que o ATGM é independente e autônomo em relação ao ATG. Se obtiver sucesso em mostrar a independência entre o ATGM e o ATG, além de livrar o desafio semântico proposto por H&T dessa primeira linha de ataque, abrirei caminho para a seguinte hipótese (embora não pretenda desenvolvê-la propriamente: se o argumento de H&T possui credores, trata-se muito mais do *Argumento do Desacordo* de Hare do que o ATG de Putnam.

2. Laurence, Margolis e Dawson

Em *Moral Realism and Twin Earth* (1999), LM&D identificam algumas dissimilaridades entre o ATGM de H&T e o ATG de Putnam e argumentam que isso ilicitamente favorece a *Intuição da Univocidade Semântica* (IUS). Eles apresentam três réplicas sustentando que, por não respeitarem a correta analogia com o ATG, H&T de alguma forma manipulam a nossa intuição em favor da conclusão que pretendem extrair com o ATGM. Por

fim, os autores argumentam que, uma vez que essas três desanalogias são corrigidas, o ATGM não mais suporta a conclusão de que 'correto-*t*' e 'correto-*tg*' possuem univocidade semântica. Para elucidar esse ponto, eles esboçam uma versão revisada do ATGM que visa ser análoga ao ATG. Sua conclusão é que, uma vez que o ponto do ATGM depende completamente da IUS e esta, por sua vez, desaparece numa descrição "corrigida" do cenário hipotético, a conclusão de que há um problema para o RMN não tem mais suporte. A seguir, considero cada uma das três réplicas argumentando que não nos dão razão suficiente para rejeitarmos o ATGM.

2.1. Argumento 1: teorias morais concorrentes no ATGM vs. teorias químicas não-concorrentes no ATG

A passagem chave do argumento é a seguinte:

> No experimento mental de H&T, o modo em que o contraste entre Terra e Terra Gêmea é feito é em termos de duas teorias morais concorrentes – consequencialismo e deontologia. As duas teorias compartilham plausibilidade, razão pela qual ambas continuam tendo fortes defensores nos círculos filosóficos. Por outro lado, no experimento da Terra Gêmea original, XYZ é uma invenção filosófica. Não há equívoco sobre isso: XYZ é uma composição química completamente diferente de H2O e, além disso, é uma composição química conectada a uma teoria química que ninguém jamais considerou ser verdadeira a respeito de água. [...] Lembre que na interpretação de H&T da Terra Gêmea Moral, em contraste com o caso padrão da Terra Gêmea, nossa intuição é de que os Terráqueos Morais Gêmeos não estão se referindo a propriedades diferentes com seus termos morais; eles apenas têm crenças e teorias diferentes sobre as mesmas propriedades morais a respeito das quais nós temos crenças. No entanto, certamente, escolher propriedades que satisfazem uma teoria moral concorrente sobre propriedades morais irá enviesar o caso em direção a essa interpretação. (LM&D, 1999, p. 156).

Em outras palavras, LM&D estão dizendo o seguinte. Há uma diferença não-trivial entre o ATGM e o ATG no que diz respeito às teorias que captam as propriedades *N* nas duas comunidades. No ATGM tanto a teoria aceita em T quanto a teoria aceita em TG representam possibilidades epistêmicas para nós enquanto apreciadores do experimento mental. Diferentemente, no ATG as teorias adotadas em T e TG não são concorrentes. Sabemos, de antemão, que a teoria química que define 'água' como sendo H_2O é a correta e que XYZ é apenas uma "invenção filosófica". De acordo com LM&D, essa diferença é de fundamental importância, pois tem o potencial de enviesar a nossa intuição a favor da univocidade semântica entre os predicados morais. Quando apreciamos o experimento mental, embora H&T nos peçam para que façamos o exercício de isolar as propriedades reguladoras dos termos morais de T e TG (assim como no ATG), consequencialismo e deontologismo continuam sendo

possibilidades epistêmicas (éticas) como pano de fundo em nossas mentes, de modo que isso facilita a nossa inclinação a dizer que 'correto-*t*' e 'correto-*tg*' são predicados semanticamente unívocos.

LM&D ressaltam, ainda, que, se H&T quisessem ser estritamente rigorosos ao manter a correta analogia com o ATG, teriam que estipular teorias não-concorrentes sobre as propriedades reguladoras dos termos morais. No entanto, se se fizer isso, argumentam, a IUS desaparece e, por conseguinte, não há mais problema para o RMN (LM&D, 1999, p. 157).

Para começar, gostaria de fazer uma observação sobre a interpretação que LM&D fazem do ATGM. Tal observação provavelmente aparecerá mais vezes no decorrer deste trabalho, pois se trata de uma forma comum, embora equivocada, de entender o argumento. Note que LM&D sustentam que o ponto da IUS é evidenciar que os termos morais, em T e em TG, refere-se às mesmas propriedades:

> [...] o ponto da Terra Gêmea Moral de H&T é que, em contraste com o experimento da Terra Gêmea padrão, nossas intuições são de que os Terráqueos Gêmeos não estão se referindo a *propriedades* diferentes com seus termos morais (LM&D, 1999, p. 156, itálico meu).

Este não é o ponto do ATGM. O que se pretende extrair com a IUS é que, embora assumamos para fins argumentativos que os termos morais referem propriedades diferentes, ainda assim há algum tipo de característica semântica que os unifica. E tal característica semântica comum é o *conteúdo conativo*. Essa é a parte positiva do ATGM. H&T pretendem não apenas recusar o RMN, mas, ao mesmo tempo, fornecer evidência a favor do expressivismo moral. Para este tipo de teoria, termos morais *não referem propriedades*. Assim, dizer que o ponto do ATGM é que os termos morais em T e TG possuem a mesma referência não é a leitura mais apropriada.

Passemos, então, à primeira objeção de LM&D ao ATGM. Como mostrou Rubin (2008), (i) a estipulação de teorias concorrentes no ATGM pode ser considerada uma virtude de tal experimento – e não um vício, como argumentam LM&D – e (ii), mesmo que elaboremos cenários hipotéticos com teorias não concorrentes, como querem LM&D, a IUS é preservada.

O experimento original de Putnam visa fornecer evidências a favor do externalismo semântico. Essa família de teorias é a pedra angular para os defensores do RMN, de modo que, se o externalismo semântico estiver em ameaça, o RMN estará em ameaça. Agora, como argumenta Rubin (2008, p. 316-317), a estipulação de teorias concorrentes no ATGM além de estar longe de representar uma fraqueza para este argumento (ao contrário, representa uma virtude), parece evidenciar um problema para o próprio ATG (argumento este que o defensor

do RMN quer preservar a todo custo). Por que? Porque o fato de o ATG não estipular teorias concorrentes sobre a regulação causal de 'água' entre T e TG parece enviesar a intuição a favor do próprio externalismo semântico. No espírito da crítica de LM&D, se poderia dizer o seguinte. Nós, enquanto leitores do experimento mental de Putnam, já associamos de antemão que teoria química correta sobre a referência de 'água' é a que afirma H_2O. Aprendemos desde os primeiros anos escolares que água é H_2O, mesmo que não manipulemos com total competência os conceitos da química, de modo que H_2O se torna parte do nosso próprio conceito de água nos causando a disposição para reconhecer que se algo é água então é H_2O. Isso implica que quando consideramos o cenário do experimento de Putnam, é necessário que façamos um exercício de suspensão de nossas crenças que pode parecer exigente, já que o que está em questão são as nossas intuições. Pode ser que quando analisamos o experimento não damos o peso justo à teoria química que diz que 'água' é XYZ e isso acabe influenciando a nossa intuição a favor do ponto que Putnam quer provar, a saber, se algo não é H_2O então não é água. Por outro lado, o ATGM ao estipular teorias concorrentes sobre a regulação causal dos termos morais, parece estar livre dessa dificuldade.

Agora, suponha que o (a) defensor (a) do RMN resolva defender o ATG dessa suposta distorção da intuição a favor do externalismo semântico. Terá, então, que substituir a ideia de teorias não concorrentes sobre a regulação causal de 'água' por teorias concorrentes. Poderá obter sucesso nesta empreitada ou não. Se obtiver sucesso, então terá mostrado que a estipulação de teorias concorrentes em experimentos mentais *à la* Putnam não são um problema, mas uma virtude (que o ATGM preserva). Se não obtiver sucesso, então terá que aceitar que o ATG tem um possível problema de forçar a intuição dos leitores em direção do externalismo semântico. Ora, não parece ser uma estratégia atrativa atacar o ATGM por não preservar a analogia com o ATG a respeito de um ponto que é justamente um problema para o ATG.

Mas a primeira objeção de LM&D vai além. Mesmo que a estipulação de teorias concorrentes seja uma virtude do ATGM (e não um vício), ainda assim, pode ser o caso que ao considerarmos o cenário descrito por H&T nossa intuição seja distorcida em favor da IUS devido ao fato das teorias que regulam o uso dos termos morais em T e TG ambas representarem possibilidades epistêmicas para nós. Neste sentido, se o ponto de LM&D estiver correto, então se construirmos um ATGM alternativo estipulando teorias de primeira ordem não-concorrentes não deveríamos ter a IUS. Caso contrário, isto é, se a intuição permanecer mesmo com a referida modificação, o argumento de LM&D não nos fornece razão suficiente para rejeitarmos o ATGM. Testemos, então, a hipótese de LM&D.

Considere a seguinte variação do ATGM. Suponha que os habitantes de T aceitam uma teoria de primeira ordem consequencialista de acordo com a qual a propriedade R-relacionada ao uso dos termos morais é a propriedade de *maximizar o agregado de felicidade*. Os habitantes de TG aceitam uma forma peculiar de egoísmo moral como teoria de primeira ordem de acordo com a qual a propriedade R-relacionada ao uso de seus termos morais é a propriedade de *satisfazer unicamente os interesses do agente*. Em ambas as comunidades os termos morais são usados para aprovar ou reprovar ações, pessoas, instituições etc; os agentes normalmente são internamente motivados a agir de acordo com os juízos morais e assim por diante. Agora, suponha que os habitantes de T e TG discutam sobre o status moral da ação de doar para a caridade. Os membros de T sustentam que doar para a caridade é correto enquanto que os membros de TG sustentam que doar para a caridade não é correto. Estão a expressar um desacordo genuíno? Tendo a crer que, por preservar o paralelismo com o ATGM original, nossa intuição é de que sim e que, portanto, a IUS é preservada.

Note que (i) a teoria de primeira ordem aceita em TG não parece ser uma possibilidade epistêmica persuasiva para nós. Por conseguinte, tal variação do ATGM não contém teorias de primeira ordem concorrentes. Ainda assim, (ii) a teoria aceita em TG é uma teoria moral – o egoísmo moral é normalmente apresentado como uma teoria moral ((Brandt, 1959; Feldman, 1978; Kagan, 1998) – portanto, o desacordo entre os membros de T e TG é possível.

Temos um caso que satisfaz a exigência de LM&D, a saber, de que as teorias de primeira ordem de T e TG *não* sejam concorrentes, e, mesmo assim, a IUS é preservada. Pelas razões apresentadas, devemos concluir que a estipulação de teorias concorrentes no ATGM não é determinante para o direcionamento ilícito da intuição a favor da univocidade.

2.2. Argumento 2: propriedades funcionais vs. propriedades não-funcionais

LM&D sustentam que o ATGM está sujeito a um segundo problema devido ao fato de H&T assumirem, diferentemente do experimento de Putnam, que no cenário moral as propriedades são funcionais.

> Outra distorção potencial influente nas intuições sobre a Terra Gêmea Moral é o fato de que se assume que as propriedades morais são propriedades funcionais. Em contraste, o experimento mental da Terra Gêmea original é construído em termos de tipos naturais não-funcionais. [...] Isso nos traz novamente ao ponto principal de que H&T podem ganhar alguma influência falsa contra o naturalismo ético meramente porque, no ponto crucial de seu argumento, eles comparam propriedades éticas a tipos naturais não-funcionais, como água. (LM&D, 1999, p. 157-159)

Uma suposição de H&T é que as propriedades morais são funcionais, tal como defendido pela teoria de Boyd. Nesta teoria, uma propriedade moral como *bondade* é uma propriedade de segunda ordem *constituída*, ou *multiplamente realizada*, por inúmeras instâncias particulares que compartilham algum mecanismo unificador (que é fornecido pela teoria moral de primeira ordem que se adota). Isso permite assumir que as propriedades que realizam a propriedade funcional, embora possam ser unificadas por características comuns, são diferentes entre si. Assim, nós podemos dizer que uma propriedade é funcional quando a sua instanciação, por um indivíduo, depende da instanciação de outras propriedades que, juntas (com parte de um *cluster*), desempenham um papel causal. Por exemplo, se uma ação instancia a *bondade*, assumindo que *bondade* é uma propriedade funcional, então ela também instancia outras propriedades que realizam a *bondade* (tais como a propriedade de *ser uma ação igualitária* ou *ser uma ação que contribui para o bem-estar*). Um tipo (ou propriedade) é não-funcional quando a sua instanciação, por um indivíduo *x*, não depende da instanciação de um conjunto de outras propriedades que a realizam. Um bom exemplo é água = H_2O.

O problema para H&T, segundo LD&M, é que, enquanto o ATG é colocado em termos de tipos naturais não-funcionais, o ATGM supõe que *bondade* é uma propriedade funcional. A preocupação é que isso possa enviesar a intuição dos leitores do experimento a supor que 'correto-*t*' e 'correto-*tg*' não são propriedades diferentes, mas são ambas realizações de uma mesma propriedade moral de ordem superior. Isso explicaria (ou facilitaria) a IUS.

Note que o núcleo da objeção é de que há problemas em se aplicar a semântica dos tipos naturais funcionais aos termos morais. LM&D sustentam há uma diferença relevante entre o ATG e o ATGM, já que o primeiro não emprega termos cuja referência é constituída pela realização múltipla de particulares enquanto o segundo emprega. No entanto, como mostra Rubin (2008), o ponto fraco desse tipo de objeção é que, se há realmente problemas ou se não é possível estender a semântica dos tipos naturais funcionais para termos morais, então nem precisaríamos do ATGM para atacar o RMN. Isso porque aplicar a semântica dos tipos naturais aos termos morais é justamente o que a teoria de Boyd sobre agregados de propriedades homeostáticas faz e o que vários defensores do RMN endossam (Boyd, 1988, p. 196; Brink, 1989, p. 157; Sturgeon, 2003, p. 534). Então, o naturalista que pretende atacar o ATGM, a não ser que tenha uma teoria naturalista bastante diferente da que estamos considerando e desde que mostre que nenhuma adaptação do ATGM lhe é aplicável, não pode se basear na suposta não aplicação da semântica dos tipos naturais para os termos morais.

Além disso, o principal problema da objeção de LM&D é que se assumirmos que *correção* é a mesma propriedade de segunda ordem em T e TG e é apenas realizada por propriedades particulares diferentes, somos levados a uma forma de Relativismo Moral que é incompatível com o RMN. Como a própria objeção deixa margem para que as propriedades realizadoras sejam diferentes (a própria teoria de Boyd deixa), podemos apontar para os casos em que as teorias particulares prescreverão cursos de ação diferentes. Isso, no entanto, é uma forma de Relativismo Moral.

2.3. Argumento 3: a dificuldade de isolar as propriedades morais

O ponto central da terceira objeção de LM&D a H&T é a suposta dificuldade em se isolar as propriedades morais entre T e TG. Ou seja, manter que o uso de 'correto-*t*' e de 'correto-*tg*' é causalmente regulado por propriedades naturais diferentes e, ao mesmo tempo, preservar a similaridade exata entre T e TG no que diz respeito a todos os outros aspectos. Como eles afirmam,

> O simples fato de que seu sistema conceitual [dos habitantes de TG] é tão parecido com o nosso e de que eles compartilham nossos interesses sociais e culturais – tem governos, músicos de rock e assim por diante – certamente enviesa a interpretação de que seus termos morais devem referir as mesmas propriedades que os nossos termos morais. (LM&D, 1999, p. 160).

Como podemos notar, LM&D estão sustentando o seguinte: dado que T e TG são exatamente similares em tudo o que não diz respeito às propriedades naturais reguladoras do uso dos termos morais (compartilham amplo conjunto de interesses culturais, governamentais etc.), é *difícil* ver 'correto-*t*' e 'correto-*tg*' como não referindo as mesmas propriedades naturais. Essa comunalidade entre os dois planetas em *quase* todos os aspectos, enviesaria a intuição dos apreciadores do experimento mental em favor da IUS. De acordo com LM&D, o que parece dificultar o julgamento dos falantes é o fato de que em muitos casos as teorias de primeira ordem são coinstanciadas em T e TG, isto é, os particulares que realizam multiplamente a *correção*, por exemplo, são os mesmos se assumirmos o consequencialismo e o deontologismo. Portanto, mesmo que a propriedade causalmente reguladora de 'correto-*t*' seja *maximizar o agregado de felicidade* apenas, os leitores não estão livres do fato de que *tratar os outros como fins em si mesmos* pode reunir o mesmo conjunto de instâncias, mesmo que tal propriedade não seja adotada como melhor teoria em T. Isso, argumentam LM&D, facilita o nosso juízo em favor de que 'correto-*t*' e 'correto-*tg*' são semanticamente unívocos. Para além disso, há uma

diferença para com o ATG original de Putnam aqui (LM&D, 1999, p. 160). No ATG, não há coinstanciação entre H_2O e XYZ, ou seja, a ocorrência de 'água' em TG é captada apenas por XYZ enquanto em T é captada apenas por H_2O e os leitores não tem motivos para supor que 'água' se refira a H_2O em TG ou a XYZ em T. Ao não respeitarem a similaridade com o experimento original de Putnam, argumentam LM&D, H&T estão propondo um experimento mental com características filosóficas claramente não neutras.

Como nota Rubin (1999), ao considerarmos a objeção de LM&D é importante notarmos o seguinte: o juízo de que os termos morais de T e TG refletem o mesmo conteúdo não é um fato semântico assumido de antemão por H&T, mas um fato sobre a reação dos falantes ao considerarem o experimento. É a falta de atenção aos detalhes, digamos, que levaria os leitores a supor que, dado que os dois planetas são quase totalmente similares, o são também no que diz respeito à referência dos termos morais, pois, na realidade, assume-se na descrição do experimento que 'correto-*t*' e 'correto-*tg*' possuem referências distintas. Este é o ponto de LM&D[21]. Para tornar o ponto mais claro, poderíamos dizer que a explicação de LM&D para a IUS é de que há uma *característica enviesadora da intuição*, que é: em T há ocorrência de 'correto-*tg*' e em TG há ocorrência de 'correto-*t*' e isso confunde o juízo dos leitores. Já no experimento de Putnam, sustentam LM&D, não haveria tal *característica*.

Pois bem, isso implica que se tal explicação para a IUS fornecida por LM&D obtêm sucesso, se colocássemos a *característica enviesadora da intuição* em outro experimento (preservando os componentes relevantes), mesmo que fosse com termos não morais, deveríamos ter o mesmo resultado, ou seja, deveríamos ter uma espécie de favorecimento da intuição em direção à univocidade. Caso contrário, tal característica não é explanatória. Sendo assim, é possível construir um experimento alternativo para testar essa hipótese. Se houver enviesamento da intuição, assim como no experimento de H&T, então LM&D estão certos e a IUS pode realmente ser uma falha. Por outro lado, se não houver, então isso significa que a hipótese explanatória fornecida por LM&D não está correta.

Rubin (1999, p. 322-3) chama atenção para o fato de que o próprio Putnam tem uma variante de seu experimento original que pode ser usada para testar a hipótese explanatória de LM&D. Lembre que LM&D chamam atenção para o fato de que no ATGM, embora não haja 'correto-*tg*' em T ou 'correto-*t*' em TG, dado que a referência de tais termos é bastante similar,

[21] É importante não confundir tal ponto com o *insight* que os próprios H&T querem extrair com o ATGM. Eles de fato argumentam que os termos morais dos dois planetas são semanticamente unívocos, mas isso se deve ao fato de haver um conteúdo não-cognitivo que é comum. Aqui, LM&D estão tentando fornecer uma explicação alternativa, isto é, como os falantes teriam a IUS mesmo na ausência da explicação não-cognitivista de H&T.

os membros de T seriam confrontados não apenas com instâncias de 'correto-*t*', mas *aparentemente* de 'correto-*tg*' também (e vice-versa). Já no ATG isso não seria o caso, pois os membros de T não têm familiaridade com XYZ e os membros de TG não tem familiaridade com H_2O. Para testar a hipótese explanatória de LM&D precisamos de um caso em que, embora a referência dos termos em questão seja diferente em T e TG, os membros dos *dois* planetas estão de alguma forma familiarizados com a entidade que constitui a referência dos termos de T e TG.

Considere o seguinte. Há um planeta chamado Terra Gêmea que é uma réplica perfeita da nossa Terra exceto pelo fato de que os habitantes de TG usam o termo 'molibdênio' para todos os propósitos pelos quais nós, habitantes de T, usamos o termo 'alumínio'. Putnam nos pede para supor que "o molibdênio é tão comum na Terra Gêmea quanto o alumínio o é na Terra e que o alumínio é tão raro na Terra Gêmea quanto o molibdênio o é na Terra" (PUTNAM, 1975, p. 225s). Além disso, devemos supor que alumínio e molibdênio são indiscerníveis em suas qualidades superficiais (tal como H_2O e XYZ). No entanto, em T o uso 'alumínio' é causalmente regulado pelo elemento com número atômico 13 (Al) enquanto o uso de 'molibdênio' é causalmente regulado pelo elemento com número atômico 42 (Mo). Por outro lado, em TG o uso de 'alumínio' é causalmente regulado por Mo e o uso de molibdênio é causalmente regulado por Al. Isso significa que em T o uso de 'alumínio' é aplicado a Al e em TG o uso de 'alumínio' é aplicado a Mo.

Agora, note que, embora o molibdênio seja raro em T, ele está presente e que, embora o alumínio seja raro em TG, também está presente. Com isso, preservamos a similaridade com o ATGM em que, aparentemente, há ocorrência de 'correto-*t*' em TG e 'correto-*tg*' em T. Isso é importante pelo seguinte motivo. LM&D estão argumentando que a intuição dos leitores do ATGM é enganada pelo fato de que a propriedade que regula causalmente o uso dos termos morais estar presente em ambos os planetas. Nessa variante do experimento, *de fato* a propriedade que regula causalmente o uso de 'alumínio' em T (Al) está presente em TG e a propriedade que regula causalmente o uso de 'alumínio' em TG (Mo) está presente em T. Assim, com a preservação da característica supostamente enviesadora da intuição, deveríamos ter o seguinte resultado: ao apreciarem o experimento, os falantes deveriam ter a intuição de que há univocidade semântica entre 'alumínio-*t*' e 'alumínio-*tg*'. Mas, será este o caso?

Suponha que habitantes de T e TG se encontrem e se engajem numa discussão sobre se um determinado particular é ou não uma instância de alumínio. Os Terráqueos sustentam que *x* é alumínio enquanto os Terráqueos Gêmeos sustentam que *x* não é alumínio. Haveria

desacordo genuíno aqui? Parece claro que não. E isso se deve a uma intuição anterior. A intuição de que 'alumínio-*t*' e 'alumínio-*tg*' não são semanticamente unívocos. 'Alumínio-*t*' se refere a Al. 'Alumínio-*tg*' se refere a Mo.

A nossa intuição deveria ser de que 'alumínio-*t*' e 'alumínio-*tg*' são semanticamente unívocos, pois construímos um experimento mental que preserva a característica supostamente enviesadora da intuição, tal como proposta por LM&D. No entanto, tal intuição não parece estar presente. Portanto, a hipótese de LM&D não pode ser a melhor explicação para a IUS. E, na ausência de uma explicação alternativa melhor do que a proposta por H&T, isto é, que temos a IUS porque os termos morais compartilham um significado conativo, devemos permanecer com a primeira alternativa.

Diante dessas réplicas aos três argumentos desenvolvidos por LM&D, podemos concluir que o apelo para as supostas diferenças entre o experimento apresentado por H&T e o experimento mental de Putnam não nos dá razão suficiente para rejeitarmos o ATGM. Pelo menos até agora, pois há outros ataques ao ATGM que adotam a mesma estratégia de LM&D. Consideremos, agora, a proposta de H. Geirsson.

3. H. Geirsson

Em *Moral Twin Earth, Intuitions and Kind Terms* (2014), H. Geirsson apresenta quatro pontos de dissimilaridade entre o ATG e o ATGM e argumenta, a partir de cada um desses pontos, que a conclusão que H&T querem extrair com o argumento não se sustenta. Como o objetivo do presente capítulo é considerar apenas o tipo de réplica ao ATGM que diz respeito à relação entre o ATG e o ATGM, terei que considerar apenas os pontos 1 e 3. Os pontos 2 e 4 representam a tese de Geirsson de que desacordos genuínos não requerem identidade extensional dos termos envolvidos no debate. Irei dedicar um capítulo inteiro a este tipo de crítica considerando, inclusive, versões mais sofisticadas do que a de Geirsson. Tal como em relação às réplicas de LM&D, acredito os pontos 1 e 3 de Geirsson não são fatais contra o ATGM, como pretendo mostrar a seguir. Passemos a eles.

3.1. Intuições semânticas vs. intuições sobre o desacordo

No experimento de Putnam simplesmente se assume de início que os termos em T e TG têm referências distintas e a questão de se realmente isso é o caso não se coloca. Já no

experimento de H&T, o problema sobre a referência distinta entre os termos morais em T e TG é o centro da questão. E, como corretamente afirma Geirsson, a ferramenta que opera o raciocínio de H&T a favor da conclusão de que 'correto-*t*' e 'correto-*tg*' têm o mesmo significado é o argumento baseado na possibilidade de desacordos genuínos. No entanto, sustenta Geirsson, é aqui que vemos uma importante diferença entre o ATG e ATGM. Pois, no primeiro experimento, quando Putnam recorre às nossas intuições ele está se referindo a *intuições semânticas*. E quando H&T recorrem as nossas intuições eles estão se referindo a *intuições sobre o desacordo*.

> É importante notar que as intuições a que Horgan e Timmons apelam aqui não tem nada a ver com o nosso domínio competente de normas semânticas. Ou seja, eles não apelam para intuições sobre a referência. Ao invés disso, depois de consultar nossas intuições sobre o desacordo, Horgan e Timmons inferem que 'água-t' e 'água-tg' não se referem a diferentes tipos ou propriedades. Putnam, por outro lado, não faz uso de intuições sobre o desacordo em seu experimento mental da Terra Gêmea e ele não conclui nada sobre a referência a partir da questão do desacordo. Há, portanto, uma clara desanalogia aqui entre o uso que Putnam faz da Terra Gêmea e o uso que Horgan e Timmons fazem da Terra Gêmea Moral. (GEIRSSON, 2014, p. 99)

O que isso implica? É importante notar que Geirsson pressupõe que H&T sustentam que as intuições geradas pelo ATGM são em essência "as mesmas" intuições provocadas pelo ATG (GEIRSSON, 2014, p. 101-102). Ora, se num caso temos intuições semânticas e noutro intuições sobre o desacordo, como argumenta Geirsson, então está claro que não se trata das mesmas intuições.

Esse tipo de objeção a H&T ainda nos deixa a dúvida de por que tal detalhe apontado por Geirsson seria um problema para o ATGM. Em outras palavras, dada essa diferença entre o ATG e o ATGM, por que o ATGM não funcionaria ou deveria ser recusado? A resposta de Geirsson parece ser simplesmente a seguinte: H&T estão sugerindo que seu argumento gera as mesmas intuições que o argumento de Putnam e, como se pode notar, ver, não se trata das mesmas intuições; portanto, devemos recusar o ATGM.

Mas será esse o caso? Em primeiro lugar, não se pode dizer categoricamente, como Geirsson o faz, que as intuições geradas pelo ATGM não são intuições semânticas. O argumento proposto por H&T é um tipo de argumento comum em filosofia com aplicações em diferentes áreas. Trata-se de uma ferramenta constituída por dois passos principais: (i) parte de intuições sobre o desacordo e (ii) extrai conclusões semânticas. A ideia é algo como: dois indivíduos estão em desacordo sobre tal e tal; se os termos usados na discussão tivessem conteúdos semânticos diferentes, então não poderia haver desacordo, pois suas proposições não seriam

incompatíveis; sendo assim, os termos não possuem conteúdos semânticos diferentes. Denomino a relação entre os dois passos desse tipo de argumento de *Tese da Conexão*: se dois falantes expressam um desacordo genuíno, então os termos usados em tal disputa possuem significado similar. Note que há uma conexão entre desacordo e conteúdo semântico. Portanto, o ATGM, ao contrário do que sustenta Geirsson, trata de intuições semânticas também.

Em segundo lugar, o ATGM parece funcionar muito bem mesmo que não se faça referência alguma ao ATG. (É importante ressaltar que esta tese pode ser considerada a respeito de qualquer uma das objeções a H&T que estou considerando no presente capítulo). Suponhamos que o ATG não existisse. O ATGM ainda assim funcionaria? Se não, então ficaria claro que há uma interdependência entre o ATGM e o ATG. Neste caso, se houvessem dissimilaridades relevantes entre os dois argumentos isso realmente poderia implicar em problemas para o argumento desenvolvido por H&T. Se sim, isto é, se o ATGM funcionasse mesmo em caso da não existência do ATG, então estaria claro que não há interdependência entre os dois argumentos. Portanto, não faria sentido atacar o ATGM apontando para supostas dissimilaridades. Acredito que temos razão para aceitar a segunda alternativa.

Como afirmei no primeiro capítulo, o ATGM tem dois precursores fundamentais, o AQA e o *Argumento do Desacordo*, de Hare. Na verdade, o ATGM pode ser visto como uma reformulação do argumento de Hare estendido a versões mais contemporâneas do naturalismo moral. Assim, é possível construir o ATGM sem fazer menção sequer ao experimento de Putnam. Considere tal hipótese brevemente. Suponha que há duas comunidades, *A* e *B*. A comunidade *A* aceita uma teoria moral de primeira ordem *x* e a propriedade R-relacionada ao uso dos termos morais de seus habitantes é *N*. A comunidade *B* aceita uma teoria moral de primeira ordem *y* e a propriedade R-relacionada ao uso dos termos morais de seus habitantes é N_1. *N* e N_1 são propriedades diferentes. Os habitantes de ambas as comunidades usam os termos morais para aprovar ações, pessoas e instituições; estão dispostos a agir de acordo com os juízos em que os termos morais ocorrem; empregam tais termos em discussões relacionadas ao bem-estar e assim por diante. Suponha que os membros das duas comunidades se engajassem numa discussão sobre o status moral de determinada ação. Membros de *A* sustentam que "tal ação é correta-*n*" enquanto membros de *B* sustentam que "tal ação não é correta-n_1". Haveria desacordo genuíno? Da mesma forma que na formulação original do ATGM, se o RMN for verdadeiro, então não pode haver desacordo genuíno, pois os falantes estão predicando propriedades diferentes sobre a ação em questão. Mas a IUS parece ser mantida, isto é, da mesma forma que não formulação de H&T, parece que os membros das duas comunidades

estão em desacordo, caso em que o conteúdo semântico dos termos morais deve ser similar. Portanto, temos evidência contrária à plausibilidade do RMN.

Eis uma versão muito simplificada do ATGM. Mas o ponto importante aqui é que se repete os pontos relevantes tendo como pano de fundo o *Argumento do Desacordo*, de Hare, e não o ATG, de Putnam. E, mesmo assim, o argumento parece preservar o conteúdo do ATGM. Então, por que a desanalogia com o ATG seria um problema para o ATGM se este último persiste mesmo sem fazermos menção ao primeiro?

3.2. Conteúdo psicológico

Outro suposto ponto de desanalogia entre os experimentos de pensamento apontado por Geirsson é o seguinte. No ATG embora os habitantes de T e TG tenham conteúdo psicológico similar a respeito de 'água', isso não tem implicações para a referência, quer dizer a referência de 'água-*t*' continua sendo H_2O e a referência de 'água-*tg*' continua sendo XYZ (na verdade, este é o principal resultado que Putnam quer extrair com o experimento). Já no ATGM, sustenta Geirsson, acontece o contrário. H&T partem do pressuposto de que o conteúdo psicológico dos falantes a respeito dos termos morais *não* é similar (pois habitantes de T e TG tem teorias diferentes sobre a formulação de seus respectivos conceitos morais) e concluem que, mesmo assim, tais termos possuem o mesmo significado (GEIRSSON, 2014, p. 101-102). Geirsson afirma que tal desanalogia com o experimento de Putnam é significante, pois:

> Em primeiro lugar, dado que Horgan e Timmons repetidamente afirmam estar se baseando nas mesmas intuições do experimento mental de Putnam, os detalhes relevantes dos dois experimentos mentais precisam ser os mesmos. Um desvio sério como esse ameaça qualquer comparação significante entre os dois experimentos de pensamento. Em segundo lugar, e mais importante, deixando de lado qualquer comparação com a Terra Gêmea de Putnam, o movimento ameaça a principal conclusão que Horgan e Timmons querem extrair. Horgan e Timmons, afirmam que, diferentemente da Terra Gêmea de Putnam, a Terra Gêmea Moral nos faz concluir que os termos relevantes *não* diferem em significado. Dado o modo como eles constroem o experimento, é difícil ver como podemos concluir isso. Pois o equilíbrio moral desempenha um papel na formação das teorias e conceitos morais relevantes e, dado que é *assumido* que os Terráqueos aceitam uma teoria teleológica enquanto os habitantes da Terra Gêmea Moral aceitam uma teoria deontológica, é assumido desde o início que os termos morais na Terra e na Terra Gêmea Moral diferem em significado estrito, isto é, a parte do significado que reside na cabeça. [...] Dado isso, parece claro que Horgan e Timmons não podem concluir, tal como fazem, que os termos morais na Terra e na Terra Gêmea Moral possuem o mesmo significado. (GEIRSSON, 2014, p. 101-102).

O segundo ponto levantado por Geirsson é o mais importante. Ele sustenta que H&T não podem extrair a principal conclusão que seu argumento visa extrair, a saber, que os termos morais em T e TG têm o mesmo significado, pois assumem de antemão que tais termos têm significado diferente na medida em que pressupõem diferentes teorias de primeira ordem. Em outras palavras, dado que H&T partem do pressuposto de que o uso de 'correto-*t*' e 'correto-*tg*' é causalmente regulado por propriedades naturais diferentes, não podem concluir há um significado comumente partilhado.

Geirsson está ignorando detalhes importantes do experimento de H&T. Em primeiro lugar, não é totalmente correto afirmar que, ao contrário do ATG em que os habitantes de T e TG têm estados mentais similares sobre 'água' (incolor, que corre nos rios e lagos etc.), no ATGM os habitantes têm conteúdos mentais diferentes. Isso porque os termos morais nos dois planetas possuem certas condições que são características intrínsecas aos termos morais, tais como: tanto em T quanto em TG os termos morais são usados para aprovar ações, pessoas e instituições; os habitantes dos dois planetas usam os termos morais para discutir assuntos relacionados ao bem-estar; os habitantes dos dois planetas são internamente motivados a agir de acordo com os juízos em que os termos morais aparecem etc. Esses são detalhes importantes do ATGM que não podem ser ignorados.

Em segundo lugar, dizer que H&T não podem concluir que os termos morais têm significado similar porque assumem de antemão que tais termos possuem referências diferentes é meramente repetir o que o naturalista teria que aceitar e ignorar a parte principal do argumento, a saber: que há uma intuição, ampla e persistente, de que há desacordo genuíno entre habitantes de T e TG e que isso fornece evidência contra o pressuposto inicial de que os termos morais não possuem significado similar. No ATGM, de fato, assume-se que os termos morais possuem referências distintas e que, por isso, deveríamos concluir que não há univocidade semântica entre 'correto-*t*' e 'correto-*tg*', *se o RMN fosse verdadeiro*. O ponto do ATGM é que não poderíamos tirar conclusões em favor da univocidade semântica somente se assumíssemos o RMN. No entanto, segue o argumento, temos evidência para supor que o RMN é falso (IUS). E, portanto, temos razão para concluir que 'correto-*t*' e 'correto-*tg*' possuem significado similar. Geirsson está ignorando esse detalhe importante do ATGM. Assim, podemos concluir que apelar para uma suposta desanalogia entre o ATG e o ATGM em relação ao conteúdo psicológico dos membros das duas comunidades também não nos dá razão suficiente para rejeitarmos o desafio semântico de H&T.

4. Conclusão

Podemos concluir o presente capítulo retomando o ponto e as conclusões básicas extraídas. Há uma linha de ataque ao problema semântico do RMN que, embora seja não ortodoxa – no sentido de que não busca refutar alguma das três premissas que compõem o ATGM em específico – é importante ser considerada inicialmente. Tal estratégia pressupõe que há algum tipo de dependência entre o experimento mental desenvolvido por H&T e o experimento mental desenvolvido por Putnam, de modo que, se as analogias entre os dois não forem preservadas, o primeiro não obtém sucesso. Nesse primeiro capítulo, tentei argumentar que essa tentativa de recusar o ATGM tem algumas falhas e que, portanto, o ATGM possui independência e autonomia em relação ao argumento original de Putnam.

A partir de agora, passarei a considerar ataques mais diretos ao ATGM, isto é, a suas premissas específicas.

CAPÍTULO 3 – DESACORDOS GENUÍNOS

1. Introdução

A P2 do ATGM nos diz o seguinte.

> P2. Se 'correto-*t*' expressa um conteúdo que é diferente do conteúdo expresso por 'correto-*tg*', então Terráqueos e Terráqueos Gêmeos não expressam um desacordo substantivo genuíno quando um diz '*x* é correto-*t*' e o outro diz '*x* não é correto-*tg*', em que ambos utilizam '*x'* para se referir à mesma ação.

A ferramenta que está por detrás desse raciocínio é o que denominarei de *Condição da Identidade da Extensão*.

> *Condição da Identidade da Extensão* (*CIE*): Dois interlocutores, S_1 e S_2, estão em desacordo genuíno sobre se um particular é um membro de um tipo K se, e somente se, S_1 usa o termo K_1 para se referir a K e S_2 usa o termo K_2 para se referir a K.[22]

A ideia é muito simples. Primeiro, considere um exemplo sobre nomes próprios. Se eu, referindo-me a Ronnie O'Sullivan, digo 'Ronnie é sete vezes campeão mundial de sinuca inglesa' e você, se referindo a Ronnie, o chaveiro do seu bairro, diz 'Ronnie não é sete vezes campeão mundial de sinuca inglesa', embora à primeira vista pareça estarmos em desacordo, na verdade não estamos, pois nossas afirmações expressam proposições diferentes, de modo que a sua proposição não nega o valor de verdade da minha proposição, ou vice -versa, e ambas podem ser verdadeiras. Agora, considere um exemplo sobre tipos, a que a CIE se refere propriamente. Imagine que estamos discutindo se determinado particular é um membro do tipo *tigre*. Eu digo '*x* é um tigre' e você diz '*x* não é um tigre'. A ideia expressa pela CIE é que estaremos discordando genuinamente se, e somente se, usamos o termo 'tigre' para se referir ao tipo *tigre*. Se eu, por alguma razão, aprendi que tigres são, na verdade, aqueles indivíduos que classificamos como leões e uso 'tigre' para referir-me a leões, então o desacordo não será possível. Eu estarei dizendo algo como '*x* é um leão' e você estará dizendo '*x* não é um tigre'.

[22] K1 e K2 podem ser, obviamente, o mesmo termo.

As proposições não são inconsistentes. Neste sentido, o que P2 faz é colocar em prática essa condição. A ideia é que, se o RMN) é verdadeiro e o conteúdo semântico de um predicado moral é captado exaustivamente por uma propriedade natural *N*, então dois indivíduos não podem estar em desacordo sobre se uma ação particular é parte da extensão de tal predicado moral se eles se referem a propriedades naturais diferentes (pois estariam se referindo a tipos diferentes, a saber, 'correto-*t*' e 'correto-*tg*').

No entanto, prossegue o ATGM, quando consideramos o cenário acima, realmente parece que habitantes de T e TG discordam genuinamente (IUS). Dado que esta intuição é amplamente compartilhada e persistente, deve ser preservada. Mas, como vimos, se o RMN é verdadeiro, então o conteúdo semântico de 'correto-*t*' e de 'correto-*tg*' é captado exaustivamente por *N* e *N**, respectivamente, o que implica, pela CIE, que eles não discordam. Assim, a conclusão negativa de H&T é:

(a) Conclusão Negativa do ATGM: o Realismo Moral Naturalista é falso.

Mas o propósito de H&T é nos entregar algo positivo também. E isso é o resultado da melhor explicação para a IUS. Eles sustentam que o melhor candidato à explicação da IUS é o expressivismo moral. De acordo com H&T, julgamos que T e TG estão engajados num desacordo moral genuíno porque, por mais que assumamos que 'correto-*t*' e 'correto-*tg*' tenham um conteúdo descritivo diferente (*N* e *N**), o significado primário desses termos é conativo (isto é, eles servem para avaliar, expressar sentimentos de aprovação ou desaprovação, i.e., atitudes não-cognitivas, sejam elas quais forem). Isso é o que há de comum a 'correto-*t*' e 'correto-*tg*', e que invoca a IUS. Assim, temos a conclusão positiva.

(b) Conclusão Positiva do ATGM: o expressivismo moral EM é verdadeiro.

Os filósofos que irei considerar neste capítulo aceitam a plausibilidade da IUS. Mas, eles não acham que isso nos força a aceitar (a) e (b). E isso só pode ser feito de uma forma: recusando P2. Ou seja, sustentando que dois indivíduos podem expressar um desacordo genuíno mesmo que os termos empregados não captem a mesma propriedade (ou não tenham a mesma extensão, referência). Em outras palavras, defendendo que é possível haver desacordo genuíno mesmo que as sentenças empregadas pelos falantes não expressem proposições conflitantes. De outro modo ainda, apresentando casos que violem a CIE.

As estratégias adotadas aqui diferem. David Copp (2000) apela para a noção de *tradução*, e argumenta que, mesmo que os termos morais usados por T e TG expressem propriedades diferentes, ainda assim, pode ser o caso que um seja a *melhor* tradução para o outro, o que implicaria que há algo no significado desses termos que é suficientemente similar para garantir que as duas comunidades discordam genuinamente. David Merli (2002) defende que, embora se possa admitir que T e TG não expressem um desacordo moral genuíno, eles podem estar expressando um outro tipo de desacordo, um desacordo prático sobre *o que fazer*. Isso supostamente explicaria a plausibilidade da IUS sem ter que descartar o RMN. David Plunkett e Tim Sundell (2013) recorrem à noção de *negociação metalinguística* para defender que, embora os termos usados pelos participantes da disputa tenham significado literal diferente, eles expressam um tipo de desacordo genuíno, um *desacordo metalinguístico*.

Neste sentido, o que temos aqui são explicações concorrentes à explicação de H&T para a IUS. Tais explicações, diferentemente das estratégias consideradas no capítulo anterior que, aceitando P2, apelavam para algum tipo de viés que nos levava e IUS, fornecem diferentes modos pelos quais podemos recusar P2. Assim, o ponto a ser decidido é: essas explicações concorrentes realmente mostram a falsidade de P2? Se sim, então o ATGM não procede e não há desafio nenhum ao RMN.

Meu objetivo neste capítulo é tentar defender P2 desses ataques. Sustentarei que essas três estratégias não são suficientes para fornecer uma defesa promissora para o RMN contra o ATGM. Vou dividir este capítulo em quatro seções. Apesar de haver uma vasta literatura a respeito do ATGM e um dos pontos centrais deste tipo de argumento seja a suposição de desacordos entre comunidades hipotéticas, não há nenhum trabalho que discuta especificamente a noção de *desacordo*. Por isso, dedicarei a primeira seção para tentar esclarecer o que é essa relação. Apresentarei algumas definições alternativas e, finalmente, irei sugerir uma definição padrão para usarmos no decorrer deste trabalho. As seções seguintes serão dedicadas a cada um dos três ataques ao ATGM. Na segunda, apresentarei o *Argumento da Tradução* empregado por Copp e desenvolverei três argumentos na tentativa de sustentar que ele não representa uma ameaça ao ATGM. Na terceira seção, descreverei a *Réplica do Desacordo Prático* de D. Merli e argumentarei que ela é vulnerável a um dilema em que ambas as opções são desvantajosas para o seu defensor. Na quarta e última seção, irei apresentar e analisar a *Réplica da Negociação Metalinguística* de Plunkett e Sundell.

2. Desacordo

Quando os filósofos discutem P2 do ATGM, a pergunta que eles fazem é: T e TG expressam um desacordo moral genuíno quando um diz que '*x* é correto-*t*' e outro diz que '*x* não é correto-*tg*'? H&T e seus apoiadores respondem que sim. Os defensores do RMN *ou* respondem que não, que T e TG não expressam um desacordo moral genuíno, *ou* respondem que sim, mas que isso de alguma forma não afeta o RMN. No entanto, surpreendentemente, não há nenhuma discussão específica sobre o que significa o estar em desacordo. E isso é um problema porque muitas vezes os filósofos têm coisas diferentes em mente quando falam que indivíduos ou grupos discordam, de modo que tanto aqueles que sustentam que T e TG discordam genuinamente quanto aqueles que o negam podem estar corretos. Neste sentido, irei dedicar esta breve seção para discutir o que é 'desacordo'[23].

Como esclarece John MacFarlane (2014) 'desacordo' tem um significado de *estado* e um significado de *atividade*. Duas pessoas podem *estar* em desacordo ou *estar tendo* um desacordo. No primeiro sentido, S_1 pode estar em desacordo com S_2 mesmo que eles não se conheçam, não estejam juntos e mesmo que estejam localizados em intervalos temporais diferentes. É neste sentido que dizemos que você ou outro alguém discorda de Aristóteles, por exemplo, sobre a natureza da alma ou sobre o que constitui um agente virtuoso. Você e Aristóteles *estão em* desacordo. No segundo sentido, atribuímos a característica de desacordo a S_1 e S_2 se eles estão engajados em algum tipo de atitude um contra o outro. Eles estão envolvidos num tipo de atividade que requer a mesma localização espaço temporal. S_1 e S_2 *estão tendo* um desacordo sobre algo.

Parece que o estado de desacordo é mais fundamental que a atividade de desacordo, pois toda atividade pressupõe o estado, mas, como vimos, nem todo estado pressupõe atividade. Assim, o que deve nos interessar aqui é o *estado* de desacordo.

Em primeiro lugar, como poderíamos caracterizar a forma lógica dessa relação de estar em desacordo? Considere,

Desacordo 1: S_1 está em desacordo com S_2.

[23] Grande parte dos pontos que apresentarei nesta seção estão em John MacFarlane (2014).

Esta caracterização é geral o suficiente para captar qualquer instância em que haja desacordo. No entanto, essa relação é demasiado geral e se aplica a todos os indivíduos, já que ninguém concorda com tudo o tempo todo. Queremos algo mais específico, tal como:

> *Desacordo 2*: S_1 está em desacordo com S_2 sobre *p*.

Mas, dado que essa caracterização pressupõe que desacordo envolve conteúdo proposicional, ela não funciona para todos os tipos de desacordo. Há desacordos que envolvem atitudes não-doxásticas, tais como desejos ou preferências. Além disso, o elemento contextual também parece ser indispensável. Assim, a seguinte caracterização parece ser mais apropriada:

> *Desacordo 3*: S_1 está em desacordo com S_2 em virtude de S_2 sustentar *x* no contexto *c*.

Como afirma McFarlane, podemos omitir a referência a S_2, pois o contexto já inclui o agente do contexto. Então,

> *Desacordo 4*: S_1 está em desacordo a respeito de *x* no contexto *c*

parece ser mais apropriado. Tal definição não é demasiadamente geral para se aplicar a todos em qualquer contexto e não parece ser demasiadamente específico para excluir formas aparentemente genuínas de desacordo. Neste sentido, quando dizemos que dois indivíduos estão em desacordo, estamos dizendo que este tipo de relação lógica se aplica a eles.

Mas o ponto mais importante não parece ser apenas como devemos entender a relação de *estar em desacordo*, mas *o que é* essa relação. Aqui, como é frequente neste tipo de pergunta filosófica, é difícil termos uma definição conclusiva sobre o que é 'desacordo'. Consideremos algumas.

2.1. Visão Simples

Talvez a visão mais intuitiva sobre o desacordo, e normalmente implícita em muitos casos, seja o que podemos chamar de *Visão Simples do Desacordo* (MACFARLANE, 2014, p. 121).

> *Visão Simples do Desacordo*: discordar com a crença de alguém de que *p* é ter crenças cujo conteúdo é conjuntamente incompatível com *p*.

A incompatibilidade do conteúdo das crenças parece ser um ingrediente importante de qualquer instância em que ocorra um desacordo, de modo que esta visão parece ser bem plausível. Em geral, é amplamente aceita pelos filósofos como uma caracterização mínima do que significa estar em desacordo. No entanto, a *Visão Simples* não parece captar todo tipo de desacordo. Considere o seguinte exemplo (MACFARLANE, 2014, p. 122). Ned, o repórter do canal 4 que é responsável pelo clima, sustenta que há uma probabilidade de 0,7 (de 1,0, suponhamos) de que irá chover amanhã. Ted, o repórter do canal 5 que é responsável pelo clima, sustenta que há uma probabilidade de 0,8 de que irá chover amanhã. Note que o conteúdo da crença de Ned e Ted não parece ser incompatível, pois ambos acreditam que irá chover amanhã. Portanto, de acordo com a *Visão Simples*, não há desacordo. Mas é difícil aceitar que Ned e Ted não estão em desacordo. Parece haver algo entre eles que nos faz considera-los como estando em desacordo. Além disso, a *Visão Simples* está limitada a atitudes doxásticas apenas. E as pessoas não discordam apenas sobre crenças, mas sobre gostos, desejos ou preferências. Assim, a *Visão Simples do Desacordo* não parece ser suficiente.

2.2. Não-Cotenabilidade

MacFarlane sugere que a noção de *não-cotenabilidade* (*non cotenability*) capta melhor os casos de desacordo[24]. Dizemos que X é cotenável (*cotenable*) com Y se ~ (X → ~Y). E que X é não-cotenável com Y se X → ~Y. MacFarlane faz uso desta noção definindo o desacordo em termos de não-cotenabilidade do seguinte modo (Cf. MACFARLANE, 2014, p. 121):

> *Não-Cotenabilidade*: S_1 discorda com S_2 se não pode, coerentemente, adotar a mesma atitude de S_2 sem excluir algumas de suas próprias atitudes correntes.

De acordo com MacFarlane, muitos casos de desacordo são casos de não-cotenabilidade. Suponha que Tom acredita que todos os banqueiros são ricos. Dom acredita que Bob é um

[24] Não há uma tradução padrão para esse termo no português. Sugiro a tradução literal "não-cotenabilidade", mas uma alternativa poderia ser "não-sustentabilidade", já que a ideia é que duas entidades não podem ocorrer conjuntamente já que uma implica a negação de outra.

banqueiro pobre. Tom e Dom estão em desacordo, pois suas crenças são não-cotenáveis. Tom não pode, coerentemente, aceitar a crença de Dom sem mudar ou excluir sua atitude doxástica corrente. O mesmo vale para Dom. Note que esta noção capta o desacordo entre Ned e Ted que a *Visão Simples* teve dificuldades. Ned não pode coerentemente aceitar a crença de que há uma probabilidade de 0,8 de que irá chover amanhã sem excluir a sua própria crença de que há uma probabilidade de 0,7 de chover amanhã. O mesmo se aplica a Ted.

A ideia é que a *Não-Cotenabilidade*, diferentemente da *Visão Simples*, se aplique também a atitudes não-doxásticas, como desejos e preferências. Por exemplo, suponha que Jane gosta de Bob enquanto Sara o odeia. Parece que Jane e Sara estão num estado de desacordo. Mas não há nenhuma proposição *p* a respeito da qual elas discordam. É possível que elas creiam sobre todas as mesmas coisas a respeito de Bob. A *Não-Cotenabilidade* capta este tipo de desacordo, pois Jane não pode coerentemente adotar a atitude de Sara em relação a Bob sem excluir a sua própria atitude em relação a Bob, e vice-versa. Neste sentido, parece que a *Não-Cotenabilidade* é uma forte candidata para explicar suficientemente bem a relação de desacordo.

Mas consideremos o seguinte caso (Cf. MACFARLANE, 2014, p. 122). Suponha que Sara ama Bob e não há nenhuma outra pessoa no mundo com a qual ela gostaria de estar. Bob, do mesmo modo, ama Sara e também não há nenhuma outra pessoa no mundo com a qual ele gostaria de estar. Parece haver um perfeito estado de acordo aqui. No entanto, note que temos um caso de não-cotenabilidade prática. Isto é, Sara não pode adotar a atitude de Bob sem excluir a sua própria atitude corrente e Bob não pode adotar a atitude de Sara sem excluir a sua própria atitude corrente. Obviamente, não queremos que uma explicação sobre a relação de desacordo nos obrigue a aceitar que há desacordo onde está claro que não há. Sendo assim, pode-se querer adotar uma noção diferente de desacordo. Pelo menos em relação a atitudes não-doxásticas.

2.3. Impedimento da Satisfação Conjunta

Em *Facts and Values,* Charles Stevenson (1963) distingue entre dois sentidos de desacordo e em relação a atitudes não-doxásticas, ele diz que um desacordo "envolve uma oposição de atitudes, ambas as quais não podem ser satisfeitas ao mesmo tempo" (STEVENSON, 1963, p. 2). Ou seja, Stevenson define o desacordo entre atitudes não-doxásticas em termos da não satisfação mútua de tais atitudes. Podemos dizer, então, o seguinte:

> *Impedimento da Satisfação Conjunta*: S_1 discorda da atitude de S_2 se a satisfação da atitude de S_1 impede a satisfação da atitude de S_2.

Isso pode nos ajudar a entender por que Sara e Bob não estão em desacordo. A satisfação da atitude de Sara (amar Bob) não impede a satisfação da atitude de Bob (amar Sara).

Note que *Impedimento da Satisfação Conjunta* e *Não-Cotenabilidade* são noções distintas de desacordo. Para ver isso, considere o seguinte exemplo. Há um cupcake na mesa e Ned e Ted querem comê-lo. Os dois tem um desejo com o mesmo conteúdo: comer o cupcake. A atitude de ambos é cotenável, quer dizer, Ned pode adotar a atitude de Ted sem excluir a sua própria atitude corrente. No entanto, ambas as atitudes não podem ser satisfeitas ao mesmo tempo, pois a satisfação da atitude de Ned impede a satisfação da atitude de Ted e vice-versa. As atitudes de Ned e Ted estão em desacordo e a noção da *Impedimento da Satisfação Conjunta* capta isso enquanto que a *Não-Cotenabilidade* não.

No entanto, embora a noção de *Impedimento da Satisfação Conjunta* tenha um alcance explanatório onde a *Não-Cotenabilidade* não tem, não podemos aceita-la como uma explicação definitiva da relação e desacordo, pois ela está limitada a atitudes não-doxásticas. Além disso, ela está limitada à *atividade* do desacordo, e não ao estado do *desacordo*. Ou seja, essa visão se aplica somente a casos em que os sujeitos estão localizados no mesmo contexto espaço temporal.

Como vimos, a *Visão Simples* é limitada a crenças. A *Não-Cotenabilidade* é melhor, mas não capta certos casos de desacordo entre atitudes não-doxásticas. O *Impedimento da Satisfação Conjunta* pode captar o desacordo onde a *Não-Cotenabilidade* tem problemas, mas está limitado a atitudes não-doxásticas. Parece que se quisermos uma explicação mais abrangente sobre o desacordo precisamos de algo que se aplique tanto a crenças como a atitudes não-doxásticas e que evite os contraexemplos que temos visto até aqui.

Talvez a melhor proposta disponível seja a de Plunkett e Sundell (2013).

2.4. Desacordo Requer Conflito de Conteúdo

Queremos uma noção de desacordo que, além de evitar os contraexemplos das noções anteriores, dê conta de pelo menos duas características: (i) explique o desacordo tanto em casos de atitudes doxásticas quanto não-doxásticas; (ii) explique o desacordo que está em jogo tanto em trocas linguísticas quanto casos em que as trocas linguísticas não estão necessariamente

envolvidas. Ou seja, queremos uma noção que explique o desacordo tanto entre crenças quanto entre desejos ou preferências e que explique o desacordo enquanto em *estado* e não exclusivamente enquanto uma *atividade* em que duas pessoas estão engajadas num contexto.

A visão de Plunkett e Sundell parece ser promissora neste sentido. Indo direto ao ponto, eles definem 'desacordo' como algo que envolve, essencialmente, "alguma incompatibilidade (do tipo relevante) entre conteúdos (sejam quais forem) aceitos (no sentido relevante) por pessoas diferentes (que podem ou não estar em conversação uma com a outra)" (PLUNKETT & SUNDELL, 2013, p. 11). Plunkett e Sundell apresentam mais rigorosamente sua definição de 'desacordo' do seguinte modo:

> *Desacordo Requer Conflito de Conteúdo (DRCC)*: Se dois sujeitos, A e B, discordam um com o outro, então há alguns objetos *p* e *q* (proposições, planos etc) tal que A aceita *p* e B aceita *q*, e *p* é tal que as demandas impostas ao sujeito em virtude de aceitá-lo são racionalmente incompatíveis com as demandas impostas ao sujeito em virtude de aceitar *q* (PLUNKETT & SUNDELL, 2013, p. 11).

Essa visão sobre o desacordo dá conta das duas exigências iniciais. Note que a suposição de que as partes do desacordo podem ou não estar em conversação uma com a outra dá conta da nossa segunda exigência. E note, também, que quando Plunkett e Sundell se referem ao conteúdo do desacordo, e à aceitação destes conteúdos, dizendo que "sejam quais forem" e "no sentido relevante" eles deixam em aberto a questão sobre se tratar de crenças ou de atitudes não-doxásticas, e isso dá conta da nossa primeira exigência.

No que diz respeito aos limites deste capítulo, não podemos aceitar a *Visão Simples* porque ela não contempla atitudes não doxásticas, o que excluiria a possibilidade de desacordos genuínos para qualquer teoria não-cognitivista. O mesmo se aplica, de modo inverso, ao *Impedimento da Satisfação Conjunta*, que contempla desacordos em que apenas atitudes não-doxásticas estão envolvidas. A *Não-Cotenabilidade* e a *DRCC* parecem ser as mais apropriadas. Para os propósitos deste trabalho, irei propor que aceitemos o *DRCC* como a melhor caracterização do que é o desacordo, pois, além de não ser vulnerável ao problema que apresentamos à *Não-Cotenabilidade*, como veremos, o DRCC desempenha um papel central numa das réplicas que consideraremos na parte final.

3. D. Copp e o Argumento da Tradução

Em *Milk, Honey, and the Good Life on Moral Twin Earth* (2000), D. Copp apresenta duas réplicas para sustentar que o ATGM não representa um desafio ao RMN. Com a primeira réplica ele mantém que, mesmo que os predicados morais usados pelos habitantes de T e TG expressem propriedades diferentes, isso não implica que o *significado* desses termos é diferente, o que seria suficiente para garantir que os habitantes das duas comunidades podem expressar um desacordo genuíno. A segunda, e mais ambiciosa réplica, consiste em fornecer uma teoria metassemântica na tentativa de mostrar que os termos morais em T e TG não expressam propriedades diferentes. Cada réplica funciona de maneira independente. Como aqui estou preocupado com o tipo de ataque ao ATGM que nega P2, vou considerar a primeira réplica apenas (considerarei a segunda, nos capítulos seguintes).

Em sua crítica ao ATGM, Copp concede a verdade de P3 a H&T, isto é, ele aceita que a IUS é "robusta e amplamente compartilhada" (COPP, 2000, p. 119). Mas ele acredita que o naturalista não está em problemas, pois pode fornecer uma explicação alternativa para P3 atacando o pressuposto de P2, a saber, a ideia de que dois falantes não discordam genuinamente se usam sentenças com conteúdo semântico diferente. Ele acredita que é possível fazer isso mostrando que, embora 'correto-*t*' e 'correto-*tg*' expressem propriedades diferentes e, portanto, tenham significado diferente no sentido "filosoficamente preferido", há um sentido em que os dois predicados morais tem o "mesmo significado" (COPP, 2000, p. 122).

Assim, o ponto central da primeira objeção de Copp, que chamarei de *Argumento da Tradução*, é que "os termos morais podem ser a melhor *tradução* para os termos morais gêmeos correspondentes" (COPP, 2000, p. 121). Isso supostamente seria evidência para o fato de que 'correto-*t*' e 'correto-*tg*', tem "*significado* similar", a despeito de referirem propriedades diferentes (COPP, 2000, p. 124). Copp acredita que essa comonalidade no significado seria suficiente para garantir que membros de T e TG podem obter sucesso em discordar genuinamente. Neste caso, dois falantes poderiam expressar um desacordo, mesmo usando sentenças com conteúdo semântico distinto.

As passagens em que o argumento se encontra são as seguintes.

> Há convergência significativa entre as extensões dos termos morais e dos termos morais gêmeos correspondentes ... [porque] consequencialistas e deontologistas concordam em muitos casos sobre quais ações são "certas" e "erradas" (COPP, 2000, p. 121).

> [...] é necessário entender que a tradução é mais similar a tentar encontrar alguém que seja suficientemente parecido com você ... do que tentar encontrar o seu gêmeo idêntico. [...] Portanto, é possível que o termo 'correto' do português seja a melhor tradução, para o português, do termo 'correto' do português gêmeo, mesmo que, por estipulação, os termos expressem propriedades diferentes. Se isso for o caso, então mesmo que os termos morais e os termos morais gêmeos correspondentes expressem propriedades diferentes, e, portanto, tenham "significados" diferentes no sentido filosoficamente preferido do termo, há também um sentido em que eles podem ter o mesmo "significado" (COPP, 2000, p. 121-122).

Primeiro, note que o argumento consiste, basicamente, em dizer que as teorias de primeira ordem que captam a referência dos termos morais possuem convergência "significativa" e, além disso, chamar atenção para uma peculiaridade da tradução, isto é, que uma tradução não requer pares idênticos, mas suficientemente próximos. Disso, Copp conclui que, embora 'correto-*t*' e 'correto-*tg*' expressem propriedades diferentes, estes termos possuem 'significado similar'. Com isso, temos o que irei chamar de Tese Referência/Significado (TRS).

> *Tese Referência/Significado* (TRS): *diferença na referência não implica diferença no significado*.

Em outras palavras, a TRS afirma que dois termos com o mesmo significado podem ter referências diferentes. A conclusão mais geral para o argumento de Copp, e implícita aqui, é de que se os predicados morais das duas comunidades têm "significado similar", então T e TG podem expressar um desacordo moral genuíno quando um diz '*x* é correto-*t*' e outro diz '*x* não é correto-*tg*', mesmo que os termos usados se refiram a propriedades diferentes. Isso, obviamente, é um desafio a P2 do ATGM, pois esta premissa nos diz que T e TG *não* expressam um desacordo genuíno se seus predicados morais possuem referências distintas.

No intuito de melhor identificar as premissas e conclusões do argumento de Copp, apresento-o na forma canônica.

> P1: A referência de 'correto-*t*' é rastreada por uma teoria de primeira ordem consequencialista (e refere a propriedade de *maximizar o agregado de felicidade*) e a referência de 'correto-*tg*' é rastreada por uma teoria de primeira ordem deontológica (e refere a propriedade de *tratar os outros como fins em si mesmos*).
> P2: A propriedade de *maximizar o agregado de felicidade* é diferente da propriedade de *tratar os outros como fins em si mesmos*.

C1: Logo, a referência de ‘correto-*t*’ e a referência de ‘correto-*tg*’ é diferente (P1, P2).

P3: Consequencialismo e deontologismo são significativamente convergentes.

C2: Se consequencialismo e deontologismo são a teoria de primeira ordem que rastreiam a referência de “correto-*t*’ e ‘correto-*tg*’ (P1), e são significativamente convergentes (P3), então a referência de ‘correto-*t*’ e de ‘correto-*tg*’ é significativamente convergente.

P4. Uma tradução aceitável consiste em encontrar pares de termos que sejam similares ‘o suficiente’ (isto é, que tenham “significado similar”), e não pares de termos que sejam idênticos.

C3. Se a referência de ‘correto-*t*’ e de ‘correto-*tg*’ é significativamente convergente [C2], então tais termos são traduzíveis um pelo outro [P4].

C4. Se ‘correto-*t*’ e ‘correto-*tg*’ são traduzíveis um pelo outro [C3], então, embora expressem propriedades diferentes [P1, P2], tem “significado” similar [P4].

Note que Copp não recusa a IUS. O que ele faz é fornecer um argumento na tentativa de mostrar que o RMN não é inconsistente com a IUS. Ele acredita que a TRS melhor explica a plausibilidade da IUS. Note, também, que Copp admite que ‘correto-*t*’ e ‘correto-*tg*’ não tem o mesmo significado no sentido “filosoficamente preferido”, mas sustenta, a partir da suposta convergência e, portanto, traduzibilidade, na extensão destes termos, que há um sentido em que ‘correto-*t*’ e ‘correto-*tg*’ têm o “mesmo significado”. Em nenhum momento fica claro qual é este outro sentido de “significado” que os predicados morais compartilham, e isso pode ser uma baixa no argumento de Copp. Mas a sua ideia é que se essa noção de “mesmo significado” em termos de traduzibilidade é suficiente para acomodar a IUS, então o naturalista tem uma boa via de escape do ATGM.

Ao analisar o *Argumento da Tradução*, talvez a primeira coisa que soe estranha é a ocorrência de C1 e C2. C1 nos diz que a referência de ‘correto-*t*’ e a referência de ‘correto-*tg*’ é diferente. Adiciona-se uma premissa apenas e a conclusão seguinte, C2, nos diz que a referência de ‘correto-*t*’ e a referência de ‘correto-*tg*’ é significativamente convergente, isto é, *pelo menos parte*, a referência não é diferente. O argumento de Copp é difícil de entender devido a essa parte e pode parecer que estou interpretando-o mal. Mas parece que ele realmente se compromete com C1 e C2. Em primeiro lugar, ele mantém que os predicados morais

expressam propriedades diferentes; isso constitui a primeira parte da TRS (C1). Em segundo lugar, uma das premissas fundamentais do seu argumento é que as teorias que rastreiam a referência dos predicados morais convergem significativamente (C2).

Poder-se-ia pensar que há uma contradição aqui. Mas seria um erro. É plausível dizer isso. Deve-se notar que aqui trata-se da referência de *tipos*, e não de nomes próprios, por exemplo. Portanto, a referência dos termos não é unívoca, como em 'Aristóteles' que refere a um único indivíduo. Assim, a extensão admite um conjunto de entidades que a compõem. E parte deste conjunto de entidades pode ser compartilhado por diferentes termos que expressam tipos. Considere o seguinte exemplo. Os subtipos 'tigre de sumatra' e 'tigre siberiano', num sentido, possuem referência diferente, senão não seria possível individuar as duas espécies. Mas noutro sentido, a referência converge em uma série de instâncias. Ambas as espécies instanciam a propriedade de *ter quatro patas*, *ter um coração*, *ter garras*, *ser da família dos felinos* etc. Como um dos principais fatores que os diferencia é o tamanho e o peso, pode-se dizer que a referência destes termos não converge quando se trata da instanciação da propriedade de, em idade adulta, *ter peso entre 200 e 310 kg*, propriedade esta que é instanciada apenas pelo tigre siberiano. Portanto, quando Copp assume que a referência de 'correto-*t*' e 'correto-*tg*' é diferente, mas converge em algum sentido, ele não está cometendo nenhuma inconsistência. Repare que ele fala em "significant overlap" entre os predicados morais. Isso é mais fácil de observar no caso dos predicados morais. É muito difícil manter que uma ação em que a honestidade ou a igualdade é realizada não seria instanciação da propriedade de *maximizar o agregado de felicidade* assim como da propriedade de *tratar os outros como fins em si mesmos*.

A partir de agora, apresentarei três argumentos com o intuito de sugerir que a TRS não fornece uma defesa promissora do RMN contra o ATGM. Como veremos, Copp não apresenta um contraexemplo a P2 do ATGM, sua explicação alternativa para a IUS não é a *melhor* explicação e sua abordagem compromete o defensor do RMN com um custo indesejado, o relativismo metaético. Consideremos cada um dos argumentos detalhadamente.

3.1. Argumento 1: ausência de contraexemplos a P2

O primeiro argumento que gostaria de expor é sobre como a não inconsistência entre C1 e C2 produz uma ambiguidade, como Copp se apropria dessa ambiguidade a seu favor e porque isso não constitui evidência contra P2 do ATGM. A ambiguidade gerada por essas duas premissas é que 'correto-*t*' e 'correto-*tg*' possuem referência diferente (já que expressam

propriedades diferentes), e, *em parte*, convergente (já que essas propriedades, consequencialista e deontológica, convergem num grande número de instâncias). O que a TRS nos diz é que diferença na referência não implica em diferença do significado, o que, em outros termos, é o mesmo que dizer que dois termos podem ter o mesmo significado embora tenham referências diferentes. Assim sendo, a TRS tem duas partes constitutivas centrais: (i) dois termos, digamos t_1 e t_2, tem referências distintas e (ii) t_1 e t_2 tem significado similar (isso está explícito em C1 e C4). Mas, para sustentar essa segunda parte, isto é, que esses termos têm significado similar, Copp apela para a convergência da referência entre t_1 e t_2 (como está em P3 e C2). Quer dizer, a estranheza aqui é que Copp argumenta a favor da tese de que a diferença na referência não implica diferença no significado notando que as referências dos termos em questão não são diferentes, ou melhor, não são *tão* diferentes.

Isso sugere que, como está, a TRS é incompleta. O que permite a Copp manter que t_1 e t_2 tem significado similar, é a convergência da referência entre t_1 e t_2. Dizer que tais termos possuem referência convergente, é dizer que a referência de tais termos, pelo menos em parte, *não* é diferente. Mas, ao final, a TRS ignora esta parte sobre a convergência (que é fundamental para a parte sobre o "significado similar" fazer sentido) e mantém que t_1 e t_2 tem referências distintas e significado similar.

Em outras palavras, o que estou tentando apontar é o seguinte: o que somos forçados a aceitar ao admitir C1 e C2 é que t_1 e t_2 tem referência diferente e, em parte, convergente; mas a TRS nos diz apenas que t_1 e t_2 tem referência diferente, ignorando, em seu resultado final, a convergência. Mas não se pode ignorar a convergência, pois a segunda parte da TRS (sobre o significado similar) não se segue. Do modo como está, o conteúdo da TRS é: t_1 e t_2 tem referência diferente e significado similar. Mas, seu real conteúdo é: t_1 e t_2 tem referência diferente, *e em parte convergente*, e significado similar.

Agora, notemos a consequência disso. O objetivo de Copp é mostrar que dois interlocutores podem expressar um desacordo genuíno, mesmo que eles usem sentenças com conteúdo semântico diferente. Isso é o que mostraria a falsidade de P2 do ATGM. Como vimos, o que confere plausibilidade a P2 do ATGM é o raciocínio de que dois interlocutores não podem expressar um desacordo genuíno se a extensão dos predicados por eles empregados for diferente. Neste sentido, se se quiser mostrar que P2 é falsa, então é necessário apresentar um caso que viole essa condição; isto é, um caso em que tenhamos dois indivíduos usando termos com extensões diferentes e, mesmo assim, expressando um desacordo genuíno. Mas será que Copp apresenta um caso desse tipo?

À primeira vista, pode parecer que sim, pois a TRS nos diz que podemos ter predicados com referência distinta e significado similar e essa comunalidade do significado seria suficiente para garantir que dois indivíduos discordam genuinamente. No entanto, um olhar rigoroso revela que Copp não apresenta um caso em que dois indivíduos estão usando termos com extensão diferente e estão em desacordo, violando, assim, P2 do ATGM. Quando analisamos cuidadosamente o que está envolvido na TRS, percebemos que a convergência da referência é um ingrediente essencial para que a comunalidade semântica se mantenha, como vimos acima. Portanto, o que Copp nos apresenta é um caso em que dois indivíduos estão usando termos com referências diferentes e, *em parte*, convergentes, e é esta parte convergente que garante a similaridade semântica. Em outras palavras, ele não nos dá um contraexemplo à P2, mas evidencia a não indispensabilidade da univocidade semântica para a ocorrência de desacordos genuínos. Assim sendo, podemos concluir que ele não mostrou a falsidade de P2 do ATGM.

3.2. Argumento 2: a melhor explicação para a IUS

O segundo argumento é sobre P3 do *Argumento da Tradução*. Como vimos, Copp sustenta que habitantes de T e TG discordam porque 'correto-*t*' e 'correto-*tg*' tem significado similar e o que garante que tais predicados tem significado similar, na sua visão, é a convergência entre consequencialismo e deontologismo. Além disso, tal convergência constitui a sua hipótese explanatória para a IUS, pois, argumenta Copp, a IUS é amplamente compartilhada devido ao fato de que 'correto-*t*' e 'correto-*tg*' convergem significativamente. Sendo assim, P3 é a principal premissa do *Argumento da Tradução*. Aqui o problema para Copp é duplo. Em primeiro lugar, o ATGM não depende do que está contido em P3 (do ATGM), como ele supõe. E em segundo lugar, mesmo se dependesse, a sua hipótese explanatória não parece ser a *melhor* explicação para a IUS.

H&T recorrem ao consequencialismo e ao deontologismo para expressar a ideia de que os predicados morais de T e TG se referem a propriedades diferentes. No entanto, tal recurso é meramente ilustrativo. O ATGM funciona da mesma forma mesmo se supormos que as teorias de primeira ordem em questão são outras. Ou, mesmo que não identifiquemos nenhuma teoria de primeira ordem. Basta supormos que os termos morais usados pelos habitantes de T e TG expressam propriedades diferentes. A especificação das teorias serve apenas para preencher melhor os detalhes do ATGM e é, portanto, dispensável. Copp ainda poderia argumentar que P3 (do ATGM) também não depende exclusivamente da convergência

entre consequencialismo e deontologismo, mas apenas do fato de que, seja qual forem as duas teorias de primeira ordem em questão, elas convergem. Embora esse tipo de resposta o livre do problema inicial, ainda assim, parece que a hipótese de Copp não é a *melhor* explicação para a IUS.

Considere o que chamo de Condição do Sucesso Explanatório (CSE).

> *Condição do Sucesso Explanatório (CSE)*: A obtêm sucesso em explicar o fenômeno B se a ausência de A torna a ocorrência de B altamente improvável.

A CSE parece ser um princípio indispensável para decidirmos sobre o que conta como boa explicação. Esta condição certamente dependerá do contexto para determinarmos mais precisamente o que conta como "altamente improvável". Por exemplo, imagine que o fenômeno a ser explicado é de que a grama está molhada e a minha hipótese explicativa é de que choveu. O fato de que choveu será uma boa explicação para o fato de que a grama está molhada se a não ocorrência da chuva torna improvável o fato de que a grama está molhada. É claro que alguém poderia ter molhado a grama ou poderia ter caído neve e derretido, o que implicaria que a ausência da minha hipótese explanatória não tornaria improvável a ocorrência do fenômeno. Mas cabe ao contexto determinar essas características. Por exemplo, se soubéssemos que não neva nesse lugar e que ninguém esteve presente para molhar a grama, a hipótese de que choveu ganha força. E se, tendo esses detalhes contextuais, o fenômeno da grama molhada for provável mesmo se supusermos a ausência da explicação de que choveu, então tal explicação não é uma boa explicação.

Agora, considere a hipótese explanatória de Copp para a IUS. Ele diz que o que melhor explica a IUS é o fato de que os predicados morais usados pelos membros de T e de TG possuem referência convergente. Sendo assim, a convergência da referência obtém sucesso em explicar a IUS se a sua ausência (da convergência da referência) torna a ocorrência da IUS altamente improvável. Em outras palavras, se a ocorrência da IUS for altamente provável mesmo na ausência da convergência da referência, então essa hipótese explanatória não é uma boa explicação para a IUS. Penso que é possível construir casos em que a hipótese explanatória de Copp está ausente (e, portanto, não deveríamos ter a ocorrência da IUS ou, pelo menos, sua ocorrência deveria ser altamente improvável) e, mesmo assim, há a ocorrência da IUS. Se esses casos forem plausíveis, então teremos evidência para acreditar que a explicação de Copp não obtém sucesso.

Considere a seguinte variação do experimento mental de H&T. Novamente, vamos supor a existência de duas comunidades Terra (T) e Terra Gêmea Moral Escravagista (TGME). Os habitantes de T acabaram adotando uma teoria de primeira ordem consequencialista C*. A referência de 'correto-t' é determinada pela propriedade N, que é rastreada por C*, e entre o conjunto de ações que fazem parte da referência de 'correto-t' está o requerimento de que não é correto submeter pessoas ao trabalho forçado e não remunerado. Os habitantes de TGME acabaram adotando uma teoria moral de primeira ordem D*. A referência de 'correto-*tgme*' é determinada pela propriedade *N*, que é captada por D*, e entre o conjunto de ações que fazem parte da referência de 'correto-*tgme*' está o requerimento de que é correto submeter pessoas ao trabalho forçado e não remunerado. Agora, imagine que indivíduos das duas comunidades se encontram e os habitantes de T dizem 'não é correto-*t* submeter pessoas ao trabalho forçado e não remunerado' enquanto que os habitantes de TGME sustentam que 'é correto-*tgme* submeter pessoas ao trabalho forçado e não remunerado*"*.

Qual é a nossa intuição sobre este caso? Há desacordo moral genuíno? Assim como no experimento original do ATGM, parece plausível que há desacordo entre membros de T e TGME. Para tornar isso mais explícito, basta imaginarmos uma acalorada discussão em 1850 entre um defensor da escravatura e um defensor da abolição. Soaria muito estranho dizer que não havia um desacordo moral genuíno entre tais indivíduos.

Agora, note que não parece haver "convergência significativa" entre as teorias de primeira ordem adotadas em T e em TGME. Portanto, a hipótese explanatória reivindicada por Copp está ausente. Mas note, também, que a IUS está presente. Portanto, temos um caso em que a hipótese explanatória proposta por Copp está ausente e, mesmo assim, a ocorrência do fenômeno a ser explicado não é altamente improvável. Assim, pela CSE, temos que concluir que a explicação fornecida por Copp não é a *melhor* explicação.

Mas qual seria, então, a melhor explicação para a IUS? Endosso o *insight* positivo do ATGM: o que melhor explica a intuição de que habitantes de T e TG discordam é o fato de que termos morais possuem um conteúdo conativo que sempre é primário. Tal conteúdo é sempre mantido embora os falantes possam variar amplamente sua compreensão sobre o que recai sobre termos como 'bom', 'certo' e 'errado'[25].

[25] Obviamente, isso requer desenvolvimentos. Por exemplo, que tipo de conteúdo conativo os termos morais possuem? Há diversas teorias não-cognitivistas que propõem respostas diferentes a essa questão. No entanto, não é o ponto aqui estabelecer qual temos mais razão para aceitar. Deve apenas ficar claro que estou assumindo que alguma abordagem não-cognitivista melhor explica a IUS.

3.3. Argumento 3: um custo indesejado

Numa resposta ao artigo de Copp, H&T (2000) sugeriram que a adoção da TRS implica num alto custo para o RMN. Ao assumir a tese de que predicados morais podem ter referência distinta e significado similar, o naturalista parece acabar tendo que aceitar algum tipo de relativismo metaético, algo que sua posição realista definitivamente pretende evitar.

É difícil fornecer uma caracterização para o relativismo metaético que faça justiça a todos os seus defensores, mas a seguinte definição fornecida por C. Gowans[26] parece ser geral o bastante para captar o que a maioria entende por relativismo metaético.

> *Relativismo Moral Metaético (RMM)*: a verdade ou falsidade dos juízos morais, ou sua justificação, não é absoluta ou universal, mas é relativa às tradições, convicções ou práticas de um grupo de pessoas.

Verdade e justificação são os conceitos principais desta definição, embora a justificação nem sempre faça parte de uma teoria relativista. Sobre a verdade, o RMM implica que juízos morais podem ser verdadeiros em relação aos padrões normativos de uma sociedade, mas falsos em relação aos padrões normativos de outra. Isso significa que juízos morais como 'O casamento entre pessoas do mesmo sexo é correto' e 'O casamento entre pessoas do mesmo sexo não é correto' não são necessariamente inconsistentes, pois ambos podem ser verdadeiros ao mesmo tempo, desde que tenham padrões normativos referentes a contextos diferentes. A relativização da justificação é uma defesa ainda mais robusta do RMM. Significa que um juízo moral pode ser justificado em uma sociedade, mas não em outra, pois tais grupos podem diferir a respeito dos padrões de justificação. Em última instância, isso significa que não há base racional para resolver disputas morais.

O ponto a ser notado aqui é o seguinte. Os predicados morais compartilham um significado comum em certo sentido, pois trata-se de predicados reconhecidos como *morais*. No entanto, dada a relatividade da verdade e da justificação para os juízos morais, o relativista terá de aceitar que a referência dos predicados morais irá variar entre sociedades que têm padrões normativos distintos.

---

[26] Cf. GOWANS, C. Moral Relativism, *The Stanford Encyclopedia of Philosophy.* EDWARD, N. Z. (Ed). URL: https://plato.stanford.edu/archives/sum2019/entries/moral-relativism.

Agora, note que o RMN é inconsistente com o RMM. Na verdade, uma das principais motivações para se adotar qualquer tipo de realismo moral é a não necessidade de lidar com os custos do relativismo. Qualquer teoria naturalista aceita que propriedades morais são naturais. Assim, se um juízo moral é verdadeiro ou falso depende de se a entidade relevante em questão (ação) tem a propriedade natural em questão (correção, incorreção, bondade). Isto é, a verdade de um juízo moral independe do que determinado grupo supõe que é verdadeiro, mas depende de como o mundo é.

No entanto, parece que a TRS implica justamente no tipo de visão que RMM tem sobre os predicados morais, isto é, que tais predicados, embora tenham um significado comum entre as sociedades (pois são parte da categoria mais geral *predicados morais*) possuem referência distinta. Desse modo, não parece que a TRS é uma boa via de escape para o defensor do RMN diante do ATGM, pois o compromete a aceitar um custo muito grande, qual seja, a plausibilidade do RMM.

Diante disso, acredito que temos boas evidências para concluir que o ataque de Copp ao ATGM não nos fornece razões suficientes para concluirmos que o RMN está livre de problemas. Como vimos, assumindo a TRS não temos nenhum caso de violação de P2 do ATGM, a hipótese explanatória de Copp para a IUS não parece ser a melhor explicação e, além disso, a aceitação da TRS implica num custo muito alto para o RMN.

4. D. Merli e a Réplica do Desacordo Prático

No final de uma seção do seu *Milk, Honey, and the Good Life on Moral Twin Earth*, D. Copp faz uma sugestão sobre o possível desacordo entre habitantes de T e de TG, porém não a desenvolve. Ele afirma que é possível ver o desacordo no experimento de H&T como desacordo *prático* sobre *o que fazer* (COPP, 2000, p. 124). Em princípio, isso abre um novo caminho de réplica ao ATGM, pois talvez seria possível conceder a H&T que os falantes não estão em desacordo genuíno sobre o que é correto, mas expressam algum tipo de desacordo prático de outra natureza. Em *Return to Moral Twin Earth* (2002), D. Merli leva a sério esta sugestão e busca desenvolvê-la detalhadamente. Ele sustenta que, embora os falantes do experimento de H&T não estejam em desacordo sobre o que é moralmente correto, eles estão num desacordo *prático* sobre *o que deve ser feito*. Além disso, Merli desenvolve uma teoria semântica sobre esses juízos práticos que combina realismo naturalista com expressivismo de normas. Ele acredita que este tipo de resposta "neutraliza o desafio geral antirrealista" (MERLI,

2002, p. 231). Nesta seção, irei analisar o ataque de Merli ao ATGM. Tal réplica já foi considerada em pelo menos dois outros lugares (RUBIN, 2013; KEYZER, 2016). Aqui, buscarei sustentar que essa estratégia para se recusar o ATGM possui um custo muito alto para o RMN e que, por isso, deve ser deixada de lado ou, pelo menos, revista. Ao fazer isso, irei me basear amplamente nos trabalhos de Rubin e Keyzer.

Podemos entender essa terceira réplica de Merli ao ATGM como tendo duas partes principais. O que ele está propondo é que se mude a localização do desacordo de *desacordo moral* para *desacordo prático* (denominemos, neste sentido, o ataque de Merli de *Réplica do Desacordo Prático*). De acordo com Merli, não devemos entender os falantes do experimento de H&T como discordando sobre o que é *moralmente correto*, mas como discordando sobre *o que deve ser feito*. Juízos normativos sobre *o que deve ser feito* são distintos de juízos morais sobre *o que é correto*, pois a primeira, sustenta Merli, é uma noção[27] mais robusta. Assim, mesmo que dois sujeitos reivindiquem conteúdos diferentes com juízos sobre o que é *correto*, de forma mais ampla, eles devem decidir sobre *o que deve ser feito*, e, com isso, eles discordam. Ou seja, embora a afirmação de que '*x* é correto-*t*' e a afirmação de que '*x* não é correto-*tg*' não expressem conteúdo proposicional conflitante, habitantes de T e TG expressam um outro tipo de desacordo, um desacordo *prático* sobre *o que fazer*. Isso constitui a primeira parte do ataque de Merli. A segunda, se refere a como devemos interpretar esses juízos normativos sobre *o que deve ser feito*. Pois os mesmos problemas metaéticos que são levantados para juízos que empregam *bondade* e *correção*, se apresentam para os juízos sobre *o que deve ser feito*. Note que não podemos meramente adotar a teoria semântica naturalista já existente para o *correto*, a proposta seria facilmente desfeita pelo mesmo ATGM e Merli não teria nada de novo a nos dizer a favor do RMN. Por isso, Merli apresenta uma teoria semântica alternativa para juízos sobre *o que deve ser feito* que supostamente não é vulnerável ao desafio proposto por H&T. Segundo Merli, isso é suficiente para livrar o RMN do ATGM[28].

Consideremos, agora, com mais detalhes a *Réplica do Desacordo Prático*. Numa das passagens chave, Merli afirma o seguinte:

---

[27] Por enquanto, uso o termo geral 'noção' para se referir à expressão 'o que deve ser feito'. Adiante, caracterizo precisamente o que Merli entende por isso.

[28] É necessário fazer uma observação sobre essa noção de *o que deve ser feito*. Merli refere-se a esse conceito, que é mais robusto do que 'moralmente correto' e capta o desacordo prático, de formas variadas. Ele fala em "how to act", "the last ought before action", "what to do", "what to do all-in" "what ought to be done", "all-in endorsement" e "guide to conduct". No entanto, seu uso mais recorrente é "o que deve ser feito". Com o propósito de simplicidade, usarei apenas 'o que deve ser feito' para expressar esta noção a que Merli se refere.

> Anteriormente, fiz alusão a um tipo de conflito que é familiar aqui na Terra: o conflito entre diferentes perspectivas avaliativas ou pontos de vista. A moralidade demanda uma ação enquanto que a prudência, etiqueta etc, demandam outras. Ao decidir como agir precisamos saber não apenas o que é correto ou o que é prudente – também precisamos saber se seguir a moralidade ou a prudência naqueles casos em que as duas demandas são distintas. Pessoas diferentes as vezes se posicionam em lados diferentes dessa questão, mesmo que concordem sobre as respostas fornecidas por essas duas espécies de avaliação. Uma pessoa pode estar comprometida a fazer a coisa *certa*, enquanto que outra a fazer o que é mais *vantajoso* ou o que melhor se adéqua à *etiqueta* mesmo que isso signifique a violação de demandas morais que eles reconhecem. *Parece que estamos engajados em conversação e desacordo sobre o que deve ser feito*. Esses conflitos também acontecem no âmbito particular do agente quando, por exemplo, ponderamos se razões morais se sobrepõe, em algumas instâncias, a outros tipos de considerações. Neste sentido, parece que uma decisão de outro tipo é requerida uma vez que sabemos o que nos é exigido a partir desses pontos de vista avaliativos concorrentes. (MERLI, 2002, p. 234. Itálico meu).

O tipo familiar de conflito anteriormente aludido por Merli é o seguinte. Na vida ordinária, de acordo com ele, numa discussão sobre como proceder em determinada situação as pessoas podem dizer que é correto agir de tal e tal modo reivindicando *conteúdos* diferentes sobre 'correto'. Por exemplo, suponha que S_1 e S_2 estejam discutindo sobre se é ou não correto mentir na situação *x*. S_1 acha que dizer uma mentira seria a coisa certa a fazer, pois, além de não ter grandes implicações por ser uma situação irrelevante, é o que demanda a etiqueta. S_2 acredita que, mesmo na situação *x*, não seria correto mentir, pois a mentira é uma ação condenável do ponto de vista moral e isso se sobrepõe à mera etiqueta. Aqui S_1 parece atribuir à 'correto' um conteúdo diferente do atribuído por S_2. Enquanto S_1 apela à etiqueta, S_2 recorre à moralidade. Ainda assim, sustenta Merli, parece que a discussão entre S_1 e S_2 é uma espécie de desacordo, pois não se trata de uma disputa sobre o conteúdo do predicado 'correto', mas sobre *o que deve ser feito*.

Sendo assim, o ponto da passagem acima, em primeiro lugar, é que há uma diferença de latitude entre, de um lado, *correto* (podendo este ser correto moral, de etiqueta, de prudência etc) e de outro lado, *o que deve ser feito* e, em segundo lugar, que essa diferença pode explicar porque dois falantes discordam genuinamente mesmo usando termos normativos com conteúdo diferente. Na visão de Merli, a noção de *o que deve ser feito* é mais abrangente e capta todas as ocorrências de *correto*. Juízos sobre o que é moralmente correto estão numa categoria mais específica, tal como juízos sobre o que é prudente, o que preserva a etiqueta, o que é melhor do ponto de vista egoísta etc. Juízos normativos *sobre o que deve ser feito* são mais gerais na medida em que envolvem a consideração de todos esses juízos mais específicos para a tomada de decisão *prática*. Embora diferentes partes apresentem juízos de correção diferentes, a instância última da discussão é sobre *o que deve ser feito*, e é esta noção mais robusta que reúne todas as outras. Assim, numa disputa ordinária sobre o que fazer, mesmo que haja desacordo

semântico sobre o conteúdo de ‘correto’’, como no nosso exemplo anterior em que tínhamos um lado sustentando que correto é aquilo que é *moralmente* correto e outro que é correto aquilo que não fere a *etiqueta*, em última instância o que se está discutindo é como se deve proceder *praticamente*.

Numa passagem no final do seu artigo, Merli esclarece melhor porque juízos normativos sobre *o que deve ser feito* são diferentes de juízos normativos sobre o que é *correto*.

> Podemos ficar um pouco desconfiados sobre a identificação de algum ponto de vista avaliativo substantivo (moralidade, prudência, racionalidade...) com *o que deve ser feito*. O raciocínio aqui é familiar: há um tipo de questão aberta que pode ser levantada sobre se os resultados de qualquer uma dessas perspectivas determinam *o que deve ser feito* (MERLI, 2002, p. 237).

O ponto aqui é que juízos morais e juízos sobre *o que deve ser feito* são diferentes porque os primeiros sempre envolvem questões abertas. Isto é, sempre faz sentido perguntar, quando aceitamos que algo é moralmente correto, se deve ser feito. Por exemplo, não mentir é moralmente correto, mas será que deve ser feito numa circunstância em que poderia causar algum mal? Essa característica é ainda mais visível quando entram em consideração uma variedade de juízos sobre o que seria correto. Enquanto isso, juízos *sobre o que deve ser feito* não envolvem questões abertas, pois são sempre o ponto definitivo a que todas as questões abertas suscitadas por juízos sobre o que é correto são dirigidas.

Tendo notado essa suposta diferença de generalidade entre juízos morais sobre *correto* e juízos sobre *o que deve ser feito*, Merli tem uma explicação alternativa para a IUS.

> E aqui, podemos pensar, é onde discordamos com os moralistas gêmeos: o que estão em questão nessa conversação é se a correção ou a correção-gêmea é o guia correto para a conduta (MERLI, 2002, p. 234).

A ideia de Merli é que essa característica do mundo ordinário pode ser usada para entendermos o que está acontecendo no cenário descrito por H&T. Temos duas comunidades que reivindicam conteúdos diferentes sobre ‘correto’ e, deste ponto de vista moral, eles não discordam. No entanto, num sentido mais geral os habitantes de T e TG estão em desacordo genuíno, pois, em última instância, o ponto da disputa é sobre *o que deve ser feito*. O ponto sobre *o que deve ser feito* não consiste apenas em saber o que a moralidade, a etiqueta ou a prudência nos dizem que é correto. Nós precisamos saber se seguimos as demandas da moralidade, da etiqueta ou da prudência. Assim, se T e TG entrassem numa disputa, mesmo que cada um fizesse uma reivindicação distinta, dado o conteúdo diferente de cada predicado moral, ainda assim eles

teriam que decidir sobre como agir, sobre *o que deve ser feito*. E isto é o elemento comum que os une na disputa. Merli acredita que este elemento comum garante que eles estejam em desacordo genuíno.

Temos uma boa réplica a H&T aqui. Mais especificamente, temos um ataque a P2 do ATGM. Como sabemos, P2 é a tese de que dois indivíduos não podem discordar genuinamente se os predicados morais usados na discussão predicam conteúdos diferentes, pois suas proposições não seriam conflitantes. O que Merli nos oferece é um modo alternativo de entender a disputa em que dois indivíduos podem discordar genuinamente mesmo que seus termos tenham conteúdos diferentes, pois o fim último que guia a sua discussão é *prático*, sobre *o que deve ser feito*. Assim, "falantes discordam mesmo sem compartilhar o termo 'correto'" (MERLI, 2002, p. 232).

Mas, como devemos entender esses juízos normativos sobre *o que deve ser feito*? Eles estão sujeitos aos mesmos problemas metaéticos que se colocam sobre os juízos em que o predicado 'correto' ocorre, tais como: possuem conteúdo cognitivo ou não cognitivo? Se possuem conteúdo não cognitivo, trata-se de que tipo de estado mental? Há fatos sobre *o que deve ser feito*? Se sim, esses fatos são independentes de nossas mentes? Tais fatos são naturais ou não naturais? E assim por diante. O sucesso da *Réplica do Desacordo Prático* depende de que tipo de resposta se dá a esses problemas metaéticos. Não basta apenas Merli apontar para uma suposta distinção de latitude entre juízos morais e juízos normativos sobre *o que deve ser feito* e dizer que o ATGM não é mais um problema para o RMN. Ele precisa fornecer uma teoria metassemântica sobre *o que deve ser feito*. Por isso, Merli aponta duas alternativas aqui. A primeira é a adoção do Realismo Naturalista tanto para juízos morais quanto para juízos sobre *o que deve ser feito*. A segunda é o Realismo Naturalista sobre juízos morais combinado com o Expressivismo de Normas sobre *o que deve ser feito*. Merli adota a segunda posição. Vamos a elas.

4.1. Realismo Naturalista

A alternativa mais natural para o defensor da *Réplica do Desacordo Prático* é adotar para *o que deve ser feito* o mesmo tipo de metassemântica realista naturalista que já é adotada para os juízos morais. Ele poderia dizer que *o que deve ser feito* expressa uma propriedade natural *N* (ou um conjunto de propriedades naturais *N*), que é rastreada pela melhor teoria normativa de primeira ordem disponível. E o uso de *o que deve ser feito* é causalmente regulado

por esta propriedade *N*. No entanto, é fácil ver porque não se pode adotar tal estratégia. Ela é vulnerável ao ATGM.

Suponha que, na Terra, o uso de *o que deve ser feito-t* é R-relacionado com uma propriedade natural N_1. Na Terra Gêmea, o uso de *o que deve ser feito-tg* é R-relacionado com uma propriedade natural N_2. *O que deve ser feito-t* e *o que deve ser feito-tg* compartilham as mesmas características formais: as pessoas são motivadas a realizar ações a que esses termos se aplicam, tais termos são usados para aprovar cursos de ação etc. Agora, suponha que habitantes de T e de TG discutem sobre se *x* deve ser feito. Habitantes de T dizem que '*x* deve ser feito-*t*' enquanto que habitantes de TG dizem que '*x* não deve ser feito-*tg*'. Se a teoria metassemântica é verdadeira, então T e TG não estão expressando um desacordo genuíno, pois estão predicando coisas diferentes sobre a situação *x*, a saber, N_1, por habitantes de T, e N_2 por habitantes de TG. Portanto, suas sentenças não expressam proposições conflitantes. No entanto, nossa intuição parece continuar sendo a mesma que no experimento original de H&T: os indivíduos discordam genuinamente. Se é assim, então deve haver algo a mais no significado desses juízos sobre *o que deve ser feito* que não é captado pela teoria Realista Naturalista em questão.

Portanto, se o defensor da *Réplica do Desacordo Prático* adotar o realismo raturalista como teoria metassemântica, não parece haver nenhum progresso do RMN diante do ATGM. Ele precisa de uma teoria alternativa. É por isso que Merli sugere manter o realismo naturalista para juízos sobre o que é moralmente *correto*, mas adotar o expressivismo de normas para juízos sobre *o que deve ser feito*.

4.2. Expressivismo de Normas

A teoria metassemântica adotada por Merli é uma conjunção do realismo naturalista para o discurso moral com o expressivismo de normas para o discurso normativo sobre *o que deve ser feito*.

> Tal abordagem combina o realismo sobre o discurso moral com o expressivismo sobre *o que deve ser feito* (*all-in endorsement*). De acordo com essa visão, a correção moral é uma questão de fato natural, mas uma resposta para a questão sobre o que deve ser feito ... não é um juízo factual, mas um *endosso* de um curso de ação ou de um conjunto de razões para a ação. Quando faço a coisa certa, estou expressando minha aceitação de certas normas, ou encorajando outros a agir de acordo ou algo nessa linha (MERLI, 2002, p. 236).

Assim, juízos morais sobre o que é *correto* continuam predicando uma propriedade natural *N* que é rastreada pela teoria normativa e primeira ordem que se adota. Mas juízos normativos sobre *o que deve ser feito* não instanciam propriedade alguma, pois sua função semântica consiste em expressar estados mentais conativos. Tais estados mentais normalmente são identificados com atitudes de aprovação ou desaprovação e são como desejos ou preferências, embora reivindiquem sobreposição aos meros desejos e preferências não relacionados à moralidade. No caso específico do expressivismo de normas (Cf. GIBBARD, 1990), tais estados mentais conativos consistem no endosso de um sistema de normas.

Mas por que não seria estranho adotar, ao mesmo tempo, realismo naturalista e expressivismo de normas? Afinal, essas teorias são uma das maiores rivalidades da metaética. A resposta de Merli é que todo expressivista ou é expressivista sobre *todo* tipo de discurso ou sobre *algum* tipo de discurso. Ele considera que a visão mais plausível é a segunda (MERLI, 2002, p. 237). Assim, nada impede que ele seja um expressivista sobre o discurso normativo sobre *o que deve ser feito* e adote outras posições em relação a outras partes do discurso, tal como um realismo naturalista sobre o discurso moral ou lógico, por exemplo.

A razão pela qual este tipo de posição supostamente evita o ATGM é clara. Como afirma Merli, "mesmo aceitando que não estamos em desacordo sobre correção ou correção-gêmea, ainda assim temos espaço para alguma forma de disputa genuína" (MERLI, 2002, p. 239). Em primeiro lugar, podemos ver a *Réplica do Desacordo Prático* como um modo de recusar P2 do ATGM e assegurando que dois indivíduos podem entrar em desacordo genuíno mesmo sem compartilhar o conteúdo do vocabulário moral. Isso é supostamente garantido pela primeira parte da réplica. Em segundo lugar, com a adoção dessa combinação de teorias metassemânticas, aparentemente, temos um modo de inviabilizar qualquer variante do ATGM. Isso porque, se o discurso sobre *o que deve ser feito* é expressivista, então há algo no significado dos predicados usados por habitantes de T e de TG que é compartilhado de modo que a possibilidade de se supor duas comunidades com termos morais com conteúdos diferentes é eliminada já de antemão. Pode-se até supor comunidades que expressam o endosso de normas diferentes. Mas ainda assim, ambas expressarão *endosso de normas* e isso é o aspecto comum suficiente para bloquear qualquer tipo de argumento similar ao ATGM.

4.3. Problemas Para a Réplica do Desacordo Prático

Um modo de atacar a *Réplica do Desacordo Prático* é resistir à sua distinção entre juízos morais e juízos sobre *o que deve ser feito*. Se poderia argumentar que juízos sobre o que é *correto* são simplesmente juízos sobre *o que deve ser feito* dizendo, por exemplo, que é injusto colocar juízos morais na mesma categoria de juízos sobre etiqueta, prudência ou razões egoístas, pois aqueles se sobrepõe a todos esses e tem o mesmo grau de robustez do que a noção de *o que deve ser feito*. Assim, se os predicados morais dos habitantes de T e de TG tivessem conteúdo diferente, não teria como recorrer a uma "outra" noção supostamente mais robusta que preservaria alguma comunalidade discursiva da linguagem de T e TG. Além disso, se se quisesse reivindicar algum tipo de expressivismo como teoria metassemântica, apenas se daria razões para recusar o RMN, pois a finalidade positiva do ATGM é justamente de fornecer razões para o expressivismo.

Um outro caminho possível a ser seguido seria insistir que ele não fornece uma refutação satisfatória de P2 do ATGM. Pois para recusar tal premissa é necessário apresentar casos em que há desacordo genuíno e que, ao mesmo tempo, a CIE é violada. Merli não apresenta um caso deste tipo, ele apenas muda o foco do desacordo concedendo que não há desacordo *moral* (pois os termos morais têm conteúdo/referência distinta), mas sim um desacordo *prático* (cujo conteúdo dos juízos sobre *o que deve ser feito* é expressivista, isto é, garantindo que tais juízos têm significado similar). Sendo assim, Merli não apenas não mostra como seria possível violar a CIE, mas ele próprio reforça a importância desta condição.

Mas não irei seguir nenhuma dessas possíveis réplicas para defender o ATGM aqui. Acredito que a *Réplica do Desacordo Prático* coloca seu defensor numa posição desconfortável devido à adoção de teorias diferentes para juízos morais, por um lado, e para juízos sobre o que deve ser feito, por outro. É importante notar que Merli não tem uma teoria híbrida completa, como fazem algumas tendências da metaética contemporânea[29]. Em outras palavras, ele não desenvolve uma abordagem que relacione, *para a mesma classe de juízos*, elementos de uma teoria realista, por exemplo, e aspectos de uma teoria expressivista. O que ele sugere é que se adote o realismo naturalista para juízos morais e expressivismo de normas para juízos sobre o que deve ser feito. No entanto, isso implica num custo muito alto, como veremos e, além disso,

[29] Na parte final desse trabalho, veremos com mais detalhes no que consiste o projeto de uma teoria metaética híbrida.

corre o risco de ameaçar o próprio realismo naturalista que visa preservar a respeito dos juízos morais. Novamente, vale ressaltar que se ele tivesse uma teoria híbrida *sobre a mesma classe de juízos*, isso não seria um problema, pois tal tipo de abordagem visa conciliar aspectos realistas e expressivistas.

Vejamos esse problema com mais detalhes. Os defensores do RMN têm objetado o expressivismo argumentando que essa teoria enfrenta, entre outros problemas: 1) não acomoda as pretensões objetivas da moralidade; 2) requer que vejamos a gramática declarativa das sentenças morais como enganosa; 3) não explica porque fazemos alusão a propriedades morais (bondade, incorreção); 4) não fornece uma explicação robusta para a aparente predicação de verdade do discurso moral; 5) não fornece uma explicação sobre sentenças [*embedded]* no contexto condicional; 6) tem sérios problemas em acomodar a aparente validade de argumentos que envolvem predicados morais. Essa lista de problemas é apontada contra o expressivista não apenas para se desacreditar essa teoria, mas como *razões positivas a favor do realismo*, em geral, e do RMN, em específico. Isto é, o realista normalmente argumenta a favor da sua posição apontando problemas em teorias concorrentes, sendo o expressivismo um dos seus principais rivais.

Neste sentido, alguns (KEYZER, 2016, p. 124) têm argumentado contra a *Réplica do Desacordo Prático* tentando colocar uma espécie de dilema para o seu defensor. Se se empregar a *Réplica do Desacordo Prático* contra o ATGM, então compra-se para si, além dos problemas usuais do RMN, os problemas teóricos do expressivismo. Se não se empregar a *Réplica do Desacordo Prático*, então o ATGM permanece como um desafio ao RMN. Qualquer uma das duas opções implica num custo muito alto.

Acredito que a dificuldade é ainda maior. Ao adotar o expressivismo de normas, o RMN pode *ou* permanecer neutro em relação aos problemas mencionados acima *ou* não permanecer neutro e assumir o ônus de lidar com eles. No entanto, parece que ele não pode permanecer neutro a tais problemas dizendo, por exemplo, que é um expressivista *local*. Pois os mesmos problemas metaéticos que se aplicam aos juízos morais aplicam-se também aos juízos sobre *o que deve ser feito*. Além disso, a plausibilidade de uma teoria metaética não pode residir somente no seu potencial de lidar com *um* problema. Como a metaética é um terreno bastante controverso, a adoção de uma teoria ou outra normalmente envolve um processo de pesagem dos custos e benefícios. Assim, uma teoria pode acomodar um problema satisfatoriamente, mas, em contrapartida, enfrentar uma série de outros problemas que, no pacote teórico geral, se sobrepõem à vantagem inicial. Desse modo, assumir o expressivismo

pode ter uma vantagem inicial de se evitar o desafio do ATGM. Mas, dado que Merli assume também o RMN, será que o conjunto de problemas acima mencionado, não se sobreporia a essa vantagem inicial?

Por outro lado, como argumenta M. Rubin (RUBIN, 2013, p. 39), se ele tomar posição a respeito dos problemas do expressivismo na tentativa de mostrar a plausibilidade da sua estratégia, ele corre o risco de minar aquilo que ele próprio quer preservar, o RMN. Pois, como afirmei, os realistas atacam o expressivismo lançando mão de um conjunto de problemas que esta teoria supostamente enfrenta; e a presença desses problemas não conta apenas como razão para desacreditarmos o expressivismo, mas como razão *positiva* para aceitarmos o realismo. Assim, ao tentar evitar tais problemas, coloca-se o RMN sob ameaça. Além disso, se o defensor da *Réplica do Desacordo Prático* tentar livrar o expressivismo de tais problemas, ele parece nos dar razões para aceitarmos o expressivismo não apenas em relação aos juízos normativos sobre *o que deve ser feito*, mas em relação aos juízos morais em geral, ameaçando, novamente, o RMN que ele quer preservar.

Este tipo de objeção à estratégia de Merli não significa que a *Réplica do Desacordo Prático* falha definitivamente. Significa apenas que a proposta de adotar RMN com expressivismo de normas parece pouco promissora, uma vez que não se trata de uma teoria híbrida. Mas isso não impede que o defensor do RMN preencha a sua resposta com uma teoria semântica diferente a respeito dos juízos normativos sobre *o que deve ser feito*. Esta teoria deverá, ao mesmo tempo, evitar o ATGM e honrar os compromissos teóricos do RMN (Cf. RUBIN, 2013, p. 37), o que parece bastante difícil. E, por mais modesto que seja recusarmos a proposta de Merli com base somente no problema acima, essa estratégia sugere ao menos duas coisas. Em primeiro lugar, certo pessimismo com relação ao sucesso deste tipo de resposta ao ATGM. Em segundo lugar, a inversão do ônus da prova. O problema agora não está nas mãos do defensor do ATGM, mas coloca-se para o defensor da *Réplica do Desacordo Prático* em fornecer uma teoria alternativa. De qualquer modo, o ATGM não parece estar em séria ameaça.

5. D. Plunkett e T. Sundell e a Réplica da Negociação Metalinguística

O ATGM é uma instância específica de um tipo mais geral de argumento (*Argumento Baseado no Desacordo*), usado em várias áreas do discurso filosófico, que parte de (i) intuições sobre o desacordo para (ii) extrair conclusões semânticas. O raciocínio é o seguinte: dois falantes estão em desacordo genuíno sobre *x*; se significassem coisas diferentes com o uso dos

termos empregados em tal discussão, então não estariam em desacordo genuíno, pois suas proposições não expressariam conteúdos inconsistentes; portanto, os termos usados por tais falantes não significam coisas diferentes. Como podemos notar, há uma conexão estreita entre *desacordo genuíno* e *similaridade do significado*. Denominemos isso de *Tese da Conexão*.

> *Tese da Conexão*: se dois falantes expressam um desacordo genuíno, então os termos usados em tal disputa possuem significado similar.

A *Tese da Conexão* desempenha papel central neste tipo de argumento. No ATGM, em particular, essa tese está explicitamente expressa em P2, que diz: T e TG não estão em desacordo genuíno se o conteúdo expresso por 'correto-*t*' e por 'correto-*tg*' é diferente.

Nos tópicos anteriores, já vimos que alguns filósofos tentam resistir à *Tese da Conexão* (especificamente no caso do ATGM) e que suas investidas não estão livres de problemas. No entanto, mais recentemente, um grupo de filósofos empregou um modelo alternativo para atacar esses *Argumentos Baseados no Desacordo*. Eles aplicam o que chamam de *análise metalinguística* às disputas hipotéticas que ocorrem nesses argumentos e sustentam que tal análise mostra que dois falantes podem expressar um desacordo genuíno mesmo que os termos usados na disputa não possuam significado similar. Isto é, a análise metalinguística supostamente mostraria que a *Tese da Conexão* é falsa. O trabalho melhor desenvolvido e especificamente direcionado ao *Argumento Baseado no Desacordo* expresso pelo ATGM é de D. Plunkett e T. Sundell. Em *Disagreement and the Semantics of Normative and Evaluative Terms* (2013), eles desenvolvem o que chamam de *negociação metalinguística* e argumentam que os "falantes podem discordar genuinamente, e frequentemente o fazem, mesmo quando *não* significam as mesmas coisas com o uso de suas palavras nas disputas que refletem esses desacordos" (Plunkett e Sundell, 2013, p. 3). Com isso, Plunkett e Sundell sustentam que um dos modelos mais famosos de argumentação das discussões recentes sobre o conteúdo semântico dos termos normativos e avaliativos é infundado.

O principal resultado da análise metalinguística aplicada aos *Argumentos Baseados no Desacordo* é que ela fornece uma ferramenta para que os defensores de certos tipos de teorias semânticas possam evitar certos ataques. No caso específico do ATGM, se a análise de Plunkett e Sundell está no caminho certo, os defensores do RMN encontram-se numa boa posição para recusar o desafio proposto por H&T, pois teriam pelo menos uma boa razão para recusar a *Tese da Conexão*.

Neste sentido, irei considerar agora tal desafio a P2 do ATGM. Vou me referir a esse ataque desenvolvido por Plunkett e Sundell como *Réplica da Negociação Metalinguística*[30]. Num primeiro momento, tentarei apresentar cuidadosamente esta réplica ao ATGM. Em seguida, passarei a discutir possíveis problemas que seus defensores precisam enfrentar.

5.1. A Réplica da Negociação Metalinguística

Iniciemos notando uma pressuposição importante que é feita por aqueles que empregam o *Argumento Baseado no Desacordo*: quando se supõe o desacordo entre os falantes desses cenários hipotéticos, pressupõe-se que eles discordam sobre o *conteúdo literalmente expresso* por suas expressões linguísticas (Cf. PLUNKETT e SUNDELL, 2013, p. 6-7). Isto é, quando os filósofos consideram as disputas[31] entre os possíveis falantes, o que determina sua posição a favor de que há desacordo genuíno ou de que não há desacordo genuíno é unicamente o conteúdo literal da sentença enunciada pelos falantes da disputa. Como notam Plunkett e Sundell, ignora-se totalmente aspectos externos à literalidade das expressões, tais como implicatura conversacional, por exemplo[32]. O resultado disso, segundo Plunkett e Sundell, é que a *Tese da Conexão* parece ser claramente verdadeira.

O ponto de Plunkett e Sundell é que, se a preocupação dos teóricos é com o significado dos termos avaliativos, essa pressuposição é arbitrária e parcial. Isso porque o conteúdo de uma expressão em termos de literalidade é apenas uma dimensão do significado, a saber, a dimensão semântica. Mas temos outra dimensão igualmente importante, a pragmática[33]. E, como

---

[30] É importante enfatizar que a *Réplica da Negociação Metalinguística* não está restrita ao ATGM, mas pretende atacar qualquer modelo de *Argumento Baseado no Desacordo*. No entanto, dados os objetivos específicos deste trabalho, irei apresentá-la tendo como alvo o ATGM. Como ficará claro, é simples generalizá-la para outras instâncias de *Argumentos Baseados no Desacordo*.

[31] 'Disputa' significa: "[...] qualquer troca linguística que *parece* evidenciar ou expressar um desacordo genuíno" (PLUNKETT e SUNDELL, 2013, p. 6).

[32] Implicatura conversacional se refere ao ato linguístico de significar ou implicar uma coisa dizendo algo diferente. Normalmente pressupõe certas características convencionais e contextuais. Ironia e metáfora são bons exemplos. São casos em que o significado da sentença do falante expressa uma coisa, mas o que o falante realmente quer comunicar tem um significado diferente. Considere o seguinte diálogo: A: Você irá na festa de João? B: Tenho que trabalhar. Note que a sentença enunciada por B não significa que B não irá na festa de João, significa apenas que ele tem que trabalhar. Mas, está claro que B está comunicando que não irá na festa. Sua sentença não *significa literalmente* que ele não irá na festa, mas ele *comunica* que não irá na festa com tal sentença. Nosso discurso ordinário está repleto dessas sutilezas linguísticas. (Cf. DAVIS, W. Implicature. *The Stanford Encyclopedia of Philosophy*. ZALTA, E. N. (Ed), URL: https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=implicature).

[33] Uma distinção precisa entre semântica e pragmática certamente não estará livre de disputa, mas, superficialmente, podemos dizer que, enquanto a semântica é o estudo da relação entre expressões linguísticas e o seu significado, a pragmática é o estudo do modo pelo qual o contexto pode influenciar a nossa compreensão dos enunciados linguísticos (Cf. SZABÓ, 2009, p. 4). Suponha que eu te diga "Bob bebeu uma garrafa de uísque e depois dirigiu até sua casa". No domínio semântico, o enunciado significa apenas que alguém, Bob, bebeu uma

sabemos, é possível comunicar pragmaticamente informações que o enunciado utilizado não comunica semanticamente (ironia, metáfora). Isso abre um caminho até então inexplorado nas discussões sobre os *Argumentos Baseados no Desacordo*. Talvez, em alguns desses casos hipotéticos descritos por esses argumentos, os falantes, embora empreguem expressões cujo conteúdo literal é diferente, comunicam pragmaticamente outro tipo de informação. Informação esta suficiente para gerar algum tipo de desacordo genuíno. Se isso for o caso, então é possível termos disputas em que os falantes usam expressões cujo conteúdo literal é diferente, mas que, mesmo assim, discordam genuinamente, já que comunicam pragmaticamente outro tipo de informação. Em outras palavras, se terá mostrado que o movimento da intuição sobre o desacordo para a conclusão sobre a univocidade semântica dos termos é apressada e não-garantida.

Plunkett e Sundell denominam essas disputas que focam apenas no conteúdo literal das expressões usadas de *disputas canônicas*[34]. Então, na tentativa de argumentar que a *Tese da Conexão* é falsa, eles tentam mostrar que há *disputas não-canônicas*. Este tipo de disputa se centra na informação comunicada pragmaticamente, ao invés de semanticamente (SUNDELL, 2015, p. 836). Uma instância notável de disputas não-canônicas são o que Plunkett e Sundell denominam de *negociação metalinguística*. Neste sentido, se as negociações metalinguísticas forem uma instância genuína de disputas não-canônicas, teremos forte evidência para rejeitar a *Tese da Conexão*. Consideremos, agora, o que são essas negociações metalinguísticas.

Para começar, temos que entender o que significa um *uso metalinguístico*. Esta noção foi primeiramente desenvolvida por Chris Barker (2002). Ele chama a atenção para cenários tais como o seguinte[35]. Considere a sentença:

(1) Feynman é alto.

---

garrafa de uísque e depois dirigiu até sua casa. Mas, pragmaticamente, posso estar comunicando algo para além disso. Dado as características contextuais, posso estar comunicando que Bob é um irresponsável e que agiu incorretamente. E você pode compreender perfeitamente isso, dado que domina essas características pragmáticas da linguagem.

[34] Uma definição mais formal de "disputa canônica" seria: uma disputa consistindo na afirmação do falante A de que *e* e na afirmação do falante B de que *f* é canônica se há dois objetos *p* e *q* (proposições, planos, etc) de modo que a afirmação de A de que *e expressa literalmente p* e a afirmação de B de que *f expressa literalmente q*, e *q* está fundamentalmente em conflito com *p* do modo apropriado para objetos de tal tipo (por *p* implicando não-*q* no caso de proposições; pela satisfação de *p* impedindo a satisfação de *q* no caso de desejos, e assim por diante) (Cf. PLUNKETT e SUNDELL, 2013, p. 9). É interessante notar que Plunkett e Sundell tomam o cuidado de não restringir o conteúdo dos enunciados a proposições apenas, pois isso excluiria de antemão as uma série de abordagens que fazem bom uso de Argumentos Baseados no Desacordo, tais como propostas semânticas expressivistas, que entendem o conteúdo das expressões valorativas como expressando atitudes de aprovação / desaprovação, planos, preferências etc.

[35] O cenário que passarei a descrever está em BARKER, 2002, p. 1-2 e PLUNKETT & SUNDELL, 2013, p. 13-14.

Normalmente, essa sentença seria usada para informar sobre a altura de Feynman. Mas há outro papel que ela pode desempenhar. Imagine que Tom e Tod estão numa festa. Tom, curioso sobre as práticas de culturas diferentes, pergunta a Tod o que as pessoas consideram como 'alto' em seu país. Na tentativa de fornecer uma explicação intuitiva, Tod observa Feynman dançando e diz:

(2) Bom... no meu país… Feynman é alto.

O uso que Tod faz de (1) aqui não é meramente descritivo. Isto é, (1) não foi usada para fornecer informação sobre a altura de Feynman. Diferentemente, (1) foi usada para fornecer informação sobre como usar um termo ('alto') de forma apropriada num determinado contexto. Note que (1) não se trata de um caso de *menção* do termo 'alto', mas de *uso*. No entanto, Tod usa (1) para dar informação sobre o significado de 'alto' num dado contexto. Os filósofos costumam fazer uma distinção entre linguagem objeto, que é usada ordinariamente para falar sobre objetos e coisas do mundo, e metalinguagem, que é usada para falar sobre as propriedades da linguagem objeto. Neste caso, dado que Tod não está usando (1) para descrever a altura de Feynman, mas para fornecer informação a Tom sobre o uso apropriado de um termo, ele parece estar no nível da metalinguagem. A este tipo de uso, Barker denomina *uso metalinguístico* de um termo (BARKER, 2002). Note, também, que literalmente (1) e (2) expressam coisas bem diferentes do que está sendo comunicado pragmaticamente. No nível semântico, (1) e (2) dizem respeito à altura de Feynman simplesmente. Mas isso não é o que os falantes estão realmente comunicando.

Plunkett e Sundell se apropriam dessa característica discursiva apontada por Barker para desenvolver o que chamam de *disputas metalinguísticas*. Imagine que junto a Tom e Tod está Ted, que contesta a afirmação de Tod. Após ouvir Tod expressar (2), Ted diz:

(3) Não. Feynman não é alto.

Assim como Tod não está fornecendo informação sobre a altura de Feynman ao enunciar (2), Ted também não está fornecendo informação sobre Feynman. Ted está contestando a afirmação de Tod sobre o uso apropriado de 'alto' num determinado contexto.

Assim, a disputa entre Tod e Ted não é sobre a altura de Feynman, mas sobre o uso correto ou apropriado de ‘alto’. Temos, portanto, uma disputa metalinguística.

A disputa entre Tod e Ted se trata de quem oferece a descrição mais acurada sobre o uso de um termo num determinado contexto. Eles estão trocando informação sobre como um determinado contexto realmente é (como as pessoas na comunidade tal usam o termo tal). Aqui, em certo sentido, há um contexto objetivo sobre o qual o desacordo entre Tod e Ted se dá, afirmam Plunkett e Sundell (Cf. PLUNKETT & SUNDELL, 2013, p. 14-15). Há uma linha demarcatória ou limite (*threshold*) sobre o que define alguém como alto ou baixo num contexto e a disputa é sobre quem fornece a descrição mais apropriada deste contexto.

Mas há um outro tipo de disputa metalinguística a que Plunkett e Sundell denominam *negociação metalinguística*. Considere o seguinte exemplo (Cf. PLUNKETT & SUNDELL, 2013, p. 14-15). Há duas características que marcam as disputas linguísticas denominadas de negociação metalinguística (PLUNKETT & SUNDELL, 2013, p. 3):

(a) Negociações Metalinguísticas empregam um tipo distintivo de mecanismo comunicativo: o *uso metalinguístico*;
(b) Negociações Metalinguísticas dizem respeito a uma questão *normativa* distintiva: como melhor usar uma palavra em relação a um contexto.

Plunkett e Sundell citam um exemplo extraído de Peter Ludlow (2008) em que este descreve um debate sobre se Secretariat (famoso cavalo de corrida americano que dominou as competições da década de setenta e quebrou uma série recordes) deveria ser colocado na lista dos maiores atletas do século XX. Podemos imaginar uma disputa do tipo:

(4.a) Secretariat é um atleta.

(4.b) Não. Secretariat não é um atleta.

Não parece que a disputa aqui seja sobre uma questão factual a respeito de Secretariat. Quer dizer, não é algo como se Secretariat superaria os cavalos de corrida de hoje ou se se destacaria em outras modalidades esportivas para além da corrida. Os falantes conhecem todos os fatos a respeito de Secretariat (seus tempos, seus recordes, suas conquistas etc). Parece que aqui a disputa é metalinguística, isto é, sobre a forma mais apropriada do uso do termo ‘atleta’.

Note, além disso, que os falantes significam coisas diferentes com o uso do termo ‘atleta’ (Cf. PLUNKETE T & SUNDELL, 2013, p. 16). O falante que enuncia (4.a) usa o termo de tal modo a incluir animais não humanos. Já o falante que expressa (4.b) usa o termo de modo a não incluir animais não humanos. Poderíamos dizer que a disputa entre (4.a) e (4.b) é algo como:

(5.a) Secretariat é *x*

(5.b) Secretariat não é *y*

em que o significado de *x* é tal que inclui animais não humanos e o significado de *y* é tal que não inclui animais não humanos.

Note que, se considerarmos essa disputa sobre Secretariat do ponto de vista canônico, temos que concluir que não há um desacordo genuíno entre os falantes. Pois, as sentenças não são inconsistentes, uma vez que uma afirmação não nega ou conflita com a outra já que o conteúdo literal do termo ‘atleta’ é diferente. No entanto, parece haver um desacordo genuíno entre os falantes sobre qual seria o uso mais apropriado do termo ‘atleta’. Como afirmam Plunkett e Sundell,

> Nessa compreensão da disputa, cada falante literalmente expressa uma proposição verdadeira dado o conceito que ele de fato expressa através de seu termo. Mas, para além disso, os falantes advogam pragmaticamente pelo conceito que estão usando e em virtude do qual eles asserem tais proposições. Assim, sua disputa metalinguística reflete um desacordo genuíno sobre como usar a palavra ‘atleta’ (PLUNKETT & SUNDELL, 2013, p. 17).

Isto é, um dos falantes comunica pragmaticamente que o uso mais apropriado do termo ‘atleta’ deve ser tal que inclua Secretariat enquanto o segundo falante nega, pragmaticamente, tal afirmação. Além disso, parece haver um desacordo genuíno aqui. Relembremos o que Plunkett e Sundell entendem por ‘desacordo’.

> *Desacordo Requer Conflito de Conteúdo (DRCC)*: Se dois sujeitos, A e B, discordam um com o outro, então há alguns objetos *p* e *q* (proposições, planos, etc) tal que A aceita *p* e B aceita *q*, e *p* é tal que as demandas impostas ao sujeito em virtude de aceitá-lo são racionalmente incompatíveis com as demandas impostas ao sujeito em virtude de aceitar *q*.

Parece que as demandas impostas ao sujeito que aceita (4.a) são racionalmente incompatíveis às demandas impostas ao sujeito que aceita (4.b). Quer dizer, (4.a) não pode, ao mesmo tempo, assumir, sinceramente, que Secretariat é um atleta e que Secretariat não é um atleta. O mesmo vale para (4.b).

Plunkett e Sundell argumentam que há casos de negociação metalinguística com termos valorativos e fornecem dois exemplos. O segundo é o que deve nos interessar mais aqui, pois eles consideram especificamente o ATGM e argumentam que o desacordo entre Terráqueos e Terráqueos Gêmeos é metalinguístico. Consideremos o primeiro exemplo (PLUNKETT & SUNDELL, 2013, p. 19).

Suponha que dois falantes discordam sobre o status do afogamento simulado (*waterboarding*). Eles sustentam, simultaneamente, o seguinte:

(6.a) O afogamento simulado é tortura.

(6.b) O afogamento simulado não é tortura.

Suponha, agora, que ambos falantes seguem diferentes orientações semânticas a respeito de como se deve definir 'tortura'. O indivíduo de (6.a) segue as Nações Unidas, que considera tortura "*todo ato que inflige sofrimento severo, físico ou mental, para obter informação ou para punir*" (United Nations, 1984, p. 85). Já o sujeito de (6.b) segue o Departamento de Justiça dos Estados Unidos, que define a tortura como "*todo ato de infligir dor ao nível da morte, falha de órgãos ou dano permanente de uma função corporal significante*" (U. S. Department of Justice, 2002, p. 340A). Portanto, os falantes significam coisas diferentes com o termo 'tortura'. Além disso, note que, se considerarmos apenas o conteúdo literal de suas afirmações, eles parecem não expressar um desacordo genuíno, pois dizem respectivamente:

(7.a) O afogamento simulado é um ato que inflige sofrimento severo, físico ou mental, para obter informação ou para punir.

(7.b) O afogamento simulado não é um ato de infligir dor ao nível da morte, falha de órgãos ou dano permanente de uma função corporal significante.

No entanto, sustentam Plunkett e Sundell, a disputa em questão aqui parece expressar um desacordo sobre qual seria a *melhor* ou mais *apropriada* definição de 'tortura', isto é, um caso de negociação metalinguística. Este parece ser o ponto central, embora os falantes não comuniquem isso através de expressão literal, mas pragmaticamente. Um falante comunica a visão de que a melhor definição de tortura é *x*. Outro comunica que a melhor definição de tortura é *y*. Trata-se de um desacordo genuíno, como nos mostra o DRCC, pois as demandas impostas ao sujeito em virtude de aceitar *x* são racionalmente incompatíveis às demandas impostas ao sujeito em virtude de aceitar *y*.

Sendo assim, concluem Plunkett e Sundell, temos um caso em que os falantes expressam conteúdos diferentes com o uso de seus termos e, mesmo assim, obtêm sucesso em discordar genuinamente. Lembre agora da *Tese da Conexão*. Ela nos diz que para haver desacordo genuíno é necessário que haja similaridade no significado. Mas, se algumas disputas são negociações metalinguísticas, então temos casos em que a conexão entre desacordo genuíno e similaridade do significado é violada. Isso mostra, na visão de Plunkett e Sundell, que um componente fundamental dos *Argumentos Baseados no Desacordo* é falso.

Notemos agora como este raciocínio se coloca como um desafio ao ATGM. Lembre que no cenário proposto por H&T temos a seguinte disputa:

(8.a) Mentir é correto-*t*.

(8.b) Mentir não é correto-*tg*.

Os dois falantes significam coisas diferentes com o uso do termo 'correto'. Para (8.a) a propriedade R-relacionada ao uso de 'correto' é a propriedade de *maximizar o agregado de felicidade*. Para (8.b) a propriedade R-relacionada ao uso de 'correto' é a propriedade de *tratar os outros como fins em si mesmos*. Portanto, se considerarmos a disputa entre (8.a) e (8.b) do ponto de vista canônico, não parece que temos um desacordo genuíno, uma vez que os falantes predicam coisas diferentes sobre 'mentir'. Lembre que o desafio ao RMN é que, no fundo, temos a intuição de que (8.a) e (8.b) discordam e se quisermos preservar esta intuição então temos que desistir de sustentar que eles significam coisas diferentes com o uso de 'correto' (algo que o RMN não pode abrir mão). No entanto, se o raciocínio de Plunkett e Sundell está no caminho correto, o defensor do RMN pode sustentar que (8.a) e (8.b) expressam um

desacordo genuíno sem ter que abrir mão do compromisso de que eles significam coisas diferentes com o uso de 'correto'. Os filósofos afirmam:

> […] podemos ver Bob, o Terráqueo, (8.a) e Chris, o Terráqueo Gêmeo, (8.b) como engajados num (talvez tácito) desacordo sobre qual conceito é o correto para se usar nesse contexto (PLUNKETT & SUNDELL, 2013, p. 20).

Ou seja, a disputa aqui, segundo Plunkett e Sundell, é sobre qual é a melhor definição para o termo 'correto'. Os falantes literalmente não expressam desacordo. Mas eles estão comunicando algo diferente. Estão comunicando diferentes visões sobre como melhor usar o termo 'correto'. E neste sentido eles estão em desacordo, como nos mostra o DRCC. As demandas impostas aos sujeitos em virtude de aceitar o que eles aceitam são racionalmente incompatíveis. Assim, temos um desacordo genuíno mesmo com os falantes significando coisas diferentes com os seus termos. Portanto, temos novamente uma violação da *Tese da Conexão*[36].

[36] É importante notar o seguinte. Em *Moral Twin Earth and Semantic Moral Realism* (2005), H. Geirsson ataca o ATGM empregando uma estratégia similar a de Plunkett e Sundell. Ele busca mostrar que é possível violar a *Tese da Conexão* recorrendo a uma conhecida distinção entre *referente semântico* e *referente do falante* (K, Donnellan, 1966). O referente semântico de um termo é a entidade que o termo refere. O referente do falante de um termo é a entidade que o falante refere ao usar o termo. Como ele nota "o que é importante sobre essa distinção … é que se pode usar um termo para se referir a um objeto ou a uma propriedade que não é o referente de tal termo" (Geirsson, 2005, p. 359). Plunkett e Sundell estão sustentando o mesmo ponto quando chamam atenção para a distinção entre semântica e pragmática. No entanto, o raciocínio que Geirsson emprega a partir deste ponto é diferente. Aplicando a distinção entre referente semântico e referente do falante ao ATGM, ele sustenta:

> Assumindo que os termos morais referem e assumindo que, na Terra Gêmea, Terráqueos e Terráqueos Gêmeos podem identificar sobre quais propriedades eles estão falando quando usam os termos morais, nada deveria impedir os Terráqueos, na Terra Gêmea, de serem capazes de usar os termos morais para se referir às propriedades morais da Terra Gêmea ... Os Terráqueos deveriam, assim, ser capazes, através da referência do falante, de usar os termos morais para se referir as propriedades morais da Terra Gêmea (GEIRSSON, 2005, p. 362).

A ideia de Geirsson parece ser a seguinte. Se os habitantes de T estão cientes de que os habitantes de TG usam 'correto-*tg*' para captar a propriedade de *tratar os outros como fins em si mesmos*, então os habitantes de T, ao invés de usarem "correto-*t*" como captando a propriedade de *maximizar o agregado de felicidade*, podem usar "correto-*t*", através da referência do falante, com a *intenção* de captar a propriedade de *tratar os outros como fins em si mesmos*, e vice versa. Assim, mesmo que seus predicados morais expressem conteúdos semânticos diferentes, eles podem usá-los com a intenção de referir a mesma propriedade. Com isso, Geirsson acredita ter mostrado que a *Tese da Conexão* é falsa e, portanto, ter refutado P2 do ATGM.
Não estou dando atenção especial a esta réplica de Geirsson aqui por dois motivos. Em primeiro lugar, porque na literatura já há um trabalho que lida diretamente com o ponto de Geirsson (veja RUBIN, 2014, p. 29-35). Em segundo, e mais importante, porque Plunkett e Sundell desenvolvem uma teoria muito mais completa e sofisticada a partir da distinção entre semântica e pragmática que se aplica a todas as instâncias dos *Argumentos Baseados no Desacordo* e não, como em Geirsson, meramente ao ATGM.

5.2. Objeções

A partir de agora, pretendo apresentar alguns problemas para a abordagem de Plunkett e Sundell e uma explicação alternativa para a IUS que considero preferível em relação à proposta dos dois autores. O objetivo final é fornecer uma defesa do ATGM contra a *Réplica da Negociação Metalinguística*.

5.2.1. Incompatibilidade com os compromissos externalistas

O primeiro problema para a teoria de Plunkett e Sundell se refere à sua inabilidade de articular de forma compatível a ideia das disputas metalinguísticas com os pressupostos externalistas. Para perceber isso, devemos notar, em primeiro lugar, que o compromisso semântico do RMN é externalista. Como vimos no primeiro capítulo deste trabalho, foi justamente o desenvolvimento das teorias externalistas do significado que prepararam o terreno para o ressurgimento do naturalismo moral. O ponto principal de argumentos a favor do externalismo semântico, tais como o *Argumento da Terra Gêmea*, de Putnam, é estabelecer que o mundo externo, e não as disposições dos falantes a respeito da aplicação dos termos, desempenha um papel relevante na determinação do significado. De acordo com o externalismo, o significado independe dos estados mentais, dos padrões de uso pelos falantes e não é relativo a um contexto. Pelo contrário, é determinado, *em parte relevante*, por características objetivas. Esse tipo de abordagem fora atraente para o RMN porque se poderia fornecer uma explicação para o conteúdo extensional dos predicados morais em sintonia com a tese realista naturalista de que fatos e propriedades morais são independentes de padrões deliberativos, crenças, visões de mundo etc. Em termos simples, essa é a significância do externalismo para o RMN.

Agora, note que a adoção da proposta sugerida por Plunkett e Sundell conflita com o pressuposto externalista, como argumenta Herman Cappelen (2018). Lembremos do exemplo dos próprios Plunkett e Sundell:

(a) O afogamento simulado é tortura.
(b) O afogamento simulado não é tortura.

A ideia de que o desacordo entre os falantes de (a) e (b) se trata de uma negociação metalinguística a respeito do melhor uso de 'tortura' pressupõe que os falantes possuem um controle efetivo sobre o significado dos termos em questão. Além disso, a expectativa de que tal tipo de desacordo chegue a algum tipo de resolução corrobora a ideia de que são os padrões de uso que irão determinar a melhor aplicação de um termo. Isso, claramente, não é a melhor forma de honrar os compromissos externalistas.

No entanto, o fato surpreendente a respeito dessa observação é que tal problema sobre a incompatibilidade com o externalismo semântico, embora seja um problema geral da abordagem de Plunkett e Sundell para qualquer exemplo de disputas metalinguísticas que eles sugiram, não parece ser um problema significativo especificamente em relação ao ATGM. Isso porque quem defende a ideia sugerida por Plunkett e Sundell poderia argumentar o seguinte. Considerando o cenário hipotético do ATGM em particular, os habitantes de T e TG estão aplicando teorias normativas de primeira ordem bem estabelecidas, e não meramente considerando definições possíveis de forma arbitrária. Isso é compatível com o RMN. A proposta de Boyd (1988), por exemplo, sustenta que é uma questão difícil e controversa a definição de qual teoria substantiva será a melhor teoria sobre a definição dos predicados morais e pode haver desacordo, isto é, comunidades que aplicam diferentes teorias. É justamente isso que acontece no cenário sugerido por H&T. Então, embora o problema da incompatibilidade entre as disputas metalinguísticas e os compromissos do externalismo semântico sejam um problema de ordem geral para Plunkett e Sundell, aparentemente, eles estão livres de tal preocupação pelo menos nesse nível específico.

Todavia, isso não significa que a proposta não seja vulnerável a outros problemas. Vejamos.

5.2.2. Atribuição de crenças incorretas

De acordo com Plunkett e Sundell, os indivíduos dos exemplos sobre se o afogamento simulado é tortura ou sobre se *x* é correto (no cenário do ATGM) estão envolvidos numa discussão cujo interesse é estabelecer como melhor aplicar um predicado, isto é, cuja razão de ser é como definir uma palavra em específico. Para além disso, eles sustentam que muitas disputas em ética são instâncias desse tipo de negociação metalinguística sobre a forma mais apropriada de definir um termo numa determinada linguagem. No entanto, como aponta Cappelen, esse parece ser o diagnóstico incorreto quando consideramos discussões morais.

> Os falantes dos tipos de conversação que Plunkett e Sundell usam como seus exemplos principais não pensam que seus argumentos e preocupações são irrelevantes para alguém que, por exemplo, fala Islandês, Chinês ou Russo. Uma forma de perceber isso é notar que eles iriam considerar-se em acordo ou deesacordo com aqueles que falam sobre o mesmo assunto em alguma dessas línguas (CAPPELEN, 2018, p. 174).

Considere as seguintes afirmações:

(a) A tortura é errada.
(b) Torture is *not* wrong.

A visão de Plunkett e Sundell implica que não há descordo entre (a) e (b), já que supostamente estão discutindo sobre como definir palavras diferentes em línguas diferentes. Mas isso parece ser atribuir crenças incorretas aos falantes de (a) e (b). Eles mantem posições sobre a prática da tortura e não sobre como melhor usar tal termo. Nesse caso, eles discordam sobre a tortura e isso é independente da língua que estão empregando para falar sobre a tortura. O mesmo se aplica para os habitantes de T e TG. Dizer que eles estão envolvidos num desacordo cujo único interesse é chegar um veredito sobre como usar o predicado moral 'correto' não parece ser a melhor abordagem sobre suas práticas morais.

Essa observação aponta para um problema maior, especificamente com relação ao modo como a proposta de Plunkett e Sundell visa explicar o desacordo entre membros de T e TG.

5.2.3. A proposta das negociações metalinguísticas não acomoda a IUS

À primeira vista, pode-se pensar que a resposta de Plunkett e Sundell dá conta da IUS e preserva os compromissos do RMN. Isto é, a teoria aceita, em primeiro lugar, que habitantes de T e TG possuem predicados morais com conteúdo semântico distinto (RMN) e, ainda assim, dá conta de acomodar a nossa intuição de que discordam genuinamente, já que se trata de uma disputa metalinguística sobre o melhor uso do termo 'correto' (IUS).

No entanto, um olhar mais cuidadoso mostra que há uma diferença entre o que é a IUS e o que a teoria de Plunkett e Sundell de fato explica. O conteúdo real da IUS é de que habitantes de T e TG estão engajados num desacordo moral genuíno. A ideia da negociação metalinguística explica que T e TG estão engajados num desacordo sobre como usar 'correto'.

Mas essas duas instâncias do desacordo são claramente diferentes. Imagine que S e S_1 estão em desacordo sobre se é correto fazer *x*. Eles devem chegar a um veredito pois precisam adotar um curso de ação, mesmo que o resultado não satisfaça as duas partes. Ao fim da discussão, eles tomarão determinado curso de ação e realização uma ação no mundo. Em contrapartida, imagine que S e S_1 estão agora em desacordo sobre como melhor aplicar o predicado 'correto'. Aqui eles não estão interessados em como tomar um curso de ação específico no mundo. Trata-se de uma disputa linguística sobre a melhor forma de usar um termo. O veredito da discussão não necessariamente irá implicar na realização de uma ação. Neste sentido, o desacordo moral e o desacordo linguístico parecem ser claramente distintos. Portanto, a proposta de Plunkett e Sundell não fornece uma explicação para o elemento central do ATGM e que coloca o RMN sob ameaça, a saber, a IUS.

5.2.4. A Ideia da Latitude como melhor explicação para o desacordo

Gostaria de propor, agora, uma explicação alternativa para o fenômeno da IUS. Explicação essa que está em concordância com o *insight* positivo do ATGM, isto é, que parte do conteúdo semântico dos predicados morais diz respeito a atitudes conativas. Tal explicação é preferível em relação à proposta de Plunkett e Sundell, já que está livre dos problemas de atribuição de crenças falsas e, de fato, busca explicar o fenômeno da IUS e não apenas desacordos linguísticos.

Trata-se da *Ideia da Latitude* e podemos compreendê-la do seguinte modo. Na ética não é requerido um acordo estrito sobre a extensão dos predicados morais como condição para a comunicação ou traduzibilidade de tais termos entre diferentes grupos ou comunidades. Quer dizer, quando se trata de termos morais, não é necessária a convergência rígida de atribuição de crenças a tais termos para que se possa reivindicar traduzibilidade ou desacordo entre os falantes. Fora do domínio da moralidade, se uma comunidade aplica um termo não moral *x* a um conjunto de instâncias *a*, *b* e *c* e outra comunidade aplica o mesmo termo não moral *x* a um conjunto de instâncias *d*, *e* e *f*, parece muito plausível dizer que os termos não são traduzíveis e que eles estão a falar sobre coisas diferentes. Considere o termo 'esquizofrenia', por exemplo[37]. Se uma pessoa, grupo ou comunidade, ao usar tal termo, atribui a crença de que estar em estado de esquizofrenia consiste, entre outras coisas, em 'ter um título de graduação acadêmica',

[37] Este exemplo é de F. Tersman (TERSMAN, 2006, p. 110s).

iremos dizer que tal pessoa não está disposta a aplicar ‘esquizofrenia’ ao conjunto de coisas que nós aplicamos ‘esquizofrenia’, já que o nosso uso de ‘esquizofrenia’ exclui a atribuição de crença de que ser esquizofrênico consiste, entre outras coisas, em ‘possuir um título de graduação acadêmica’. Portanto, nosso vocabulário, embora ortográfica e foneticamente similar, não é traduzível e, se dissermos respectivamente ‘*x* é esquizofrênico’ e ‘*x* não é esquizofrênico’ muito provavelmente não estaremos expressando um desacordo genuíno, pois falamos sobre coisas bem diferentes.

Predicados morais, por outro lado, não requerem tal uniformidade de atribuição de crenças. F. Tersman (2006) denomina tal característica dos termos morais de *Ideia da Latitude*.

> De acordo com a ideia da latitude, deveríamos exigir menos acordo e permitir mais erros e visões idiossincráticas no caso da ética comparada com outras áreas de investigação. Pois a ideia da latitude nos permite atribuir a uma pessoa uma convicção moral que não compartilhamos, mesmo na ausência de um *backgound* de acordo substancial e mesmo que esse veredito não possa ser excluído com referência a alguma falta cognitiva (especificável independentemente) (TERSMAN, 2006, p. 111s).

Na ética é muito plausível dizer que, num nível superficial, não há simetria sobre a correta aplicação dos predicados morais. Se houvesse, não teríamos nenhuma controvérsia moral. Pelo contrário, as pessoas discordam sobre questões morais o tempo todo. Além disso, podemos adentrar um nível mais profundo da moralidade em que as controvérsias são muito mais desconcertantes. O status moral da pena de morte, do aborto, do suicídio assistido, do melhoramento humano não é algo que suscita controvérsia apenas num nível ordinário. Filósofos do mais alto grau de especialidade nestes assuntos manifestam desacordos aparentemente irreconciliáveis. Mas isso é uma peculiaridade da ética, de acordo com a *Ideia da Latitude*. Note que não parece nem um pouco razoável dizer que os predicados morais usados pelas pessoas que manifestam diferentes atribuições de crenças morais não são traduzíveis ou que elas estão falando sobre coisas muito diferentes. Elas estão discutindo sobre moralidade. Neste sentido, é aceitável que habitantes de T e TG, seja nos três casos variantes acima, seja no experimento original de H&T, expressam desacordo moral, mesmo que a extensão dos termos seja diferente.

Na verdade, a explicação dos desacordos morais é um dos maiores apelos explanatórios da *Ideia da Latitude*, como sugere Tersman:

> Uma característica da ideia da latitude é precisamente que ela também nos permite estar em desacordo moral genuíno com pessoas cujos vereditos morais divergentes não podem ser excluídos com referência a alguma deficiência cognitiva. Ou seja, a ideia da latitude nos permite traduzir o ‘moralmente correto’ ou ‘justo’ de outrem com

> o nosso, mesmo que nossas diferenças sobre como aplicar tais termos não possam ser atribuídas a tais deficiências (TERSMAN, 2006, p. 80).

A *Ideia da Latitude* ajuda a explicar o fenômeno do desacordo porque podemos reconhecer como possuidores de crenças *morais* mesmo aqueles que têm posições muito diferentes de nós. Se levarmos a sério a ideia de que há propriedades naturais *N* que representam exaustivamente o conteúdo dos predicados morais e regulam estritamente o nosso uso de tais termos, em muitos casos teríamos que aceitar que não há desacordo moral genuíno onde parece claramente que há (tal como no cenário do ATGM). Além disso, a *Ideia da Latitude* permite resolver desacordos morais mesmo onde não há um ponto de partida comum entre os agentes. A tolerância garantida permite que estejamos em desacordo moral genuíno mesmo onde não parece haver ponto de partida comum entre os debatedores (TERSMAN, 2006, p. 81). Assim, a *Ideia da Latitude* justifica porque é legítimo que tenhamos a intuição do desacordo nos casos T e TG, mesmo que os habitantes de cada mundo tenham teorias morais diferentes.

E, para além disso, *Ideia da Latitude* explica *melhor* a IUS do que a abordagem sobre negociações metalinguísticas. A proposta de Plunkett e Sundell visa preservar o pressuposto do RMN de que o conteúdo dos predicados morais é fornecido exclusivamente pela propriedade natural *N* e transfere o desacordo para um nível metalinguístico. Como vimos, tal estratégia atribui crenças falsas aos falantes. A *Ideia da Latitude*, por outro lado, não atribui crenças erradas aos falantes, mas sustenta que na ética é admissível uma abordagem mais lata sobre o significado dos predicados morais. Essa abordagem mais lata está em sintonia com uma proposta expressivista. Isso porque no RMN as crenças determinam unicamente o conteúdo semântico dos predicados e, caso sejam diferentes, o conteúdo será diferente. Já no expressivismo, mesmo que as crenças morais sejam diferentes, há um significado ainda similar que é preservado, o conteúdo conativo. A *Ideia da Latitude* preserva isso. Além disso, a *Ideia da Latitude* acomoda a IUS, ao contrário da proposta de Plunkett e Sundell que localiza a explicação do desacordo no nível meramente linguístico.

6. Conclusão

No decorrer deste capítulo busquei analisar a estratégia de réplica ao ATGM especificamente direcionada a P2. Na primeira seção, na tentativa de dar conta de uma lacuna normalmente presente nas discussões sobre o ATGM, analisei o conceito de desacordo e propus que seguíssemos a definição de Plunkett e Sundell, segundo a qual, *Desacordo Requer Conflito*

de Conteúdo (DRCC). Na seção seguinte, considerei a resposta desenvolvida por D. Copp. Com base em três argumentos, sustentei que devemos recusá-la. Na terceira seção, abordei a *Réplica do Desacordo Prático* de D. Merli argumentando que a adoção de tal estratégia implica num custo muito alto para o defensor do RMN. Na quarta e última seção, apresentei detalhadamente a proposta de Plunkett e Sundell de que o desacordo entre os membros de T e TG trata-se de uma disputa metalinguística. Sugeri alguns problemas e propus uma explicação alternativa para a IUS argumentando que é preferível em comparação à teoria de Plunkett e Sundell. A linha geral desse percurso fora apresentar uma defesa de P2 do ATGM contra essas réplicas.

CAPÍTULO 4 – A FAVOR DA *INTUIÇÃO DA UNIVOCIDADE SEMÂNTICA*

1. Introdução

O ATGM tem duas pressuposições sobre semântica: (i) que as intuições semânticas geradas pelo experimento são o resultado da competência semântica/conceitual dos falantes e que (ii) essas intuições são confiáveis. Isso é o que autoriza o movimento de P3 para C. Para além disso, o ATGM é abdutivo. Quer dizer, ele faz uma inferência à melhor explicação a partir dessas intuições. Se pressupõe confiabilidade das intuições de falantes competentes e se é abdutivo, isso significa que a intuição suscitada pelo cenário hipotético pode ser, a princípio, derrotável. Ou seja, mesmo que a intuição seja amplamente compartilhada, forte e persistente, isso não garante sua plena segurança. Pode ser que a *Intuição da Univocidade Semântica* (IUS) seja resultado de algum tipo de ilusão, ignorância ou falta de clareza dos indivíduos que apreciam o cenário estipulado por H&T. Se isso for o caso, então P3, que é absolutamente fundamental para o sucesso do ATGM, é falsa.

Esta é a estratégia que Neil Levy (2011), Andrea Viggiano (2008) e John Sonderholm (2013) adotam. Eles recusam o ATGM tentando desmascarar a origem da intuição produzida pelo experimento. Eles argumentam, por caminhos diferentes, ou que a intuição resultante não é a da *Univocidade Semântica* ou que podemos ter explicações alternativas, mais convincentes e não incompatíveis com o RMN, para a sua presença. Sua estratégia consiste em sustentar que nossas intuições sobre o cenário do ATGM são resultado de limitações epistêmicas, de detalhes escondidos que falhamos em considerar ou da falta de clareza sobre a localização temporal do experimento. Quando todas essas contingências são eliminadas e podemos apreciar o experimento livres de qualquer viés, argumentam, deveremos ver claramente que a IUS foi apenas uma ilusão.

Meu objetivo deste capítulo é tentar recusar esta linha de réplica ao ATGM. No que segue, procurarei argumentar que os argumentos de Levy, Viggiano e Sonderholm não nos dão razões definitivas para concluirmos que P3 é falsa e que, por isso, devemos recusar o ATGM. Na primeira seção, irei considerar as críticas de Levy. Na segunda, o desafio proposto por Viggiano. E, na terceira, o ataque de Sonderholm. Irei concluir que nenhuma dessas objeções enfraquece a IUS e que, portanto, P3 permanece livre de problemas[38].

[38] Gostaria de adicionar aqui um esclarecimento sobre intuições. Irei pressupor que intuições semânticas ou intuições sobre o desacordo possuem significância filosófica, isto é, que é legitimo partirmos de intuições

2. N. Levy: psicologia e divergência futura

Uma das pressuposições fundamentais do ATGM é que os termos morais estão R-relacionados a propriedades naturais diferentes em T e TG[39]. Uma propriedade natural *N* captada por uma teoria moral de primeira ordem consequencialista em T - digamos, a propriedade de *maximizar o agregado de felicidade* - e uma propriedade natural *N* captada por uma teoria moral de primeira ordem deontológica em TG - digamos, a propriedade de *tratar os outros como fins em si mesmos*. Chamemos isso de *Estipulação Básica*. Em "*Moore on Twin Earth*" (2011), N. Levy sustenta que há dois modos de encararmos a *Estipulação Básica*. Numa interpretação, somos obrigados a rejeitá-la e concluir que o uso dos termos morais em T e TG é R-relacionado com a mesma propriedade moral. Se formos por este caminho, então a pergunta sobre o desacordo genuíno putativo entre os habitantes de T e TG nem se coloca, pois a rejeição da *Estipulação Básica* bloqueia o próprio avanço do ATGM contra o RMN. Numa outra interpretação, aceitamos a *Estipulação Básica*, mas somos obrigados a concluir que os habitantes de TG estão fazendo algo muito diferente do que engajados com moralidade. Portanto, não haveria desacordo genuíno entre membros das duas comunidades. Seja verdadeira a primeira ou a segunda interpretação, o ATGM não representa um problema para o RMN, pensa Levy.

No que segue, argumentarei que o desafio de Levy não coloca uma ameaça definitiva ao ATGM. Dividirei esta empreitada em duas partes. Na primeira, irei discutir sua interpretação de que o uso dos termos morais dos membros de T e TG é R-relacionado com a mesma propriedade natural *N*. Na segunda, tratarei da sua ideia de que não há desacordo entre habitantes de T e TG porque estes últimos não são agentes morais propriamente. A conclusão será de que nenhuma dessas interpretações de Levy nos dá razões suficientes para recusarmos P3.

2.1. Divergência Psicológica

ordinárias na elaboração de argumentos em defesa de teorias filosóficas. No entanto, estou ciente de que há ampla discussão sobre o papel e o alcance das intuições em filosofia.

[39] Um predicado moral estar R-relacionado a uma propriedade *N* significa que o uso de tal predicado por falantes de uma linguagem é determinado pelo conjunto de instâncias em *N*.

Comecemos pela primeira, e menos ameaçadora, interpretação da *Estipulação Básica*. Segundo Levy, se o uso dos termos morais nas duas comunidades é R-relacionado com propriedades naturais diferentes deve haver alguma razão para isso. E ele mesmo expõe a explicação de H&T.

> Horgan e Timmons nos fornecem uma resposta em seu experimento de pensamento: os termos morais são regulados por uma propriedade deontológica em TGM [Terra Gêmea Moral] porque seus habitantes diferem de nós *psicologicamente*. É porque eles são mais propensos à culpa e menos propensos à simpatia que uma teoria deontológica é verdadeira em seu mundo (LEVY, 2011, p. 142. Itálico meu).

Levy assume que o modo correto de interpretar esta passagem é dizer que a diferença psicológica entre T e TG é relevante para as circunstâncias de ação apenas, e não propriamente em relação às teorias morais nos dois mundos. Isso implicaria, segundo ele, que uma teoria deontológica é apenas superficialmente verdadeira em TG, mas que, no fundo, ambas as comunidades possuem a mesma teoria moral. A hipótese de Levy é a seguinte: é possível que as características psicológicas dos habitantes de TG fazem com que eles se sintam angustiados ao violar uma regra ou que acreditem que o custo de quebrar uma regra é muito maior do que o custo de não perseguir as melhores consequências (LEVY, 2011, p. 142). Isso explicaria porque eles adotam uma teoria deontológica ao invés de uma teoria consequencialista. No entanto, se isso é verdadeiro, há uma implicação importante: os habitantes de TG, em última instância, seguem uma teoria deontológica porque isso *maximizaria sua felicidade*. Mas note: "se é o caso que guiar o comportamento com referência a uma teoria deontológica maximiza a sua felicidade, a justificação fundamental para guiar o seu comportamento dessa maneira é, na verdade, consequencialista" (LEVY, 2011, p. 142). Portanto, habitantes de TG são, fundamentalmente, consequencialistas. E, dado que habitantes de T são, de fato, consequencialistas, o uso dos termos morais em T e TG está R-relacionado com a mesma propriedade natural.

De acordo com Levy, isso *explica* porque a IUS nos parece tão plausível. Nossa intuição é justamente de que os habitantes dos dois mundos manifestam um desacordo substancial normativo; e, se a teoria moral de primeira ordem é a mesma em T e TG, então não há conflito nenhum entre nossas intuições e as estipulações do RMN. Assim, a teoria de Boyd, tal como exposta no primeiro capítulo, não enfrenta problemas, pois os habitantes das duas comunidades estão em desacordo moral genuíno porque estão em acordo semântico sobre a referência do termo moral. Levy está tentando inviabilizar a própria possibilidade do ATGM, pois, se ele está certo, então P1, que estabelece que 'correto-*t*' e 'correto-*tg*' expressam

significados diferentes, é falsa. E se somos obrigados a rejeitar a *Estipulação Básica*, o argumento de H&T nem se coloca como objeção ao RMN.

Não vejo este ponto como particularmente problemático para H&T, como Levy parece ver. Isso porque mesmo que as diferenças psicológicas entre os habitantes de T e TG não sejam condição suficiente para justificar a adoção de teorias de primeira ordem diferentes nos dois mundos, isso não invalida a pressuposição de H&T, pois *é legítimo* assumir que T e TG adotem teorias diferentes. Vejamos.

Levy está sustentando que a razão dada por H&T para a preferibilidade de uma teoria consequencialista ou deontológica pelos habitantes de T e TG não é suficiente para mostrar que essas teorias são diferentes, pois mesmo os que seguem regras (TG) podem estar apenas querendo maximizar a sua felicidade. Devemos ser cuidadosos ao apreciar este ponto. Note que Levy não está sustentando que ambas as teorias normativas são convergentes. Ele está dizendo que a explicação fornecida por H&T não elimina a possibilidade de que ambas as comunidades tenham uma mesma teoria normativa fundamental. Mas, mesmo que ele estivesse se comprometendo com uma posição mais radical de dizer que consequencialismo e deontologismo podem ser encaradas fundamentalmente como teorias convergentes, ainda assim, isso não refutaria o ATGM (argumentarei a favor deste ponto na parte final desta seção).

Suponhamos que Levy esteja certo de que a diferença psicológica entre os habitantes de T e TG não seja um bom candidato para justificar porque eles aceitam teorias de primeira ordem diferentes. Isso refuta a *Estipulação Básica*? Acredito que não, pois pode haver explicações alternativas da razão pela qual os habitantes dos dois mundos adotam teorias diferentes. Levy encara de forma muito estrita a seguinte passagem de H&T:

> As diferenças na regulação causal, podemos supor, são devidas, ao menos em parte, a certas diferenças no temperamento psicológico que distingue Terráqueos Gêmeos de Terráqueos (H&T, 1992a, p. 165).

De acordo com a sua interpretação, a suposta diferença psicológica entre os membros de T e TG é a *única* causa para a escolha de teorias normativas diferentes. Assim, os habitantes de T adotam uma teoria consequencialista *somente* porque são mais propensos à simpatia uns pelos outros e os habitantes de TG adotam uma teoria deontológica *somente* porque são mais propensos à seguir regras. Mas, considere a seguinte passagem de H&T:

> Na verdade, duvidamos que realmente haja uma única caraterística temperamental – algum perfil particular de sentimentos – que opera em questões de moralidade para os Terráqueos em geral (H&T, 1992a, p. 173, nota 20).

Note que os próprios H&T duvidam que exista alguma característica temperamental que caracteriza definitivamente os habitantes de T. E, além disso, gostaria de chamar atenção para o próprio trecho de H&T a que Levy faz alusão. Nele, H&T afirmam que as diferenças psicológicas explicam apenas *em parte* a adoção de teorias normativas diferentes. Portanto, sua explicação em termos de diferenças psicológicas não é definitivamente a única causa da preferibilidade de teorias normativas diferentes, como Levy parece entender. Essas passagens deixam claro que ao citar a diferença psicológica entre os habitantes dos dois mundos, H&T estão estipulando tal tendência psicológica como uma, e apenas uma, explicação possível e não como a única *causa* da escolha de teorias diferentes. Se isso estiver certo, então podemos inclusive aceitar o que Levy está dizendo. Quer dizer, podemos aceitar que a diferença psicológica não é condição suficiente para demarcar a diferença na escolha de teorias de primeira ordem. Mesmo assim, isso não significa que temos que concluir que é a mesma propriedade que regula o uso dos termos morais em T e TG.

Mas, se poderia perguntar, se as diferenças psicológicas são apenas uma das possíveis explicações, então por que H&T não fornecem outras? Acredito que por duas razões, sendo a segunda a mais importante. Em primeiro lugar, porque não é parte central do experimento dizer precisamente porque T e TG assumem teorias normativas diferentes. É plausível pensar que, ao fazerem tal movimento, H&T estejam querendo apenas deixar o experimento mais completo. Em segundo lugar, porque *é legítimo* assumir que os membros de T e TG adotem teorias de primeira ordem diferentes, pois esta é uma peculiaridade importante do próprio RMN. Uma das teses centrais do RMN é sobre a importância das teorias normativas para a definição (*a posteriori*) dos termos morais e a identificação dos fatos e propriedades naturais que são idênticos aos fatos e propriedades morais (Boyd, 1988, p. 203s; Brink, 2001, p. 162; Sturgeon, 1985, p. 61). Para o RMN, os fatos e propriedades morais são agregados homeostáticos multiplamente realizados por fatos e propriedades naturais. E o que mantém esses fatos e propriedades unidos são propriedades naturais de ordem superior, tal como a propriedade de *maximizar o agregado de felicidade*. Mas é um ponto controverso qual propriedade natural melhor capta as instâncias de realização da propriedade moral. Portanto, as teorias morais de primeira ordem disponíveis desempenham um papel muito importante para o fornecimento de explicações possíveis. Considere esse exemplo de Sturgeon:

> Então, onde se deve procurar por tal explicação completa ou (se houver) por tal redução? Minha resposta é que tal explicação terá que ser derivada da *nossa melhor teoria moral*. [...] Se o utilitarismo hedonista de atos for verdadeiro, por exemplo, então podemos definir bom como prazer e ausência de dor, e uma ação correta é aquela que produz ao menos tanto bem quanto qualquer outra e será ali que os fatos morais se enquadram. Se, de forma mais plausível, alguma outra teoria moral acabar sendo verdadeira, teremos um outro tipo de explicação e (se a teoria tomar a forma correta) diferentes definições redutivas (STURGEON, 1985, p. 61).

Como afirma Sturgeon, a adoção de uma propriedade natural determinada como definidora da propriedade moral é fornecida pela teoria normativa. Mas nós não estamos num estágio de desenvolvimento em teoria moral avançado a tal ponto de podermos adotar, sem controvérsia, esta ou aquela teoria moral como sendo a teoria correta. Por isso, seremos obrigados a escolher abdutivamente entre teorias concorrentes. Isso significa que é natural que se escolham teorias diferentes para a definição da propriedade natural *N*. Na medida em que a discussão avança, podemos eliminar algumas teorias e ficar com outras e, esperamos, chegar ao ponto de adotar uma teoria como sendo *a melhor*. Isso é o que o próprio RMN pressupõe.

Neste sentido, quando H&T supõem a *Estipulação Básica*, estão apenas reivindicando esta característica do naturalismo moral. O ATGM não precisa fornecer uma teoria a mais para a explicação dos motivos pelos quais uma comunidade (T) ou outra comunidade (TG) aceitam tal e tal teoria. Este é um fato contingente e bem difícil de explicar. Talvez um bom candidato seja *razões*. Os habitantes de T podem ter chegado a um estágio de discussão moral no qual aceitam o consequencialismo por um conjunto complexo de razões *r*. Os membros de TG podem ter chegado a um estágio de discussão moral no qual aceitam o deontologismo, da mesma forma, por um conjunto complexo de razões *r*. O ATGM não precisa especificar essas razões. Quem tem tal ônus é justamente quem que se compromete com o RMN. Ao defensor ao ATGM, basta notar que não há consenso sobre qual seria a melhor teoria normativa a se adotar e que, por conseguinte, possivelmente haverão escolhas diferentes.

A conclusão é de que esse ponto de Levy sobre as diferenças psicológicas entre os membros de T e TG não ameaça a *Estipulação Básica*. Acredito que esse modo de atacar o ATGM apelando para a impossibilidade de uma comunidade alternativa como TG é muito interessante e pode colocar sérios problemas para H&T. No entanto, considerando que Levy apela somente para a não suficiência da divergência psicológica, meu raciocínio é: dado que a divergência psicológica não é condição suficiente para assumirmos que T e TG não tem teorias diferentes, e Levy não apresenta nenhum outro motivo, e, além disso, dado que o próprio RMN

nos certifica de que é legítimo fazer tal pressuposição, então os habitantes de T e TG podem ter teorias normativas de primeira ordem diferentes.

Vamos supor que Levy esteja indo mais fundo e queira sustentar, contra o ATGM, que o consequencialismo e o deontologismo são teorias convergentes (como mencionei acima, não acredito que ele tem isso em mente, e nenhuma passagem do seu artigo sugere isso; mas façamos tal suposição apenas com o intuito de afastar possíveis ameaças). Dado que são convergentes, ele poderia dizer, então certamente teríamos que ter a *Intuição da Univocidade Semântica*, pois 'correto-*t*' e 'correto-*tg*' tem o mesmo conteúdo semântico.

Mesmo assim, H&T não estariam em apuros, pois o ATGM independe de qual teoria de primeira ordem se supõe que as comunidades hipotéticas aceitam. Como afirma A. Viggiano,

> […] O argumento de H&T pretende ser totalmente independente de qualquer suposição sobre a natureza da relação da fixação da referência e sobre as propriedades que se supõe estarem R-relacionadas com o termo 'correto' na Terra ou na TGM [Terra Gêmea Moral]. O exemplo específico que eles dão (em que as duas propriedades são caracterizadas através de uma teoria consequencialista e uma teoria deontológica, respectivamente) é apenas para fazer com que uma afirmação geral se torne mais vívida (VIGGIANO, 2008, p. 218).

Neste sentido, é possível construir o mesmo argumento supondo que T e TG aceitam teorias morais diferentes. Por exemplo, suponha que os habitantes de T aceitam uma teoria moral de primeira ordem contratualista de acordo com a qual a propriedade R-relacionada ao uso do termo 'correto' em T é a propriedade de *ser justificável e consensualmente aceito pelos membros de T*. E suponha que os membros de TG aceitam uma teoria de primeira ordem de virtudes de acordo com a qual a propriedade R-relacionada ao uso do termo 'correto' em TG é a propriedade de *ser realizada por um agente virtuoso*. As propriedades em questão são diferentes, portanto a *Estipulação Básica* é preservada. Agora, suponha que membros de T e TG entram em uma controvérsia moral: habitantes de T afirmam que '*x* é correto' e habitantes de TG afirmam que '*x* não é correto'. Eles estariam expressando um desacordo moral genuíno? Não vejo porque nossa intuição seria diferente da intuição do experimento original de H&T. Mas essa intuição conflita com o RMN. Se esta teoria é verdadeira, então T e TG não deveriam expressar desacordo moral genuíno, pois o significado dos termos morais é exaustivamente captado pela propriedade natural *N*, e *N* é diferente nos dois mundos. A conclusão é que, como a intuição do desacordo parece legítima, deve ter algo a mais no significado dos termos morais que não é captado pelo RMN. Portanto, mesmo que se altere a suposição das teorias de primeira

ordem aceitas em T e TG é possível gerar o mesmo tipo de problema. Então, se Levy estivesse atacando o ATGM com base na tese da convergência entre consequencialismo e deontologismo, o argumento de H&T também resistiria.

Assim sendo, essa primeira parte do ataque de Levy ao ATGM não representa uma ameaça definitiva e podemos aceitar a *Estipulação Básica*. Mas, mesmo assim, ainda não estamos livres de complicações. Se aceitarmos a *Estipulação Básica*, argumenta Levy, temos um problema ainda maior pela frente. Passemos a ele.

2.2. Divergência Futura

Levy argumenta que se resolvermos aceitar a *Estipulação Básica* seremos forçados a concluir que, a despeito de os habitantes de T e TG usarem termos ortográfica e fonologicamente iguais, os habitantes de TG não tem um vocabulário moral genuíno e nem sequer pensamento moral. Se Levy está certo sobre isso, parece difícil recusar a sua conclusão de que não há univocidade semântica entre 'correto-*t*' e 'correto-*tg*' e que, por conseguinte, P3 do ATGM é falsa.

No ATGM se supõe que as comunidades T e TG não deveriam discordar se o significado dos termos morais fosse exaustivamente captado pela propriedade natural *N*, pois tais propriedades são distintas. Mas, intuitivamente, nos parece que apresentam um desacordo substantivo genuíno. Portanto, conclui-se que há algo a mais no significado dos termos morais do que a propriedade natural *N* é capaz de rastrear. Levy sustenta que se levarmos a sério a *Estipulação Básica* e "entendermos propriamente" (LEVY, 2011, p. 140) o que ela implica, temos de concluir que não há desacordo genuíno entre as comunidades, mesmo que aceitemos que a *Intuição da Univocidade Semântica* seja persuasiva. Assim, ele insiste que

> [...] os enunciados conflitantes dos habitantes da TGM não expressam o seu desacordo para com nossos enunciados […] porque o seu uso dos termos morais é regulado por um conjunto de propriedades distinto do conjunto de propriedades que regula nosso uso de termos ortograficamente (LEVY, 2011, p. 140).

E, além disso, apresenta um argumento (*debunking argument*) a fim de desmistificar a origem da nossa intuição e explicar sua razoabilidade "aparente".

Para tornar mais explícito e convincente o ponto de Levy de que não há desacordo entre as duas comunidades linguísticas porque os habitantes de TG não são agentes morais, considere a seguinte comparação com a Terra Gêmea de Putnam. Aqui, somos apresentados de antemão ao fato de que o uso do termo 'água' é causalmente regulado pela propriedade de ser H_2O em T e que o uso do mesmo termo em TG é causalmente regulado pela propriedade *XYZ*. Seja qual for a propriedade *XYZ* e, por mais que aquilo a que os habitantes de TG denominam 'água' seja fenomenologicamente similar a H_2O, nossa intuição é de que *não há* água em TG. Por isso, os enunciados '*x* é água' dito por um habitante de T e '*x* não é água' dito por um habitante de TG não são conflitantes, pois um está dizendo '*x* é H_2O' e outro que '*x* não é XYZ'. Se transpusermos isso para o ATGM, podemos obter uma conclusão paralela e, se compreendi corretamente, essa seria a principal razão de Levy para sustentar que não há desacordo genuíno entre os habitantes de T e TG.

Mas por que, então, a *Intuição da Univocidade Semântica* é extensamente compartilhada e persistente? Para o ponto anterior ser plausível, Levy precisa apresentar uma explicação de porque somos amplamente "iludidos" quando submetidos ao cenário descrito por H&T. Essa é a parte central do desafio de Levy ao ATGM e que chamo de *Argumento da Não Localização Temporal*. Este argumento é suficientemente descrito na seguinte passagem.

> A história desses dois planetas deve divergir no decorrer do tempo. As diferenças psicológicas entre nós e os habitantes da TGM seriam suficientes para forçar uma divergência, o que colocaria os planetas em trajetórias bastante diferentes; e quando adicionamos a isso o fato de que as diferenças psicológicas acarretam diferenças nas instituições e práticas, a divergência se torna bem radical. Portanto, se encontrássemos os habitantes da TGM no futuro de sua (e nossa) história, as diferenças entre nós e eles seriam grandes. Essas diferenças seriam tão grandes que acredito que seria correto dizer que haveria pouca razão para pensar os nossos termos morais tem a mesma referência (LEVY, 2011, p. 143-144).

Aqui Levy usa a expressão 'diferenças psicológicas' como sinônimo para 'propriedades distintas que regulam o uso dos termos morais'. Como vimos, diferenças psicológicas não são causa para o fato de que os termos morais nas duas comunidades sejam regulados por propriedades distintas, mas apenas uma sugestão explanatória. Mas isso não refuta o presente ponto de Levy. Sua objeção seria melhor colocada dizendo que, dado que diferentes propriedades *N* causalmente regulam o uso dos termos morais dos habitantes de T e de TG, então, no decorrer da história, os mundos tomariam trajetórias diferentes, pois haveria implicações diferentes para como se organizam as instituições e práticas sociais em geral. A

ideia de Levy é que, o ATGM se beneficia do fato de não estabelecer claramente em que período da história de T e TG os leitores devem imaginar o suposto desacordo moral (LEVY, 2011, p. 143). Se formos apresentados ao experimento num período em que os habitantes de T e TG acabaram de definir a melhor teoria moral de primeira ordem sobre a definição dos termos morais - consequencialista em T e deontológica em TG - a IUS parece ser uma opção bastante atrativa[40]. No entanto, se imaginarmos o suposto desacordo num período razoavelmente posterior da história, dado que a diferença nas propriedades que regulam o uso dos termos morais em T e TG implicariam, segundo Levy, na organização prática e institucional "radicalmente diferente", ficaria bem mais difícil encararmos os termos morais como sendo semanticamente equivalentes. Portanto, a IUS se tornaria bem implausível. Então, em termos simples, o que explica a aparente razoabilidade e ampla aceitação da intuição do desacordo genuíno é o fato de o ATGM não estar precisamente localizado no tempo; o fato de H&T deixarem este ponto em aberto, supostamente distorce nossa intuição a favor de que 'correto-*t*' e 'correto-*tg*' tem significado similar.

Se quisermos uma organização canônica do *Argumento da Não-Localização Temporal*, podemos expressá-lo do seguinte modo.

> P1. Se as propriedades *N* R-relacionadas ao uso dos termos morais em T e TG forem diferentes, então, ao longo do tempo, isso implica que T e TG tomarão trajetórias distintas e serão "radicalmente diferentes" (por exemplo, no que diz respeito à organização das instituições, práticas morais, etc).
>
> P2. As propriedades naturais R-relacionadas ao uso dos termos morais são diferentes em T e TG (*Estipulação Básica*).
>
> > C1. Portanto, ao longo do tempo, as diferenças entre T e TG serão "radicalmente diferentes".
>
> P3. A IUS parece plausível se, e somente se, o ATGM não estiver localizado temporalmente.
>
> P4. Horgan e Timmons não especificam a localização temporal do ATGM.

[40] Essa é *quase* a nossa condição atual enquanto leitores do ATGM; digo 'quase' porque não há consenso entre os especialistas em filosofia moral sobre qual teoria é melhor, se é que devemos escolher entre as disponíveis atualmente no mercado.

C2. Portanto, a IUS parece plausível.

P5. A IUS não é plausível num cenário em que há "diferenças radicais" entre T e TG.

P6. Se localizarmos o ATGM no futuro, haverão "diferenças radicais" entre T e TG.

C3. Portanto, se supusermos o ATGM num cenário futuro, a IUS não será plausível.

Note que as premissas centrais deste argumento são P1 e P3. É necessário uma explicação adicional sobre P3. Por que a IUS faz sentido apenas se o argumento de H&T não estiver localizado temporalmente? Acredito que a melhor explicação é a seguinte. Ainda não atingimos uma posição epistêmica em que possamos saber em definitivo qual teoria moral de primeira ordem melhor capta a propriedade natural R-relacionada com o nosso uso dos termos morais. Então, tanto a teoria consequencialista quanto a deontológica representam possibilidades epistêmicas para nós hoje. Portanto, ao sermos submetidos ao ATGM hoje, dado que ele não está localizado temporalmente, faz muito sentido supormos uma proximidade maior entre T e TG, pois tendemos a imaginá-lo em nossa condição epistêmica atual. O grande compromisso de Levy é que o que resulta da diferença das propriedades morais que regulam o uso dos termos morais em T e TG é que, no futuro, as diferenças entre os dois mundos serão "radicalmente diferentes". Se isso for realmente verdadeiro, parece muito plausível supormos que a IUS não passa de uma fantasia.

D. Brink (2001) parece sugerir algo similar ao argumento de Levy, embora faça apenas uma sugestão em nota e não desenvolva o ponto suficientemente a explicar a origem da "ilusão" da intuição tal como Levy. Ele escreve:

> A diferença entre Terra e Terra Gêmea Moral não parece ser meramente composicional. Presumivelmente, diferentes pessoas, ações e instituições irão satisfazer padrões consequencialistas e deontológicos. Se as pessoas tem o mesmo comprometimento com a moralidade na Terra e na Terra Gêmea Moral, os padrões divergentes irão fazer com que as pessoas de cada planeta considerem pessoas, ações e instituições de forma diferente; ao longo do tempo, isso deve afetar o curso da história individual e social da Terra e da Terra Gêmea Moral [...] Diferenças mais extensas entre a Terra e a Terra Gêmea Moral podem complicar o argumento de Horgan e Timmons (BRINK, 2001, p. 165s, Nota 21).

O ponto de Brink não é exatamente o mesmo de Levy. Levy está sustentando que a diferença ao longo do tempo nas duas comunidades irá diferenciá-las de forma tal que, num cenário futuro, não teríamos a IUS. Brink está sustentando que essa diferença no futuro é uma dissimilaridade com o experimento de Putnam, pois, enquanto a diferença entre T e TG no experimento de Putnam é meramente composicional (H_2O e XYZ), supostamente a diferença entre T e TG no experimento de H&T implicaria, ao longo do tempo, em divergências práticas mais significativas. Brink não nos diz exatamente que tipo de complicação isso poderia trazer para o argumento de H&T. Portanto, vejo apenas duas alternativas: ou ele está sustentando que o experimento de H&T não preserva analogia com o experimento de Putnam (como argumentam Margolis, *et*. *al*., 1999) ou está sustentando que essa diferença, a longo prazo, implicaria no afastamento da IUS. A primeira opção, como vimos no capítulo anterior, não é plausível. Vamos supor, então, que Brink esteja sustentando algo similar a Levy, embora não esteja sendo explícito, para ver se este tipo de argumento realmente expõe um problema para H&T.

Na minha visão, o *Argumento da Não-Localização Temporal* não é um problema para o ATGM. Portanto, essa segunda parte do desafio de Levy a H&T pode ser recusada também. Abaixo, argumento que temos boas razões para recusar P1 e P3 do referido argumento. Meu ponto é que (i) temos razão para acreditar que, mesmo que diferentes propriedades N regulem o uso dos termos morais em T e TG, ainda assim, não haverá "diferenças radicais" entre os dois mundos ao longo do tempo e que (ii) mesmo que houvessem tais diferenças, a IUS seria mantida.

Mas antes de ir ao ponto, primeiro quero chamar atenção para um problema inicial do desafio de Levy, que é sobre a sua interpretação do ATGM. Nas últimas linhas da passagem acima, ele tenta fornecer uma explicação do porquê a nossa IUS seria diferente da defendida por H&T se tivéssemos apreciado corretamente os detalhes não explícitos do experimento. Note que, ao invés de colocar a questão como se os termos morais tivessem um *significado* similar, ele coloca a questão como se os termos tivessem a mesma *referência*. Como notou M. Rubin (2014, p. 292), essa não é uma interpretação justa do ATGM. Ao fazer isso, Levy cria um viés contra a IUS. Em primeiro lugar, 'correto-*t*' e 'correto-*tg*' são predicados. Portanto, é uma questão em aberto se esses termos realmente têm referência. Eles podem ter características relevantes para o significado que não a extensão. Considere, por exemplo, uma tese expressivista sobre o significado dos termos morais. Na verdade, é justamente este o ponto que H&T querem extrair com a IUS e o seu modo de expor a questão preserva essa possibilidade.

Levy coloca a questão como se tudo o que fosse necessário para termos a IUS fosse a mesma referência. Isso já é enviesar a questão a favor do RMN, pois se esse fosse o caso, e lembre que as referências de 'correto-*t*' e 'correto-*tg*' realmente são diferentes (*Estipulação Básica*), então de fato a nossa intuição estaria totalmente errada. O que Levy exclui, com isso, é a possibilidade de que, para um expressivista, mesmo que dois termos tenham extensões diferentes, isso não implica que não compartilhem características semânticas relevantes (lembre do experimento de Hare sobre os Missionários e Canibais, (HARE, 1952, p. 148-149). Portanto, a questão sobre a plausibilidade da nossa intuição seria melhor colocada como se os termos 'correto-*t*' e 'correto-*tg*' tivessem o mesmo *significado*, e não a mesma referência.

Dito isso, sigamos adiante. Levy (assim como Brink) assume que haveriam diferenças relevantes em T e TG no decorrer da história, dada a *Estipulação Básica*. No entanto, ele não nos diz exatamente como isso aconteceria. Não mostra porque, do fato de os termos morais serem R-relacionados com propriedades distintas em T e TG, se segue que os mundos seriam organizados de maneira diferente e nem quais diferenças seriam essas. Isso não é meramente um detalhe irrelevante do argumento, pois não é nada óbvio porque qualquer diferença na história de T e TG nos levaria a concluir que a IUS é ilusória, como afirma Rubin (RUBIN, 2014, p. 293s). O ponto de Levy é que a divergência histórica entre T e TG mostraria a não univocidade entre 'correto-*t*' e 'correto-*tg*'. Mas isso não parece ser uma condição suficiente para a não univocidade. Vejamos porquê.

Rubin apresenta uma objeção a Levy supondo situações contrafactuais em que há diferença entre T e TG e tentando imaginar como isso traria implicações para a univocidade entre 'correto-*t*' e 'correto-*tg*'. Ele diz:

> Suponha que nós, Terráqueos, descobríssemos que a história da Terra Gêmea Moral não teve eventos análogos à Reforma Protestante, Revolução Americana, Segunda Guerra Mundial ou bombas atômicas a grandes cidades. A ausência de tais eventos iria indicar que as histórias dos dois planetas divergiram grandemente. Mas, ainda assim, esse fato não nos dá razão nenhuma para duvidar que 'correto-*t*' e 'correto-*tg*' sejam unívocos (RUBIN, 2014, p. 293).

No entanto, a objeção de Rubin não atinge Levy diretamente. Essas suposições contrafactuais são apenas teorização de cenários possíveis. O ponto de Levy, é que o decorrer da história seria "radicalmente diferente" em T e TG *como implicação direta da divergência entre as propriedades N que regulam o uso dos termos morais*. Em outras palavras, essas diferenças decorreriam diretamente da *Estipulação Básica*. Os exemplos supostos por Rubin não mostram que essas diferenças seriam implicação da *Estipulação Básica*. (Ou melhor, não é nem um

pouco plausível supor que houve, ou não houve, uma Reforma Protestante ou Revolução Americana em determinado mundo porque se aceitava uma teoria moral de primeira ordem consequencialista ou deontológica). Portanto, Rubin não atinge a posição de Levy diretamente.

Agora, consideremos P1 do *Argumento da Não-Localização Temporal* mais especificamente. Ela afirma que, dada a *Estipulação Básica*, as duas comunidades tomariam trajetórias "radicalmente diferentes". Penso que não há boas razões para supor que isso seja o caso. Pelo contrário, acredito que temos motivos relevantes para concluir que T e TG *não* seguiriam trajetórias diferentes.

É plausível supor que T e TG iriam seguir rumos diferentes apenas em um cenário possível: se as propriedades R-relacionadas ao uso dos termos morais nos dois mundos fossem *de fato* "radicalmente diferentes". Por exemplo, se a teoria moral adotada em T fosse consequencialista e a propriedade R-relacionada ao uso dos termos morais fosse a propriedade *maximizar o agregado de felicidade* e a teoria moral adotada em TG fosse uma teoria bem estranha segundo a qual a propriedade R-relacionada ao uso dos termos morais é a propriedade de *maximizar o maior agregado de sofrimento possível*. Obviamente, os mundos iriam diferir "radicalmente" no curso do seu desenvolvimento. Mas, claramente, pode ser objetado que a teoria de primeira ordem em TG não é propriamente uma teoria *moral*. Sendo assim, suponha que os habitantes de TG aceitam um tipo de Egoísmo Moral e que a propriedade R-relacionada ao uso dos termos morais é a propriedade de *ser um ato que promove os próprios interesses do agente mesmo ao custo dos interesses de outros*. Neste caso, não se pode dizer que a teoria de primeira ordem em TG não é uma teoria moral (ao menos, ela aparece no catálogo de teorias morais em uma série de respeitáveis filósofos morais: Feldman, 1978, p. 80s; Kagan, 1998, p. 194s; Ross, 1930, p. 7, 16; Sidgwick, 1907, 107s). Neste caso, é bastante plausível supor que os mundos iriam diferir em suas trajetórias, pois é fácil imaginar circunstâncias em que algo que seria considerado 'correto' em TG seria, ao mesmo tempo, reprovado em T. Considere, por exemplo, ações que visam apenas o benefício individual com custos para terceiros. A ideia de que os mundos poderiam diferir em suas trajetórias parece aceitável nesses casos.

Todavia, isso não é o caso no ATGM. As propriedades R-relacionadas ao uso dos termos morais no experimento de H&T co-referem instâncias em uma ampla variedade de casos, como sugere Rubin (RUBIN, 2014, p. 294s). Torturar alguém apenas por diversão, roubar, insultar e enganar, por exemplo, evidentemente seriam co-instâncias da propriedade R-relacionada ao uso do termo moral 'incorreto' em T *e* em TG. Ajudar pessoas em situações vulneráveis, ser educado, honesto, tratar as pessoas com igualdade e assim por diante, seriam

co-instâncias de ‘correto’ *nos dois mundos*. Note que torturar uma pessoa, por exemplo, é *tratá-la como um meio* e não *maximiza do agregado de felicidade*. O mesmo é válido para os outros exemplos. Assim, podemos imaginar inúmeras instâncias em que ambas as teorias morais de primeira ordem, de T e TG, não apresentariam variabilidade na classificação das instâncias de ‘correto’, ‘incorreto’ e ‘bom’. Isso é um obstáculo à ideia de Levy de que os mundos iriam “divergir radicalmente” ao longo do tempo. Mas note, dizer que há um conjunto significativo de casos em que a extensão de ‘correto-*t*’ e ‘correto-*tg*’ irá coincidir não implica em eliminar a *Estipulação Básica*. A propriedade N de *tratar os outros como um fim em si* e a propriedade *N* de *maximizar o agregado de felicidade* continuam sendo diferentes. Seria até possível aceitar que habitantes de T e TG forneceriam respostas diferentes em algumas circunstâncias. Mas dada a convergência ampla entre os dois mundos, esses casos particulares não parecem ser suficientes para, a longo prazo, distanciar “radicalmente” a organização institucional e prática moral de T e TG. Parece que os casos em que haveria comunalidade na extensão dos termos morais prevalecem significativamente. Portanto, considerando que Levy não nos mostra como os mundos iriam divergir “radicalmente”, dada a *Estipulação Básica*, e dado que temos razões a favor de que *não* iriam divergir, é mais plausível concluir que a convergência entre as duas comunidades seria mantida. E que, assim, a IUS seria preservada, mesmo que considerássemos o ATGM no futuro.

Agora, consideremos a segunda parte do argumento de Levy, a parte em que ele fornece uma explicação de que nossa intuição parece plausível porque falhamos em apreciar completamente os detalhes do ATGM. Ele sustenta que se localizarmos o argumento de H&T num cenário futuro, a IUS não nos parecerá uma alternativa atraente, pois os mundos terão seguido rumos diferentes como resultado da diferente extensão dos termos morais. Em primeiro lugar, o argumento que apresentei acima contra P1 já desempenha um papel importante para bloquear a conclusão de que as comunidades irão diferir substancialmente e que isso irá afastar a nossa intuição. Mas vamos supor que Levy tenha alguma razão e que T e TG acabem tomando rumos diferentes ao longo do tempo. Isso removeria a IUS? Penso que não.

Para sustentar meu ponto, irei introduzir, agora, o *Argumento Invertido da Terra Gêmea Moral*. Este argumento é uma variação do ATGM e tem inspiração no trabalho de vários filósofos (DREIER, 1990, p. 7; SMART, 1981, p. 458; LENMAN, 1999, p. 445; HENNING, 2011, p. 721; RUBIN, 2014, p. 295).

Mesmo nos casos em que ‘correto-*t*’ e ‘correto-*tg*’ irão diferir em sua extensão, não devemos esquecer que o *papel prático* desempenhado pelos termos morais, seja em T ou em

TG, permanece o mesmo. Na verdade, esta é uma estipulação do próprio experimento de H&T. Isso significa, em última instância, que mesmo nos casos em que a extensão de 'correto-*t*' e 'correto-tg' varie, e mesmo que isso aconteça num futuro *t*, essas características continuarão presentes. E, o mais importante, elas irão contar a favor da IUS.

Lembre que H&T estipulam pelo menos quatro características que constituem esse papel prático (H&T, 1992, p. 164). Chamemos essas características de *Condições Semânticas Formais* (CSF). São elas:

(i) os termos morais são usados para *avaliar* ações, pessoas, grupos e instituições;
(ii) os termos morais são usados para discutir sobre considerações que tem a ver com o *bem estar*;
(iii) os falantes que usam o vocabulário moral estão normalmente *dispostos a agir* de modo correspondente aos seus juízos sobre o que é 'correto' ou 'incorreto';
(iv) os falantes tomam os atos aos quais aplicam os termos 'correto' e 'incorreto' como sendo especialmente importantes para decidir *o que fazer*;

Rubin (2014, p. 295), apresenta ainda uma quinta CSF que parece bem plausível:

(v) os falantes estão dispostos a sentir culpa e vergonha quando realizam uma ação a qual aplicam o termo 'errado' (e sentir desaprovação e ressentimento quando outros realizam tal ação).

Essas características dos termos morais não são mero recurso argumentativo de H&T para "direcionar" a nossa intuição a favor da univocidade semântica. Elas são indissociáveis ao discurso e pensamento moral. Portanto, mesmo que a extensão dos termos morais seja diferente em T e TG, e mesmo que localizemos o ATGM num futuro *t*, se tratarmos os termos do experimento em questão como termos *morais*, essas CSF estarão presentes. E a sua presença é suficiente para gerar a IUS.

Dado que o meu ataque a Levy depende da tese de que CSF são indissociáveis dos termos morais, se poderia recusar meu ponto dizendo que as cinco características acima não são

parte necessária do significado dos termos morais. Por isso, temos que notar a força do *Argumento Invertido da Terra Gêmea Moral*[41].

Suponha que há uma Terra Gêmea Moral*. O uso dos termos morais pelos habitantes de TG* está R-relacionado com a mesma propriedade R-relacionada com o uso dos nossos (T*) termos morais. Suponha que em T* e TG* o uso do termo 'incorreto' é causalmente regulado por uma propriedade captada por uma teoria moral de primeira ordem consequencialista segundo a qual a propriedade natural N é a *não maximização da felicidade*. Agora, suponha que as CSF são completamente irrelevantes para os habitantes de TG*. Quer dizer, embora os habitantes de TG* apliquem 'incorreto', e outros termos morais correspondentes, aos mesmos estados de coisas que nós aplicamos, eles não consideram a aplicação de tal termo a uma ação como uma consideração a favor de evitar tal ato (3), de decidir o que fazer (4) ou de sentir vergonha ou culpa quando realizam um ato dessa natureza (5). Além disso, a aplicação desses termos a pessoas, instituições, grupos ou ações estão totalmente desvinculadas de suas discussões sobre o bem estar (2) e eles claramente não estão avaliando algo quando aplicam esses termos (1). Note que ambos os termos, 'incorreto-*t**' e 'incorreto-*tg**', tem a mesma extensão, pois seu uso está R-relacionado com a mesma propriedade natural N. A única diferença é que, enquanto nós consideramos as CSF como características relevantes dos termos morais, os habitantes de TG* não consideram.

Agora, imagine que os habitantes de T* afirmam '*x* é incorreto' enquanto que os habitantes de TG* afirmam '*x* não é incorreto'. Estão eles expressando um desacordo moral substantivo genuíno? Dado que é a mesma propriedade N que está R-relacionada com o uso dos termos morais para os habitantes de T* e TG* e supondo que a propriedade natural capta o significado dos termos morais exaustivamente - tal como sustenta o RMN - então os enunciados conflitantes das duas comunidades claramente expressam um desacordo moral genuíno. Mas qual é a nossa intuição diante desse cenário hipotético? Estou convencido de que a nossa intuição é de que T* e TG* *não* expressam desacordo genuíno quando afirmam os enunciados conflitantes. Os habitantes de TG* estão predicando algo muito estranho de *x*, embora o que quer que seja que estão predicando seja expresso por um termo ortograficamente similar ao nosso termo 'incorreto'. Não parece que os habitantes de TG* estão atribuindo uma propriedade *moral* ao ato *x*, mas algo diferente do que nós atribuímos quando dizemos que um ato é 'errado'.

[41] Este experimento é similar ao experimento de H&T em tudo, exceto no seguinte: onde o experimento de H&T supõe comunalidade (CSF) este supõe divergência; e onde o argumento de H&T supõe divergência (propriedades *N* R-relacionadas ao uso dos termos morais) este supõe comunalidade.

Em outras palavras, poderíamos dizer que os habitantes de TG* não são *morais*, mas estão engajados em algum tipo de prática específica que não a moralidade.

A grande lição que devemos tirar disso é que essas CSF não são apenas um dispositivo argumentativo a favor da intuição que H&T pretendem extrair com o ATGM, mas características essenciais dos termos morais. Portanto, as CSF estarão presentes sempre que termos *morais* estiverem em questão. Sua ausência num cenário de aparente desacordo moral torna claro que, na verdade, não há desacordo porque uma das partes não está engajada com moralidade. Sua presença é suficiente para identificarmos um dado discurso como sendo *moral*. Então, quer imaginemos o cenário estabelecido pelo ATGM no presente ou no futuro essas características terão que ser preservadas. Se forem preservadas num cenário futuro, então não há o afastamento da IUS, mesmo que as comunidades apresentem diferenças. Portanto, a tese de Levy de que o argumento de H&T só parece plausível porque se beneficia da não localização temporal, não é problema para o ATGM.

Podemos, então, concluir que o desafio de Levy a H&T não nos obriga a descartar P3 do ATGM, isto é, não nos força a concluir que a IUS é fruto da nossa falha em apreciar certos detalhes que passam despercebidos quando somos expostos ao cenário do experimento. Sua tentativa de mostrar que, ao assumirmos a *Estipulação Básica*, somos necessariamente levados a dois caminhos, ambos incertos e que trazem problemas para o ATGM, não parece aceitável. Por um lado, sua tentativa de mostrar que as diferenças psicológicas entre membros de T e TG implicam que as duas comunidades têm a mesma teoria normativa fundamental, o que explicaria o suposto desacordo entre elas, não é aceitável, pois: (i) as diferenças psicológicas não são a única explicação possível para se assumir teorias normativas diferentes, (ii) não são relevantes e centrais para o ATGM e (iii) o próprio naturalismo moral nos certifica de que é legítimo assumir que T e TG podem ter teorias normativas diferentes. Por outro lado, sua tentativa de desmistificar a origem da IUS sugerindo que o ATGM se beneficia da não localização temporal, que "distorce" a nossa intuição contra o RMN, também não parece trazer grandes problemas para o ATGM, pois: (i) temos mais razão para pensar que T e TG não seguirão rumos "radicalmente diferentes", dada a co-extensionalidade moral entre as teorias normativas adotadas nas duas comunidades, do que seguirão e que isso seria suficiente para afastar a nossa intuição de que há um desacordo moral genuíno entre eles; e (ii) o *Argumento Invertido da Terra Gêmea Moral* nos mostra que, mesmo se T e TG seguissem rumos diferentes no decorrer do tempo, ainda assim, as CSF dos termos morais teriam que ser preservadas, o que manteria a nossa IUS. Assim, o defensor do ATGM não tem nada a temer do desafio de Levy.

3. A. Viggiano e o conteúdo real da nossa intuição

Ethical Naturalism and Moral Twin Earth (2008), Andrea Viggiano argumenta que o conteúdo *real* das nossas intuições sobre o experimento de H&T é o de que não há desacordo genuíno entre membros de T e TG e que a IUS parece plausível apenas devido às nossas limitações epistêmicas. Ele sugere que a intuição do desacordo é enviesada pelo nosso estágio atual de investigação moral e supõe um cenário que visa eliminar tal dificuldade, o cenário do *Fim da Investigação Moral*. Aqui, as limitações epistêmicas seriam excluídas e poderíamos apreciar o experimento de forma pura e imparcial. Segundo Viggiano, nesta condição se tornaria claro porque não há desacordo entre T e TG e porque a IUS é uma farsa.

No que segue, (i) apresentarei detalhadamente o ataque de Viggiano ao ATGM e (ii) argumentarei que ele não obtém sucesso. Irei sustentar que seu *Argumento do Fim da Investigação Moral* é vulnerável a uma série de contraexemplos que evidenciam a fraqueza do seu ponto.

3.1. O Argumento do Fim da Investigação Moral

Antes de irmos diretamente ao argumento de Viggiano, quero tornar explícito o seguinte ponto: as CSF, apenas, embora sejam condições necessárias para classificarmos um termo ou conceito como sendo *moral*, não são suficientes. Para ver isso, considere o seguinte caso. Variando em certa medida o exemplo de P. Foot (FOOT, 1958, p. 512), suponha que há uma comunidade alternativa (chamemos *Mundo Estranho*).

> *Mundo Estranho (ME)*: os habitantes deste mundo usam termos tais como ‘certo’, ‘errado’, ‘bom’ etc. e, além disso, tal como o nosso uso desses termos, eles cumprem as CSF; isto é, os falantes tendem a usar tais termos para fazer avaliações, sentem culpa quando realizam algo a que é apropriado a aplicação do termo ‘errado’ e assim por diante. No entanto, os habitantes desse mundo aplicam o termo ‘errado’ apenas aos atos, muito estranhos, de “olhar para ouriços na luz do luar”. Podemos dizer que o termo ‘errado’, tal como usado pelos habitantes dessa comunidade, é R-

> relacionado a propriedade de "olhar para ouriços na luz do luar" e esta propriedade é o que regula o uso de tal termo.

Note que, mesmo que tais termos cumpram as CSF, que são condições necessárias para os termos serem considerados morais, não diríamos que, em *Mundo Estranho*, 'errado' é um termo *moral* (e este é o ponto que Foot está tentando esclarecer com este tipo de exemplo). Na verdade, dificilmente diríamos que os habitantes de *Mundo Estranho* tem *pensamento moral*, pois se quiséssemos fornecer uma resposta para a questão 'sobre o que é a moralidade?' ou 'um dado vocabulário é moral quando diz respeito a que?', embora pudéssemos apresentar várias respostas diferentes, não parece que defenderíamos seriamente que moralidade tem a ver com olhar para ouriços na luz do luar (a resposta de Foot é que um vocabulário conta como moral se tem a ver com causar danos e benefícios (FOOT, 1958, p. 510)).

Como nota Rubin (RUBIN, 2014, p. 297), o que o experimento de Foot nos ensina é que não basta apenas um predicado cumprir as CSF para ser considerado como moral; os habitantes de ME fazem avaliações com 'errado' condenando quem olha para ouriços na luz do luar, tendem a sentir culpa ou vergonha se eles mesmos o fazem, se sentem internamente motivados com juízos sobre 'errado', e assim por diante e, mesmo assim, seu vocabulário não é moral. O que mais, então, um predicado deve cumprir para ser genuinamente moral? Há *Condições Semânticas Substantivas* (CSS) que também devem ser cumpridas. Essas CSS especificam que tipo de entidade pode ser extensão de um predicado. No nosso exemplo do *Mundo Estranho*, 'olhar para ouriços na luz do luar' especifica que tipo de entidade (neste caso, ações) são extensão de 'errado'. Ou, se assumirmos uma teoria de primeira ordem consequencialista do tipo que temos considerado, a CSS de 'correto' é 'maximizar o agregado de felicidade' e isso é o que especifica quais entidades (ações) poderão ser categorizadas como (ou extensão de) 'corretas'. Assim, embora possa não ser muito claro que tipo de CSS os predicados morais devem cumprir, considerando cenários possíveis como ME, é plausível aceitar que algum tipo de CSS, para além das CSF, eles devem cumprir.

Tendo notado que tanto as CSF quanto as CSS são necessárias para a classificação de um termo ou conceito como sendo *moral*, podemos passar ao argumento de Viggiano contra H&T. O ponto de desacordo entre Viggiano e H&T pode ser colocado nos seguintes termos: para H&T, o conteúdo real da nossa intuição é de que há desacordo genuíno entre T e TG e, para Viggiano, o conteúdo real da nossa intuição é de que *não* há desacordo genuíno entre T e TG. Para sustentar sua tese, Viggiano apresenta um argumento em dois passos.

O primeiro consiste na apresentação de duas variantes do experimento de H&T e na explicação do que isso significa. A primeira variante é esta:

> *Terra Gêmea Moral Subjetivista (TGMS)*: suponha que o uso de 'correto' em TG é R-relacionado com uma propriedade natural *N* que é captada por uma teoria de primeira ordem subjetivista S*. A propriedade que regula o uso de 'correto-*tg*' é a propriedade de *ser um ato permitido de acordo com as regras de comportamento da comunidade do agente*. O uso de 'correto' em T é R-relacionado com uma propriedade natural *N* captada por uma teoria de primeira ordem consequencialista C*. A propriedade que regula o uso de 'correto-*t*', é a propriedade de *maximizar o agregado de felicidade*. Os dois mundos convergem em relação às CSF.

E a segunda:

> *Terra Gêmea Moral Egoísta (TGMEG)*: Suponha que o uso de 'correto' em TG é R-relacionado com uma propriedade natural *N* que é captada por uma teoria de primeira ordem egoísta E*. A propriedade que regula o uso de 'correto-*tg*' é a propriedade de *ser um ato que promove os próprios interesses do agente mesmo ao custo dos interesses de outros*. O uso de 'correto' em T é R-relacionado com uma propriedade natural N captada por uma teoria de primeira ordem consequencialista C*. A propriedade que regula o uso de 'correto-*t*', é a propriedade de *maximizar o agregado de felicidade*. Os dois mundos convergem em relação às CSF.

A partir disso, Viggiano sustenta: "penso que o que os falantes competentes estariam inclinados a dizer sobre esses dois casos é que 'correto', tal como usado na TGM [Terra Gêmea Moral], não é sinônimo com o nosso termo 'correto' e nem mesmo um termo moral" (VIGGIANO, 2008, p. 219). Ou seja, de acordo com Viggiano, falantes competentes na mesma condição do experimento de H&T teriam a intuição de que 'correto-*t*' e 'correto-tg' não são semanticamente equivalentes, não são traduzíveis e, mais do que isso, que 'correto-*tg*', nos dois cenários variantes, não é um predicado *moral* genuíno. Mas porque teríamos essas intuições? A resposta a esta pergunta revela o ponto das duas variantes do experimento de H&T no argumento de Viggiano: as duas variantes mostram que há *Condições Semânticas Substantivas*

(CSS) para os termos morais (VIGGIANO, 2008, p. 220). Ele descreve um conjunto aproximado de tais CSS:

> A ética é sobre o que satisfaz as necessidades básicas humanas (associadas com a sobrevivência e bem-estar básico) ou vários tipos de desejos (wants) ou sobre o que é condição para o desenvolvimento das capacidades humanas distintivas (não sobe tradições estabelecidas). E assumir o ponto de vista ético requer o abandono de um ponto de vista inteiramente centrado na consideração dos interesses particulares. (VIGGIANO, 2014, p. 220-221).

Essas CSS supostamente mostram porque relutamos em aceitar que os habitantes de T e TG discordam nos casos variantes. Dado que essas CSS são, segundo Viggiano, características indispensáveis do uso dos termos morais e especificam que tipo de extensão um predicado moral como 'correto' pode ter, e dado que o termo 'correto-*tg*' nos dois cenários variantes não cumpre tais CSS (note que uma das CSS é o abandono do ponto de vista subjetivo egoísta), não temos a intuição de que há desacordo entre membros de T e TG na TGMS e na TGMEG. Ou, pelo menos, assim pensa Viggiano.

Passemos, então, ao segundo passo do argumento. Note que no experimento de H&T as CSS não são um conjunto genérico sobre os quais temos informações iniciais (tal como no experimento de Viggiano). As CSS são bem definidas: *maximizar o agregado de felicidade* em T e *tratar os outros como um fim em si* em TG. Agora, lembre também do que chamei de *Estipulação Básica* do experimento de H&T: as propriedades R-relacionadas ao uso dos termos morais nos dois mundos são diferentes (consequencialista em T e deontológica em TG). A partir disso, Viggiano apresenta uma explicação para o fato de que a IUS nos parece tão plausível. Eis a passagem chave:

> Sugiro que a plausibilidade da afirmação de H&T de que as pessoas na Terra e na Terra Gêmea Moral discordam moralmente umas com as outras é devida a diferenças entre o presente e o estado imaginado das nossas investigações éticas. O estado das nossas investigações éticas imaginado no exemplo é bastante diferente do seu estado atual: em particular, ele permite excluir algo (abordagens deontológicas sobre a correção) que não podemos excluir atualmente. Mesmo que a teoria consequencialista referida no exemplo seja realmente a teoria que acabemos adotando no fim das nossas investigações normativas, ainda não chegamos a tal conclusão, de modo que o requerimento semântico híbrido não pode ainda se revelar num experimento de pensamento *à lá* Terra-Gêmea: nossas intuições presentes não excluem a hipótese de que a correção tem a ver com a conformidade a um conjunto de regras (VIGGIANO, 2008, p. 222-223).

A ideia é a seguinte. Dado o estado presente do nosso avanço moral, consequencialismo e deontologismo são ambos possibilidade epistêmica como teorias que rastreiam a propriedade natural *N* R-relacionada ao uso dos termos morais. Mas a exigência feita pelo experimento de H&T é de que excluamos uma dessas teorias do nosso mundo (deontologismo) e imaginemos um cenário em que os dois mundos tem teorias bem definidas (ou, para colocar em outros termos, um cenário em que os dois mundos têm CSS diferentes sobre o uso dos termos morais). O resultado é que, dado que não podemos excluir a CSS deontológica, tal como exigida pelo experimento, pois atualmente ela nos é epistemicamente possível, acabamos por concluir que os termos morais podem estar R-relacionados tanto com a propriedade *N* consequencialista quanto com a propriedade *N* deontológica. Isso, supostamente explicaria a atratividade da intuição de que membros de T e TG apresentam um desacordo moral genuíno.

A partir desta parte explanatória, Viggiano sugere um modo de driblarmos essa dificuldade que, segundo ele, torna claro porque os habitantes de T e TG *não* tem um desacordo real e que, portanto, a IUS é apenas um artefato da nossa ignorância. Ele diz:

> Suponha que as nossas investigações éticas tiveram sucesso e que, partindo das nossas conclusões éticas, obtivemos sucesso em especificá-las, de modo que obtemos uma teoria moral completa. Consideramos várias opções sobre a correção (incluindo, podemos supor, as abordagens deontológicas) e acabamos dizendo que as ações corretas são aquelas que maximizam o prazer das pessoas envolvidas. Podemos supor que esta teoria (tal como aquela sobre as características químicas da água) foi amplamente aceita e que até mesmo (novamente, tal como aquela sobre a água) se tornou parte do conhecimento padrão de pessoas educadas. O que diríamos, sob tais suposições, sobre as pessoas na TGM [Terra Gêmea Moral] cujas investigações éticas sobre ações corretas também obtiveram sucesso, mas cuja conclusão foi de que as ações corretas são aquelas que se conformam a um conjunto específico de regras? Penso que diríamos que seus juízos sobre 'correto' não expressam crenças sobre a correção de forma alguma: as pessoas na TGM falam sobre outra coisa (VIGGIANO, 2008, p. 222).

Este seria o cenário do *Fim da Investigação Moral*. Aqui não há teorias concorrentes sobre qual propriedade N está R-relacionada ao uso dos termos morais. A teoria de primeira ordem está muito bem definida. E, por conseguinte, a CSS está muito bem definida (*maximizar o agregado de felicidade* em T e *tratar os outros como um fim em si* em TG). A ideia de Viggiano é que se chegássemos ao *Fim da Investigação Moral* e concluíssemos que é a teoria consequencialista que capta corretamente a propriedade natural N (e a CSS correspondente), então a teoria deontológica seria naturalmente excluída como candidata à teoria que evidencia a propriedade natural *N* (e a CSS correspondente) dos termos morais. Ou, se chegássemos a conclusão de que é a teoria deontológica que capta corretamente a propriedade natural *N* (e a CSS

correspondente) dos termos morais, então a teoria consequencialista seria excluída. O resultado é: nossa intuição sobre T e TG não seria direcionada pela nossa condição atual de ambiguidade epistêmica sobre a teoria de primeira ordem correta e poderíamos dizer, seguramente, que não há um desacordo genuíno entre os habitantes dos dois mundos, pois ambos expressam coisas muito diferentes com seus predicados 'correto-*t*' e 'correto-*tg*'. Portanto, a IUS, que sustenta que o significado dos termos morais dos membros de T e TG é similar, é uma farsa, ou, mera consequência da nossa ignorância epistêmica atual. Para voltar ao primeiro passo do argumento, assim como os habitantes de T e TG, nos cenários subjetivista e egoísta, não expressam desacordo genuíno porque uma parte (TG) não cumpre as CSS sobre o uso correto do vocabulário moral, aqui os membros de T e TG também não expressam desacordo genuíno porque uma parte usa o predicado 'correto' para expressar algo bem diferente do que é expresso pelo uso de 'coreto' da outra parte.

Para tornar mais claro ainda o ataque de Viggiano ao ATGM, podemos organizar o seu *Argumento do Fim da Investigação Moral* do seguinte modo:

> P1. Há CSS dos termos morais que são correspondentes à teoria normativa de primeira ordem que se assume.
>
> P2. No ATGM, as teorias normativas de primeira ordem são diferentes em T e TG.
>
> C1. Portanto, no ATGM as CSS são diferentes para 'correto-*t*' e 'correto-*tg*'.
>
> P3. Num cenário epistemicamente avançado (*Fim da Investigação Moral*) em que se teria uma teoria de primeira ordem final, se teria também uma CSS final correspondente.
>
> P4. Se se teria uma CSS final, então qualquer predicado que não fosse R-relacionado à esta CSS final, não seria um predicado *moral* genuíno.
>
> C2. Dado que no ATGM as CSS finais seriam diferentes em T e TG, então os predicados morais de um mundo não são semanticamente equivalentes ou traduzíveis aos predicados morais de outro mundo.

O raciocínio de Vigginao parece ser o seguinte. Se se assumisse que a teoria final correta é o consequencialismo e a CSS correspondente é *N*, então qualquer predicado que não esteja R-

relacionado a *N* não seria um predicado moral. Se se assumisse que a teoria final correta é o deontologismo e a CSS correspondente é *N**, então qualquer predicado que não esteja R-relacionado a *N** não seria um predicado moral. A IUS sugere que predicados R-relacionados a N *e* a *N** ambos são morais, semanticamente equivalentes e traduzíveis. Como a suposição do *Fim da Investigação Moral* mostra, tais predicados não podem ser ambos morais, não são semanticamente equivalentes e traduzíveis. Portanto, a IUS é enganosa.

Passemos, agora, a (ii).

3.2. Contra o Argumento do Fim da Investigação Moral

Como vimos, o ponto de Viggiano é que se uma teoria moral de primeira ordem for excluída, e a sua CSS correspondente também for, então termos R-relacionados a propriedade *N* (CSS) de tal teoria não são termos morais. Em primeiro lugar, já parece difícil aceitar um cenário em que se teria a teoria moral correta em definitivo. Mas, dado que é uma estipulação do seu argumento, concedamos isso a Viggiano. O que é mais difícil de aceitar, em segundo lugar, é que essas CSS regulam o uso dos termos morais de tal modo que, se descartamos uma determinada teoria moral, então se usarmos ou aplicarmos os predicados morais de acordo com as CSS de tal teoria, não estamos falando sobre moralidade, mas sobre algo diferente (e, consequentemente, não teríamos desacordo moral genuíno com um usuário da teoria moral supostamente correta). Parece que, a ética requer uma noção mais lata sobre essas CSS. Caso contrário, teríamos que excluir uma série de discursos morais como não morais e aceitar que uma série de desacordos morais na verdade não são desacordos genuínos porque uma das partes, ou as duas, simplesmente não está engajada com o discurso moral. É por essas razões que o ataque de Viggiano ao ATGM é vulnerável a contraexemplos que, espero, mostram sua implausibilidade.

Que tipo de condições um bom contraexemplo ao *Argumento do Fim da Investigação Moral* deve cumprir? Pelo menos duas:

(a) a teoria moral de primeira ordem adotada pelos habitantes de TG (e a sua CSS correspondente) é uma teoria que a nossa melhor teorização moral excluiria e;
(b) mesmo assim, temos a intuição de que há um desacordo moral genuíno entre nós e os habitantes de TG.

Como mostra Rubin (RUBIN, 2014, p. 304s), é possível construir cenários deste tipo, o que implica que o *Argumento do Fim da Investigação Moral* não é sólido. Assim, contra Viggiano, proponho três contraexemplos, em primeiro lugar. E, em segundo, busco apresentar uma justificativa adicional do porquê a intuição de que há desacordo com os habitantes de TG, mesmo nesses casos em que a teoria de primeira ordem já foi excluída, nos é tão plausível.

Considere o seguinte experimento, tal como proposto por Rubin (Rubin, 2014, p. 304s):

> *Terra Gêmea Moral Anti-Consequencialista (TGMAC)*: no fim de sua investigação moral, os habitantes de TGMAC acabaram adotando uma teoria de primeira ordem deontológica D* radicalmente anti-consequencialista. O uso de 'correto' em TGMAC é R-relacionado com a propriedade *N* captada pela teoria D* e entre o conjunto de ações proibidas pelo código de TGMAC está o requerimento de que é errado fornecer ajuda aos outros. O uso de 'correto' em T é R-relacionado com uma propriedade natural *N* captada por uma teoria de primeira ordem consequencialista C* e entre o conjunto de ações proibidas pelo código de T está o requerimento de que *não* é errado fornecer ajuda aos colegas. Os dois mundos convergem em relação às CSF.

Agora, suponha que os habitantes dos dois mundos se encontrem e discutam o status moral de uma ação *x* que consiste em, sem sacrifício algum, salvar a vida de uma pessoa inocente (talvez puxando a alavanca e desviando um trem desgovernado da colisão com tal indivíduo, porém dessa vez, felizmente, para um trilho livre onde não há cinco pessoas inocentes). Os habitantes de TGMAC diriam '*x* não é correto-*tg*' enquanto os habitantes de T diriam '*x* é correto-*t*'. O ponto é: eles expressam um desacordo moral genuíno?

Como responde Rubin, corretamente, "parece-me óbvio que as duas partes estão expressando um desacordo moral substantivo genuíno" (RUBIN, 2014, p. 305). Se expressam desacordo, então o requisito (b) é cumprido. Acredito que o requisito (a) também é cumprido, pois note que Viggiano não poderia responder algo como: mas meu *Argumento do Fim da Investigação Moral* explica porque temos tal tipo de intuição, pois somos enviesados pela situação atual em que há ambiguidade epistêmica sobre a teoria moral de primeira ordem correta e a situação exigida pelo experimento onde não há tal ambiguidade. De fato, uma teoria moral

tal como a adotada em TGMAC é excluída do nosso menu de possíveis candidatos à teoria moral correta e isso, para citar o próprio Viggiano, "foi amplamente aceito e até mesmo se tornou parte do conhecimento padrão de pessoas educadas" (VIGGIANO, 2008, p. 222). Mesmo assim, não parece certo que não haveria desacordo moral genuíno com alguém que endossasse tal teoria. Portanto, o requisito (a) também é cumprido. Este contraexemplo mostra que P4 é, no mínimo, problemática, pois a teoria de primeira ordem de TGMAC já foi excluída como nossa melhor teoria moral (portanto, sua CSS também), mas não nos parece que 'correto-*t*' e 'correto-*tg*' não são traduzíveis ou que 'correto-*tg*' não é um predicado moral genuíno.

Considere mais um caso similar que é evidência para o mesmo ponto.

> *Terra Gêmea Moral Escravagista (TGME)*: no fim de sua investigação moral, os habitantes de TGMES acabaram adotando uma teoria de primeira ordem D*. O uso de 'correto' em TGMN é R-relacionado com a propriedade *N* captada pela teoria D* e entre o conjunto de ações do código moral de TGMES está o requerimento de que é correto submeter negros, contra a sua vontade, ao trabalho forçado e não remunerado. O uso de 'correto' em T é R-relacionado com uma propriedade natural *N* captada por uma teoria de primeira ordem C* e entre o conjunto de ações do código moral de T está o requerimento de que *não* é correto submeter negros, contra a sua vontade, ao trabalho forçado e não remunerado. Os dois mundos convergem em relação às CSF.

Como no experimento anterior, parece óbvio que essa teoria possível são definitivamente excluídas do nosso menu de teorias morais corretas (ou da teoria moral correta). Portanto, a condição (a) é cumprida. Além disso, parece correto dizer que numa discussão possível entre habitantes de TG e habitantes T sobre o status moral de uma ação em que um deles reivindicasse que *x* é correto e o outro que *x* não é correto, haveria desacordo moral genuíno; não parece que membros de T estariam falando sobre moralidade e membros de TG estariam falando sobre algo completamente diferente. Portanto, a condição (b) também é cumprida.

Esses contraexemplos nos dão boas razões para recusarmos P4 do argumento de Viggiano. Nos três casos temos uma teoria final (e uma CSS correspondente final) sobre o uso dos predicados morais e, mesmo assim, não parece correto dizer que os predicados não R-

relacionados com tais CSS não são predicados morais. Portanto, a conclusão de que 'correto-*t*' e 'correto-*tg*' não são traduzíveis e que a IUSé enganosa não é verdadeira.

Além disso, tais contraexemplos sugerem algo a mais. Para que não tivéssemos a intuição de que há desacordo entre os dois mundos (nos casos apresentados acima) as CSS teriam que constranger o nosso uso dos termos morais num sentido estritamente rigoroso. Quer dizer, para não nos parecer que discordamos dos habitantes de TGME, por exemplo, o nosso uso dos termos morais teria que ser constrangido pela CSS revelada pela teoria C*, *e apenas pela CSS relevada por C**. Teria que ser claro que qualquer outra CSS revelada por outra teoria possível não seria uma CSS moral e que qualquer predicado R-relacionado a tal teoria não seria um predicado moral. Mas, como os contraexemplos mostram, nenhuma CSS constrange tão rigidamente o nosso uso dos termos morais como parece pensar Viggiano. Isso sugere que P4 pode ser atacada por outra frente ainda. Como nota Rubin (RUBIN, 2014, p. 303s), Viggiano não nos dá muitas razões para que aceitemos a sua premissa de que um predicado moral R-relacionado a uma CSS já excluída não seria um predicado moral. O que nos resta supor é que ele está pensando em analogia com os dois casos variantes que ele mesmo apresenta (*Terra Gêmea Moral Subjetivista* e *Terra Gêmea Moral Egoísta*). Segundo ele, não há desacordo nesses casos variantes, pois o uso dos termos morais em TG (nos dois casos) não cumpre o conjunto de CSS que ele mesmo supõe serem parte do uso dos termos morais. Mas será que isso é verdadeiro? Se for, então teríamos que aceitar que não há desacordo genuíno com egoístas e subjetivistas ou relativistas. Este não é um fato que estamos dispostos a aceitar prontamente sem razões adicionais. E, dado que Viggiano não nos dá essas razões, é plausível sustentar que seu ponto exclui como desacordo legítimo ocorrências que realmente parecem desacordos legítimos.

Assim sendo, podemos dizer que o ataque de Viggiano ao ATGM, ao desafiar P3, não é suficientemente persuasivo para nos fazer recusar o argumento.

4. J. Sonderholm e a não confiabilidade da intuição

A estratégia de Jorn Sonderholm (2013) contra H&T tem finalidade similar a de Levy e Viggiano: recusar o ATGM tentando desmascarar a verdadeira origem da intuição gerada pelo experimento. Mas o modo como ele faz isso é diferente. A ideia básica de sua crítica é a seguinte: dado que o cenário do ATGM é "radicalmente diferente", não devemos confiar nas

intuições suscitadas pelo experimento e, para além disso, se excluirmos a diferença entre a nossa condição epistêmica atual e a condição requerida pelo experimento veremos claramente que a IUS é uma farsa. Sonderholm conclui, assim, que o ATGM não é uma ameaça ao RMN.

Nesta seção tentarei fornecer razões para recusarmos o desafio de Sonderholm ao ATGM. Primeiro, apresentarei em detalhes o seu ataque. Depois argumentarei porque é possível recusá-lo.

4.1. Não Confiabilidade e Explicação

O ataque de Sonderholm a H&T é duplo. Por um lado, ele apresenta um argumento para sustentar que devemos ser céticos em relação à intuição causada pelo ATGM. Vou chamar isso de *Argumento do Ceticismo Sobre a Intuição.* Por outro, seu ataque desempenha um papel explanatório. Ele busca mostrar porque é natural que os leitores de H&T tenham a IUS, dadas as condições estipuladas pelo próprio experimento e dada a teoria de Boyd. Vou chamar isso de *Argumento Sobre a Explicação da Intuição*. A pressuposição comum à investida de Sonderholm é de que a nossa situação presente e a situação exigida pelo experimento de H&T é significativamente diferente: nossa situação é aquela em que há amplo desacordo sobre qual teoria moral de primeira ordem capta corretamente a propriedade natural *N* que regula o uso dos termos morais (deontológica, consequencialista, uma teoria de virtudes, alguma teoria comum às três anteriores ou uma teoria diferente de todas essas) enquanto que a situação exigida pelo experimento é aquela em que não há tal desacordo (somos convocados a pensar num cenário possível em que a propriedade R-relacionada ao uso dos termos morais é consequencialista).

Assim, o ponto do primeiro argumento de Sonderholm consiste em sustentar a não confiabilidade da intuição gerada pelo experimento de H&T dada essa diferença modal. A passagem principal é esta:

> Minha sugestão é que, devido à nossa presente situação em relação a 'correto', é difícil, através da nossa imaginação, ter uma noção adequada de como seria estar numa posição em que se tem o conhecimento estipulado por [Horgan e] Timmons. Se a situação descrita por [Horgan e] Timmons fosse apenas marginalmente diferente da nossa situação presente seria relativamente fácil, através da nossa imaginação, ter uma noção acurada do que seria estar em tal posição. Mas a situação descrita por [Horgan e] Timmons é significantemente diferente da nossa situação atual e, portanto, há uma grande chance de que nossa imaginação nos dará uma imagem distorcida do que seria estar em tal posição. Devido a ser difícil ter uma noção apropriada do que seria estar em numa situação em que há amplo acordo de que o uso de 'correto' é causalmente

> regulado por uma propriedade natural *N*, nosso juízo sobre qual seria a nossa intuição em tal situação deve ser tratado com algum ceticismo (SONDERHOLM, 2013, p. 85).

Podemos organizar o *Argumento do Ceticismo Sobre a Intuição* do seguinte modo.

P1. No presente, há desacordo amplo entre os especialistas em ética sobre qual propriedade natural regula causalmente o uso dos termos morais.

P2. O ATGM nos pede que imaginemos um cenário em que o uso dos termos morais em T é causalmente regulado por uma propriedade natural *N* e que o uso dos termos morais em TG é causalmente regulado por uma propriedade natural *N**.

C1. Portanto, o ATGM nos pede que imaginemos uma situação "significativamente diferente" do que a nossa situação atual (P1, P2).

P3. A IUS conta como evidência contrária ao RMN se, e somente se, é um guia confiável sobre a semântica dos termos morais.

P4. A intuição de um experimento de pensamento é confiável se, e somente se, o cenário do experimento não é "radicalmente diferente" do cenário presente.

C2. A intuição gerada pelo ATGM não é confiável (P4, C1).

C3. Se a intuição gerada pelo ATGM não é confiável, então a IUS não é um guia confiável sobre a semântica dos termos morais (P3, C2).

C4. Se não é um guia confiável sobre a semântica dos termos morais, então a IUS não conta como evidência contrária ao RMN. (P3, C3).

Note que o núcleo do argumento é a conjunção de C1, P4 e C2. A ideia aqui é, basicamente, que dada a nossa situação e a situação exigida pelo ATGM, é difícil imaginar que a nossa intuição, neste cenário contrafactual "significativamente diferente", não seria distorcida por influência da situação atual. Essa distorção tem como resultado justamente a IUS. Quer dizer, não estivéssemos em uma situação de desacordo amplo sobre a propriedade R-relacionada com o uso dos termos morais, estaríamos em condição melhor (menos distorcida ou enviesada) para termos uma intuição mais acurada; mas já que somos instados a imaginar

qual seria nossa intuição num cenário contrafactual bem diferente, temos a intuição que temos, ou seja, a intuição de que habitantes de T e TG estão em desacordo moral genuíno.

Até aqui temos o argumento de Sonderholm para explicar porque a intuição não é confiável. Mas ainda falta tornar mais explícito *porque* temos a IUS. É por isso que o argumento é estendido para uma tarefa além: *a tarefa explanatória.* Essa explicação é dada em dois passos. O primeiro deles já foi dado, ou seja: a intuição é resultado (distorcido ou enviesado) da nossa situação atual de desacordo amplo sobre qual propriedade natural causalmente regula o uso dos termos morais e a necessidade de imaginar um cenário "radicalmente diferente". Mas há um segundo passo que estabelece que é normal que haja a intuição do desacordo genuíno. Este segundo passo, que chamo de *Argumento Sobre a Explicação da Intuição*, não é separado do primeiro; é uma extensão.

O *Argumento Sobre a Explicação da Intuição* é estabelecido em duas passagens principais. A primeira é a seguinte:

> […] "correto" é aplicado a um número de coisas que, num nível (nível que pode ser denominado "o nível das propriedades superficiais"), não tem nada em comum. As pessoas, em um tempo e lugar, aplicam "correto" a instâncias *f* (digamos, encerrar a vida de bebês recém-nascidos que tem deformidades físicas severas) enquanto que as pessoas em outro tempo e espaço aplicam "correto" a instâncias de não-*f*. O fato de "correto" ser aplicado a coisas que são frequentemente bem diferentes em termos das propriedades superficiais não exclui, no entanto, o fato de que "correto" é causalmente regulado por uma propriedade funcional fundamental. As coisas às quais "correto" é aplicado podem compartilhar uma propriedade fundamental: talvez a propriedade de ser condutivo ao bem-estar humano. É não problemática a ideia de que coisas que são diferentes no nível das propriedades superficiais possam, em diferentes contextos sociais e econômicos, ser condutíveis ao bem-estar humano (SONDERHOLM, 2013, p. 86).

O que deve ficar claro aqui são duas coisas: (i) no nível superficial os termos morais não rastreiam nenhuma propriedade fundamental unitária; (ii) isso não é inconsistente com o fato de que os termos morais rastreiam uma propriedade fundamental unitária. Parece que Sonderholm entende "nível das propriedades superficiais" como expressando a variabilidade de instâncias a que os termos morais são aplicados e aqui é muito comum vermos um termo moral ('incorreto') e sua negação sendo aplicados às mesmas ações (é incorreto consumir carne animal, alguns sustentam; não é incorreto consumir carne animal, outros sustentam). O discurso moral ordinário é repleto disso. Portanto, de acordo com Sonderholm, parece que nesse nível superficial não há nenhuma característica mais fundamental que os termos morais captam. No entanto, o fato de haver esse emprego divergente dos termos morais na moralidade comum não implica que os mesmos não rastreiem uma propriedade unitária fundamental. Eles podem

rastrear uma propriedade consequencialista (*maximizar o agregado de felicidade*) ou deontológica (*tratar os outros como fins em si mesmos*). O ponto é que ainda não estamos numa posição epistêmica favorável para dizer com certeza que esta ou aquela propriedade está R-relacionada com o uso dos nossos termos morais. Portanto, não há inconsistência entre (i) e (ii). Em outros termos - e este é o ponto de Sonderholm - não há inconsistência entre (i) e a teoria de Boyd ou ao RMN.

A partir disso, Sonderholm faz uma conjunção entre a premissa do desacordo epistêmico presente sobre quais propriedades regulam o uso dos termos morais e a premissa de que os termos morais não rastreiam propriedades fundamentais unitárias no nível superficial para obter a explicação da IUS.

> Agora, dado que, presentemente, não há acordo entre os especialistas sobre qual propriedade funcional regula causalmente o uso de "correto", e dado que "correto", no nível das propriedades superficiais, não rastreia nenhuma propriedade unificadora, penso que é esperado que as pessoas não tenham intuições de designação rígida sobre "correto". Seria estranho se falantes competentes do português, que na maioria são inconscientes da noção de uma propriedade funcional fundamental e que veem "correto" sendo aplicado a coisas que não parecem ter propriedades unificadoras, tivessem intuições de designação rígida sobre esse termo. Dado que a impressão é de que não há uma propriedade unificadora que o uso de "correto" rastreia, seria estranho que falantes competentes do português tivessem a intuição de que seu termo "correto" capta, em todos os mundos possíveis, a mesma propriedade que ele capta no mundo atual (SONDERHOLM, 2013, p. 86).

Como foi estabelecido na citação anterior, no nível dos não especialistas a maioria das pessoas não está familiarizada com a ideia de propriedade fundamental reguladora, e, além do mais, as propriedades superficiais do uso dos termos morais não rastreiam nenhuma propriedade comum, pois são aplicadas muitas vezes a ações antagônicas. Some a isso o fato de que até mesmo no nível dos especialistas em teoria moral não há acordo sobre qual propriedade fundamental regula o uso dos termos morais. Temos que concluir que não estamos *familiarizados* com a ideia de propriedade reguladora unitária. Portanto, Sonderholm conclui, seria muito estranho se se tivesse a intuição de que *não* há desacordo genuíno entre T e TG, pois para não haver desacordo teria que estar bem claro que a propriedade R-relacionada aos termos morais nos dois mundos é diferente. Como não está, então a ideia de que habitantes de T e TG discordam genuinamente parece ser bem atrativa.

Neste sentido, podemos resumir o *Argumento Sobre a Explicação da Intuição* do seguinte modo.

P1. Quando se considera o cenário do ATGM, se tem a IUS.

P2. Se se tem a IUS, então ou a teoria de Boyd (RMN) é falsa ou há alguma outra explicação.

P3. Há uma outra explicação para a IUS.

C1. Portanto, a teoria de Boyd (RMN) não é falsa.

A principal premissa é P3 e as evidências a seu favor estão na explicação acima ((a) há desacordo amplo entre os especialistas sobre a propriedade que regula o uso dos termos morais e (b) no discurso comum as propriedades superficiais não rastreiam nenhuma propriedade fundamental unitária). Finalmente, o ataque de Sonderholm ao ATGM pode ser sintetizado ao seguinte: (i) o ATGM deve ser tratado com ceticismo, pois a IUS não é confiável e (ii) há uma boa explicação do porquê temos a IUS.

4.2. Problemas para Sonderholm

Assim como Levy e Viggiano, Sonderholm fornece um argumento abdutivo como candidato à melhor explicação da IUS. O que melhor explica o juízo intuitivo é a distância modal entre a nossa condição epistêmica atual e a condição requerida pelo experimento de H&T. Irei adotar uma estratégia similar para recusar o ataque de Sonderholm. Vou sustentar que a melhor explicação para a intuição são as CSF. Em uma réplica a Sonderholm, H&T (2015) usam este tipo de estratégia. Neste sentido, vou dividir em duas partes minha defesa do ATGM: (i) irei endossar algumas críticas dos próprios H&T bem como propor alguns argumentos adicionais contra Sonderholm; (ii) vou construir dois casos para testar o alcance da hipótese explanatória de Sonderholm e da hipótese explanatória que estou reivindicando (a que apela para as CSF). A conclusão será de que esses dois conjuntos de críticas, juntamente com a pressuposição metodológica mais geral sobre estratégias abdutivas (que esclareço abaixo), fornece boas razões para recusarmos o ponto de Sonderholm e aceitarmos que o ATGM persiste. Quero deixar claro que a minha defesa de H&T não precisa necessariamente mostrar que a argumentação de Sonderholm é completamente falsa. Meu objetivo é muito mais simples

e modesto. O ponto é apenas sustentar que a explicação da IUS dada por Sonderholm enfrenta uma série de problemas e que, se tivermos uma explicação melhor, devemos adotá-la.

Comecemos esclarecendo qual é a pressuposição metodológica que adoto (H&T, 2015, p. 367). Normalmente, consideramos nossa intuição sobre cenários hipotéticos um guia razoavelmente confiável quando fazemos teorizações filosóficas. Há muitos exemplos de teses filosóficas que se baseiam neste tipo de dispositivo argumentativo. Mas isso não significa que devemos aceitar a intuição como critério determinante a favor da verdade de uma teoria. Podem existir casos em que a intuição é fruto de algum viés, ausência de informação etc. Nesses casos, os argumentos que buscam desmistificar a origem da intuição (*debunking arguments*) nos mostram porque devemos recusar certas ideias que nos parecem certas. No entanto, nos casos em que a intuição é amplamente compartilhada e persistente uma explicação que a justifica é, a princípio, preferível do que uma explicação que a desmistifica. Quem apresenta o argumento para desmistificar a intuição, nestes casos, tem o ônus da prova muito maior de mostrar porque tantas pessoas apresentam um tipo de "ilusão". Neste sentido, uma presunção inicial está a favor do defensor do ATGM e contra estratégias que buscam dar conta da intuição apelando para explicações alternativas que visam mostrar como e porque nós nos enganamos *quando consideramos o cenário do ATGM e temos a IUS.*

Mas, note, isso não é dizer que uma intuição amplamente compartilhada e persistente não possa ser fruto de algum tipo de ilusão. O proponente do argumento desmistificador pode estar certo. No entanto, se o seu argumento apresenta uma série de pressuposições controversas e não é razão claramente convincente contra a intuição em questão, então esse pressuposto metodológico de que, a princípio, temos mais razão para aceitar um argumento que justifique uma intuição amplamente compartilhada do que para aceitar um argumento que a desmistifique, ganha força. Assim, esse primeiro ponto metodológico que adoto funciona melhor em conjunto com outros argumentos (que estão abaixo) do que isoladamente. Assim, minha estratégia é, partindo do pressuposto de que a IUS é amplamente compartilhada e persistente, tentar mostrar que a explicação alternativa de Sonderholm enfrenta problemas (mesmo que não seja obviamente falsa) e que, por isso, devemos adotar a melhor alternativa em termos de alcance explanatório.

Concordo com H&T que a melhor explicação sobre a IUS "é que tais termos e conceitos são primariamente governados por normas semânticas que tem a ver como o papel que tais termos e conceitos desempenham no pensamento e discurso" (H&T, 2015, p. 365). Essas normas semânticas (ou CSF), relembremos, são: (a) termos morais são usados para

avaliar ações, pessoas, instituições, etc., (b) para discutir considerações a respeito do bem-estar, (c) possuem practicalidade intrínseca, (d) desempenham papel importante em contextos em que se busca decidir o que fazer, (e) há certos sentimentos normalmente associados à realização de ações a que termos morais são aplicados (tal como culpa ou ressentimento). H&T ainda adicionam uma característica a mais: (f) juízos morais são categóricos no sentido de que (i) não são mero resultado de desejos e aversões e que (ii) ao discutir assuntos morais, as pessoas afirmam sua posição moral categoricamente, quer dizer, reivindicando verdade ou correção (em oposição à mera opinião). Desse modo, se os termos empregados pelos habitantes de T e TG são termos *morais*, então mesmo que putativamente refiram propriedades *N* diferentes, terão que contemplar tais características. E essas características são o que melhor explicam o nosso juízo de que há desacordo entre T e TG. Mas por que deveríamos aceitar esta explicação e não a de Sonderholm? Porque esta é a *melhor* explicação. Vejamos.

Se quisermos mostrar que o desafio geral de Sonderholm não funciona, temos que mostrar que seus passos específicos têm dificuldades. Como argumentam H&T (2015, p. 365ss), o ataque de Sonderholm ao ATGM sofre de uma série de problemas. Comecemos com a afirmação de Sonderholm sobre a "distância modal" entre a situação presente e a suposição do ATGM. Ele sustenta que essa distância modal mostra (i) porque a intuição não é confiável e (ii) explica porque é plausível, mesmo que errado, termos a IUS. Todavia, Sonderholm precisa apresentar razões do porquê isso é o caso, pois "certamente não é óbvio ou auto-evidente que a distância modal em questão tem qualquer uma dessas características" (H&T, 2015, p. 368). E é aqui que seus argumentos apresentam problemas.

Para provar seu ponto de que, dada a distância modal (i) e (ii) se seguem, Sonderholm apela para um exemplo de E. Kraemer (1990) sobre como uma intuição pode ser indeterminada num cenário hipotético. O ponto de Kraemer parece ser que a intuição pode ser mais ou menos forte, mais ou menos compartilhada ou consistente - e, por conseguinte, confiável - *dependendo do tipo de exemplo* que se tem em mãos e da nossa condição epistêmica atual diante do cenário hipotético. O exemplo de Kraemer é o seguinte. Suponha que descobrimos um mundo cujos habitantes têm um mecanismo de hereditariedade reprodutiva bem complexo e diferente do que associamos, no nosso mundo, ao DNA. Vamos supor que é um composto químico KLM. Assim, enquanto nós associamos o termo 'gene' ao DNA esses habitantes estranhos associam 'gene' a KLM. O ponto é: estaríamos inclinados a dizer, ou teríamos a intuição, que essas criaturas não têm 'genes', dado que não tem DNA? A resposta de Kraemer é que não se sabe. Ele diz:

> Não tenho certeza do que diríamos. Talvez a intuição semântica comum seria de que essas criaturas não têm genes; mas, talvez, não seria. Tal indecisão é instrutiva, pois sugere que pode ser difícil predizer de forma acurada como nossas intuições semânticas irão mudar dadas mudanças radicais no corpo epistêmico e no entendimento comum de tal corpo epistêmico (KRAEMER, 1990, p. 470).

Note que o que Kraemer está dizendo é o seguinte. Não se tem certeza sobre que tipo de intuição as pessoas teriam, pois não sabemos como as pessoas reagiriam ao ser submetidas a uma condição epistêmica que não é a condição atual. A conclusão que Sonderholm extrai disso é:

> Considero o exemplo de Kraemer como sendo tal que … as intuições linguísticas de falantes competentes não são mais um guia confiável do significado, dado o fato de que elas dizem respeito a um cenário radicalmente contrafactual em que as coisas são muito diferentes do que são no mundo atual. No mundo atual, por exemplo, não há acordo amplo sobre a ideia de que traços hereditários são mediados através de mecanismos não associados com o DNA. Reflexão sobre o cenário sugerido por Kraemer é reflexão sobre um mundo tão distante no espaço modal que as intuições geradas pela reflexão não constituem evidências fortes para nenhuma visão particular sobre o significado do termo "gene" (SONDERHOLM, 2013, p. 83).

Ou seja, para Sonderholm, o exemplo de Kraemer mostra que a nossa intuição, num cenário diferente da situação epistêmica atual, não é confiável. Neste sentido, e este é o ponto de Sonderholm quanto ao ATGM, a IUS também não seria confiável, já que supõe uma mudança de condição epistêmica similar.

É muito estranho que Sonderholm se baseie nesse ponto de Kraemer para sustentar que a intuição *presente* das pessoas sobre o ATGM não é confiável, como notam H&T (2015, p. 68s), pois o próprio Kraemer não diz que a intuição não é confiável, mas apenas que *não tem certeza* sobre que tipo de intuição as pessoas teriam sobre o exemplo do gene. O que se poderia inferir, de modo análogo do ponto de Kraemer, não é que a intuição não seria confiável, mas apenas que não se teria certeza sobre que tipo de intuição seria correto extrair do ATGM. Tendo a compreender isso do seguinte modo: ao sermos submetidos ao cenário do ATGM não podemos afirmar com segurança nem que as pessoas teriam a IUS e nem que não teriam. Isso é tudo o que Sonderholm pode extrair do ponto de Kraemer.

Mas o seguinte fato conta contra Sonderholm: ao se submeter as pessoas ao cenário do ATGM não é tão controverso se elas teriam a IUS ou não teriam. Elas normalmente têm. (As estratégias de se atacar o ATGM tentando explicar a origem dessa intuição pressupõem justamente que ela é amplamente compartilhada). Portanto, se tivermos uma boa explicação

sobre a IUS (e parece que a explicação que apela para as CSF é uma boa explicação), somada ao fato de que ela é amplamente compartilhada e persistente, podemos ter razões para aceitá-la, mesmo que Kraemer esteja certo sobre a dificuldade em se determinar qual seria a nossa intuição em *alguns* cenários hipotéticos. O resultado disso, como sustentam H&T, é que o exemplo de Kraemer faz muito pouco a favor de Sonderholm.

Há uma dificuldade adicional para Sonderholm. C1 do *Argumento do Ceticismo Sobre a Intuição* afirma que o cenário que o ATGM nos submete é "significativamente diferente" da nossa situação atual. Mas será que é realmente? Acredito que a diferença não seja significativa. Ou, pelo menos, não a ponto de nos dar razão suficiente para recusarmos a confiabilidade da intuição. Para ver isso consideremos o *Argumento da Terra Gêmea* de Putnam (cujo *insight* os defensores do tipo de naturalismo moral considerado aqui frequentemente aceitam, pois aceitam, e expandem para os termos morais, a teoria externalista da referência). Aqui, assim como no ATGM, também há diferença epistêmica entre a situação atual e a situação exigida pelo experimento. Esta diferença diz respeito ao elemento químico que compõe a água (H_2O em T e XYZ em TG). Note que a diferença epistêmica é muito mais "significativa" no experimento de Putnam do que no ATGM. Naquele, nos é exigido que imaginemos uma comunidade que associa o termo 'água' a uma substância química diferente de tudo o que conhecemos e que jamais tivemos contato. Neste, nos é exigido apenas que suponhamos que 'correto' é captado por uma teoria consequencialista em T e uma teoria deontológica em TG. Sonderholm pode até ter razão ao sustentar que não temos muita intimidade com a ideia de que essas teorias *definitivamente* regulam o uso dos termos morais nas duas comunidades. No entanto, quando se empreende esforços a favor de uma teoria consequencialista ou deontológica a pressuposição é que ela capte suficientemente todas as instâncias de correção. Quer dizer, uma teoria moral, além de ter um objetivo prático (proporcionar um método de tomada de decisões), tem um objetivo teórico, que é responder o que as ações têm que as faz serem certas ou erradas. E este objetivo teórico visa a sistematização da moralidade, isto é, busca mostrar não porque esta ou aquela ação é certa ou errada, mas porque *qualquer* ação que tenha tais e tais características é certa ou errada. Portanto, o pano de fundo de uma teoria moral de primeira ordem é ser *definitiva* (captar aquilo que confere status moral para qualquer ação), mesmo que ainda não tenhamos chegado à condição epistêmica de dizer com certeza qual seria esta teoria.

Dito isto, é plausível dizer que a diferença epistêmica entre a condição atual e a condição exigida no cenário do experimento é muito maior no experimento de Putnam do que no ATGM. E, se diferença epistêmica é condição suficiente para sermos céticos em relação à

intuição, como mantém Sonderholm, então temos muito mais razão pra sermos céticos em relação à intuição do experimento de Putnam. Mas, o experimento de Putnam é amplamente aceito. A teoria resultante que se extrai do experimento pode ser, e é, atacada de diversas frentes, mas pelo menos é concedido ao experimento valor explanatório. A razão disso é que a intuição do experimento de Putnam é largamente compartilhada e consistente e, portanto, há ampla aceitação de que é confiável. Assim sendo, se não recusamos a intuição de um experimento cuja diferença epistêmica entre a situação atual e a exigida pelo cenário imaginado é realmente significativa, então não devemos recusar a intuição de um experimento cuja diferença epistêmica é menos significativa. Portanto, devemos dar crédito ao *insight* do ATGM.

O argumento de Sonderholm tem um custo muito grande. Por isso, torna-se difícil de aceitar. Se aceitarmos o que ele diz, então temos que banir completamente muitos argumentos filosóficos que se baseiam em experimentos de pensamento, pois muitos deles nos exigem que pensemos em situações em que nossa condição epistêmica é alterada, seja em menor ou maior grau. Há vários exemplos - para além do já mencionado argumento da terra gêmea, de Putnam - como o argumento da máquina de experiências contra o hedonismo, o experimento do cérebro na cuba, o problema do trolley e assim por diante. Teríamos que recusar todos eles se formos recusar o ATGM com base na premissa de que nos fazem a exigência de nos colocarmos numa situação epistêmica diferente. Mas não parece necessário fazer isso, pois conseguimos alargar a nossa imaginação de forma suficientemente aceitável para a maioria dos casos. Portanto, não parece necessário recusar o ATGM apenas com base na diferença modal.

Essa inferência duvidosa que Sonderholm faz do exemplo de Kraemer se torna mais problemática quando consideramos a parte explanatória de seu ataque a H&T. Lembre que ele afirma que essa distância modal entre T e TG explica porque temos a intuição, supostamente errada, da univocidade semântica. Primeiro, notemos, sobre o exemplo de Kraemer, que o correto seria dizer que (a) temos intuições semânticas sobre o caso do ‘gene’, mas que (b) essas intuições são *equívocas* no sentido de que podem ser de que há genes no mundo hipotético e que não há genes no mundo hipotético, e não sabemos qual seria a opção mais difundida. Assim sendo, se Sonderholm usa o exemplo de Kraemer como evidência contra o ATGM, e no exemplo de Kraemer a intuição é equívoca, então, para o argumento por analogia funcionar, é preciso que a intuição sobre o ATGM também seja equívoca. Mas será que é? Parece que não.

H&T fornecem o seguinte argumento para recusarmos a parte explanatória do desafio de Sonderholm (2015, p. 369s). Para preservar o paralelo com o exemplo de Kraemer, suponha que em T os habitantes realmente considerem que o termo moral ‘correto’ seja causalmente

regulado pela propriedade natural N (consequencialista). Ou seja, em T não há controvérsia se é uma teoria de primeira ordem consequencialista ou deontológica que melhor capta as instâncias do certo e errado. Sigamos o experimento em todos os outros aspectos. Teríamos a intuição equívoca neste caso? Ou seja, estaríamos em uma condição em que não saberíamos qual intuição se teria tal como no exemplo de Kraemer? Parece que não, e o amplo compartilhamento e persistência da IUS reflete este fato. Portanto, a reflexão no ATGM não implica na equivocidade da intuição como a reflexão no experimento de Kraemer.

Isso coloca em problemas a parte explanatória do argumento de Sonderholm. Lembre que Sonderholm afirma que a diferença modal entre os dois mundos, além de tornar a intuição não confiável, explica porque temos a intuição da univocidade semântica. Segundo ele, o exemplo de Kraemer mostra isso. Assim, se a distância entre os dois mundos é o que explica a intuição, e no caso do exemplo de Kraemer a intuição é equívoca, no ATGM a intuição deveria ser equívoca também. Mas, como vimos, não é. As pessoas compartilham largamente a intuição de que há desacordo entre T e TG. Portanto, a distância entre os mundos não é um bom candidato para explicar o porquê da nossa intuição do desacordo.

Sonderholm ainda poderia argumentar que não é apenas a diferença modal que explica a intuição, mas que é a diferença modal *mais* o fato de que o uso de 'correto' no discurso comum não rastreia nenhuma propriedade unitária. Daí que as pessoas teriam dificuldade em reconhecer, no cenário do ATGM, que os termos morais são regulados por uma propriedade unitária em T e TG e que, por isso, teriam dificuldade em reconhecer que essas propriedades são diferentes nos dois mundos e isso favoreceria a IUS.

Mas, como apontam H&T (2015, p. 370), aceitar que, porque as pessoas aplicam o termo 'moral' a situações que não tem nada em comum no nível das propriedades superficiais elas não reconhecem que há algo que une todas as instâncias em que aplicam o termo, é problemático. Por que categorizamos, então, uma variedade tão grande de ações sob os domínios de 'certo' e 'errado'? Se não reconhecêssemos nenhum aspecto comum unitário entre ações boas e más, certas e erradas, isso nos impediria de fazer qualquer categorização. Portanto, parece que reconhecem pelo menos alguma unidade. É muito mais sensato dizer que, ao menos implicitamente, as pessoas reconhecem que correto, errado instanciam algum tipo de propriedade comum, mesmo que não tenhamos uma resposta em mãos sobre a ontologia de tal propriedade. Neste sentido, parece que essa segunda alternativa de Sonderholm explicar a IUS também implica em problemas.

Até aqui apontei uma série de problemas que a explicação de Sonderholm parece ter. Isso já nos dá razão para preferirmos a explicação da intuição que apela para as CSF do que a hipótese alternativa. Mas, ainda assim, quero dar um passo além. Sonderholm nos apresenta a hipótese da distância modal entre os mundos como sendo explanatória da intuição. Portanto, a posição dele implica, se há tal distância modal temos a IUS (os habitantes de T e TG discordam genuinamente). Para mostrar que sua hipótese não é suficientemente explanatória, e que a alternativa que apela para as CSF é melhor, vou construir dois casos. No primeiro, a hipótese explanatória de Sonderholm de que há desacordo está presente, mas não há desacordo. No segundo, a hipótese explanatória de Sonderholm de que há desacordo não está presente, mas há desacordo. E, para evidenciar a relevância explanatória das CSF, no cenário em que não há desacordo essas CSF não estão presentes e no cenário em que há desacordo elas estão presentes. A conclusão não pode ser outra: a hipótese explanatória que apela para as CSF é *melhor* do que a hipótese de Sonderholm.

Vamos ao primeiro cenário. Suponhamos um caso, muito similar ao construído por H&T, em que os habitantes de T adotam uma teoria moral de primeira ordem consequencialista como especificadora da propriedade N que regula o uso dos termos morais em T e os habitantes de TG adotam uma teoria moral de primeira ordem deontológica como especificadora da propriedade *N* que regula o uso dos termos morais em TG. Imaginemos que em uma das duas comunidades, seja em T ou em TG, as CSF não estão presentes no uso dos termos morais. Ou seja, embora os membros desta comunidade apliquem, por exemplo, 'correto' a pessoas, ações e instituições que *maximizam o agregado de felicidade*, eles não manifestam nenhuma das características que elencamos como sendo as CSF dos termos morais. Embora eles digam que '*x* é correto*"*, e façam tal predicação acertadamente de acordo com a teoria de primeira ordem em questão, eles não estão dispostos a agir de acordo, tais juízos não desempenha papel importante para tomar decisões sobre o que fazer, não tratam tais juízos como categóricos, não sentem culpa quando realizam uma ação cujo predicado associado é 'incorreto', etc.

Agora, suponhamos que os habitantes de T afirmam que '*x* é correto' e os habitantes de TG afirmam que '*x* não é correto'. Estão eles em desacordo moral genuíno? Antes de responder, acrescentemos a seguinte informação relevante. Suponhamos que a nossa condição epistêmica *real*, enquanto apreciadores do experimento em questão, é de que *não* há consenso definitivo sobre a teoria moral de primeira ordem que melhor capta a propriedade R-relacionada ao uso dos termos morais, seja entre os agentes morais comuns ou entre os especialistas em ética. Ou seja, suponhamos que há uma diferença modal entre a nossa situação epistêmica atual

e a situação requerida no experimento (que é de consenso sobre a propriedade R-relacionada ao uso dos termos morais). Com isso, preservamos a hipótese explanatória de Sonderholm.

Agora podemos perguntar, novamente: os habitantes de T e TG estão em desacordo moral genuíno quando uns afirmam que '*x* é correto' e outros que '*x* não é correto'? A resposta é, claramente, '*Não*'. E, se apreciamos corretamente a significância do *Argumento Invertido da Terra Gêmea Moral*, sabemos o motivo. Uma das duas comunidades não faz moralidade. Mesmo que apliquem termos que nós consideramos morais a pessoas, ações, instituições etc., eles não usam termos *morais*, pois as CSF não estão presentes. Uma das duas comunidades não é constituída por agentes morais, mas *schmorais*.

Mas o que tudo isso significa? Significa que temos um caso em que as características que Sonderholm atribui à causa da nossa IUS (intuição de que há desacordo moral genuíno entre T e TG) estão presentes e, mesmo assim, não temos tal intuição. E note, também, que as CSF, que estou reivindicando como a verdadeira causa explanatória da IUS, estão presentes. Portanto, parece que o que melhor explica a nossa intuição não é a hipótese de Sonderholm, mas as CSF.

A fim de tornar mais explícito ainda o meu ponto contra Sonderholm, vou construir um segundo cenário em que, mesmo com a hipótese explanatória de Sonderholm ausente, temos a IUS.

Suponhamos um caso em que os habitantes de T adotam uma teoria moral de primeira ordem consequencialista como especificadora da propriedade *N* que regula o uso dos termos morais em T e os habitantes de TG adotam uma teoria moral de primeira ordem deontológica como especificadora da propriedade N que regula o uso dos termos morais em TG. Mas agora as CSF fazem parte do uso dos termos morais de ambas as comunidades. Quer dizer, quando aplicam termos como 'correto' a ações, pessoas, instituições etc., os habitantes das duas comunidades aprovam ou recomendam, estão dispostos a agir de acordo, a considerar tais juízos como importantes para a tomada de decisões sobre o que fazer e assim por diante.

Além disso, vamos adiantar a nossa condição epistêmica real para um futuro no qual, depois de muita discussão e análise, se chegou a um consenso de que a melhor teoria moral de primeira ordem é consequencialista e capta a propriedade natural *N* que está R-relacionada ao nosso uso dos termos morais. Portanto, na nossa condição real, não há controvérsia sobre qual teoria moral de primeira ordem devemos adotar (consequencialista, deontológica ou algo diferente de tudo o que conhecemos).

Agora, perguntemos: se os habitantes de T afirmam que '*x* é correto' e os habitantes de TG afirmam que '*x* não é correto' estão eles em desacordo moral genuíno? A resposta é: sim, eles estão em uma disputa moral genuína. Note que, neste caso, eliminamos a hipótese explanatória de Sonderholm que supostamente explicaria a intuição de que há desacordo. E, se temos a IUS mesmo quando supomos que a nossa condição epistêmica real é de que não há consenso sobre qual teoria moral de primeira ordem é melhor, num cenário futuro possível em que há consenso, a IUS deve ser mais persistente ainda.

O que temos neste segundo cenário é o seguinte. Eliminamos a hipótese que Sonderholm argumenta ser a explicação da IUS, mas, mesmo assim, a intuição permanece. Curiosamente, as CSF estão presentes no uso dos termos morais dos habitantes de T e TG neste caso. Portanto, temos que concluir que o que realmente explica a IUS não é a distância modal entre as duas comunidades, mas as CSF dos termos morais.

Novamente, gostaria de ressaltar que pode ser que a argumentação de Sonderholm não seja totalmente falsa. Mas ela apresenta vários problemas. E, diante da explicação concorrente da intuição, temos mais razão para aceitarmos a concorrente e recusarmos a de Sonderholm. Então, o ataque de Sonderholm não invalida o ATGM. E, para concluir, lembremos da pressuposição metodológica que estamos seguindo. Argumentos que visam desmistificar a intuição sobre cenário hipotéticos, em muitos casos, obtêm sucesso. Mas quando esses argumentos apresentam uma série de problemas, e a intuição em questão é amplamente compartilhada e persistente, a explicação vindicativa ganha força. A hipótese explanatória de Sonderholm enfrenta vários obstáculos, como vimos. Neste caso, temos mais razão para recusá-la do que para adotá-la.

5. Observações finais: o desconforto do naturalista

Tenho argumentado que nenhuma das estratégias consideradas até aqui obtêm sucesso em enfraquecer a IUS. Mesmo diante dos desafios empregados por Levy, Viggiano e Sonderholm, não temos razão suficiente para concluir que P3 é falsa e que, por conseguinte, o defensor do ATGM está numa posição difícil. Pelo contrário, vimos que mesmo que membros de T e TG se encontrem num futuro determinado ou se estiverem numa posição epistêmica avançada, a intuição de que haveria desacordo moral genuíno permanece.

Para concluir, consideremos um último movimento que talvez o RMN poderia querer fazer. Vamos supor que Levy e Viggiano estivessem certos. Se isso fosse o caso, então não

haveria desacordo entre Terráqueos e Terráqueos Gêmeos, pois eles estariam predicando coisas diferentes com seus termos morais. 'Correto-*t*' para os membros de T e 'correto-*tg*' para os membros de TG. Isto é, haveriam fatos e propriedades morais, estes fatos e propriedades seriam naturais, mas não haveria uma propriedade natural *N* definitiva que seria a extensão dos predicados morais; essas propriedades naturais N seriam diferentes nos dois mundos, tal como esses filósofos aceitam. Esta manobra é sugerida especialmente por E. Gampel (1997, p. 152-154) e consiste, literalmente, em aceitar o *relativismo* entre os mundos. Mas, será que o RMN pode recorrer a este tipo de estratégia? Será que aceitar o relativismo entre T e TG e dizer que não há desacordo entre as duas comunidades é livrar-se do problema?

Acredito que isso seria um custo demasiado alto para o RMN. Ele teria que aceitar que se um membro de T dissesse '*x* é correto', um membro de TG poderia também dizer 'sim, *x* é correto, mas *x* não é correto*', enquanto que o mesmo habitante de T poderia replicar 'claro, concordo com você que *x* não é correto*, mas *x* é correto'. Isso é tudo o que resta ao RMN. Mas, parece que há uma questão a mais que deve ser resolvida aqui, caso contrário a discussão moral não avança. Além disso, como apontam H&T,

> [O relativismo] será bastante não atrativo para os naturalistas metaéticos como Boyd. Esses filósofos defendem o realismo moral, uma posição inflexivelmente não-relativista. E, em qualquer caso, o relativismo é a última opção defensiva contra o Argumento da Terra Gêmea Moral. Se o naturalismo ético for forçado em tal direção, então merece morrer (H&T, 1992, p. 70).

Quer dizer, seria completamente inesperado para o RMN recorrer a uma posição que sua própria teoria realista pretende expurgar. Isso colocaria o RMN numa posição bastante desconfortável. Portanto, ainda que quiséssemos aceitar o ponto de Levy e Viggiano, teríamos que nos comprometer com uma posição difícil.

CAÍTULO 5 – METASSEMÂNTICAS ALTERNATIVAS

1. Introdução

Quando H&T apresentaram o ATGM contra o RMN eles especificaram uma teoria naturalista em particular como alvo, o Naturalismo Semântico Causal (NSC) de Boyd. E até mesmo na literatura subsequente que discute o argumento há um comportamento padrão de apresentar o ATGM como um ataque à teoria de Boyd apenas. Mas o NSC de Boyd não é a única teoria naturalista disponível. Diante disso, pode-se pensar que, mesmo que o desafio de H&T obtenha sucesso em refutar o NSC de Boyd, há outras versões de RMN que *não* são vulneráveis ao ATGM. No entanto, H&T enfatizam que a Terra Gêmea Moral não é meramente um experimento mental que se aplica a uma forma específica de RMN, mas é uma ferramenta que pode ser generalizada e possui diversos alcances.

> […] A Terra Gêmea Moral é mais do que um experimento mental específico direcionado a uma tese semântica específica do NSC. É, para além disso, uma *fórmula* para experimentos mentais. Pois, para qualquer versão potencial de RMN de acordo com a qual (i) os termos morais possuem alguma relação R com certas propriedades naturais que, coletivamente, satisfazem alguma teoria moral normativa específica T, e (ii) os termos morais supostamente *referem* às propriedades naturais com as quais eles estão em tal relação R deve ser possível construir um cenário da Terra Gêmea Moral [...] (H&T, 1992, p. 167).

Por que, então, a teoria de Boyd fora proposta como alvo do argumento? Os próprios autores respondem: "nenhuma versão interessante de naturalismo semântico sintético, significativamente diferente do naturalismo semântico causal ao estilo de Boyd, mesmo que remota, atualmente está disponível" (H&T, 1992, p. 170).

No entanto, atualmente, temos várias teorias naturalistas alternativas. E, tal como o RMN dos *Realistas de Cornell* tinha como importante ponto positivo evitar o AQA de Moore, essas novas abordagens reivindicam como ponto decisivo em seu favor o fato de escaparem ao desafio semântico do ATGM. Esgotados os argumentos contra P2 e P3 do ATGM, essa parece ser a estratégia restante para o defensor do RMN. Tal linha de réplica consiste em sustentar que 'correto-*t*' e 'correto-*tg*' não expressam conteúdos semânticos diferentes, mas sem se comprometer unicamente com alguma forma de não-cognitivismo antirrealista, obviamente. Defender esse tipo de tese seria negar a verdade de P1 do ATGM.

Atualmente, há três principais desenvolvimentos que visam ao ataque de P1. São eles: (a) fornecer uma Semântica Moral Normativamente Enriquecida (SMNE); (b) apelar para uma teoria semântica bi-dimensional; (c) e recorrer à ideia de magnetismo referencial (*reference magnetism*). As propostas aqui são bastante sofisticadas e complexas, de modo que, para os limites do presente capítulo, terei que fazer algumas delimitações. Portanto, irei considerar a abordagem (a) e (b) apenas. Mais especificamente, irei explicitar a SMNE de David Brink (2001) e a teoria metaética híbrida de David Copp (2001). Tentarei argumentar, contra Brink, que a SMNE não representa grandes avanços para o RMN, já que ou tem custos indesejados ou é vulnerável a variantes do ATGM. Além disso, irei sustentar que o hibridismo metaético de Copp, embora tenha alguns custos, é preferível em comparação a SMNE de Brink e pode ser a melhor estratégia de escape para o ATGM que temos visto até aqui, *desde que se aceite a plausibilidade de teorias metaéticas híbridas*. No entanto, este é um ponto sobre o qual terei que permanecer neutro, dado o seu alcance e os limites desse capítulo.

Comecemos, então, com a SMNE de Brink.

2. Semântica Moral Normativamente Enriquecida

Embora uma das teses mais famosa de Putnam em *The Meaning of Meaning* seja de que a fixação da referência dos termos depende de aspectos externos aos estados mentais, é importante notar, também, a ênfase dada no papel das *intensões referenciais* dos falantes ao usar determinado predicado (PUTNAM, 1975, p. 225). A intenção referencial garante a continuidade da referência entre diferentes falantes. Uma das principais ideias da teoria da referência direta é que há um primeiro ato em que determinado termo é associado a determinado objeto (“batismo”) e o uso subsequente de tal termo por falantes diferentes garante que tal termo se refere ao mesmo objeto devido à participação na corrente histórico causal que conecta ao uso anterior (KRIPKE, 1980). O que permite que falantes do futuro usem o termo para captar o mesmo referente que seus antecessores, o que os conecta na mesma corrente histórico causal, é a intenção do falante em se referir ao mesmo objeto. Além disso, as intensões referenciais também explicam porque um dado termo capta uma característica do ambiente dos falantes e não outra. Considere o famoso exemplo de Keith Donnellan (1966). Obtenho sucesso em me referir ao sujeito A, e não ao sujeito B, com a descrição ‘o homem que está bebendo Martini’ mesmo que A esteja apenas segurando um copo de Martini com água dentro, e B esteja

realmente bebendo Martini, embora não visível de onde estou. O que garante isso é a minha intenção referencial.

Em *Realism, Naturalism, and Moral Semantics* (2001), Brink argumenta que se enfatizarmos o papel das intensões referenciais dos falantes na teoria da referência direta, obtemos um resultado diferente da teoria da regulação causal de Boyd. Resultado este que nos coloca em posição para responder ao desafio colocado pelo ATGM ao RMN. Como vimos anteriormente, dada a *Intuição da Univocidade Semântica* (IUS), parece haver um significado comum entre 'correto-*t*' e 'correto-*tg*'. O RMN, representado pela teoria metassemântica de Boyd, não está em posição de fornecer uma explicação para tal significado comum, de modo que ou rejeita a IUS, o que não parece ser uma opção atraente, ou aceita o relativismo moral, o que seria irônico já que sua teoria visa justamente evitar qualquer tipo de relativismo. Neste sentido, seria um grande avanço para o defensor do RMN se se pudesse fornecer uma explicação para a comunalidade semântica entre 'correto-*t*' e 'correto-*tg*', mas do ponto de vista naturalista, isto é, sem se comprometer com nenhum tipo de conteúdo expressivista sobre os predicados morais. Essa é a estratégia que Brink busca perseguir ao enfatizar o papel das intensões referenciais dos falantes. Consideremos sua teoria em maiores detalhes.

Há duas teses principais aqui:

(i) As intensões referenciais do falante determinam quais predicados de uma linguagem são ou não *morais*.
(ii) As intensões referenciais do falante fixam a propriedade que constitui o conteúdo semântico dos predicados morais.

A combinação de (1) e (2), sustenta Brink, "[...] identifica aquele significado ou referência comuns sobre o qual as duas comunidades [T e TG] tem um desacordo" (BRINK, 2001, p. 174-175). Tal conteúdo semântico compartilhado entre 'correto-*t*' e 'correto-*tg*' é revelado pela intenção referencial de se usar o termo no sentido *moral*. Assim, a questão principal é o que significa usar certos predicados como sendo predicados *morais*. Ou, por que deveríamos considerar os predicados dos habitantes de TG como sendo predicados *morais*?

Sua resposta requer uma especificação do que ele chama de *ponto de vista moral*. O ponto de vista moral deve ser específico o suficiente para se limitar àquilo que consideramos juízos *morais* apenas – e não todo e qualquer tipo de juízo avaliativo (como os juízos estéticos, por exemplo) – e geral o suficiente para que uma variedade de visões morais, mesmo as não

ortodoxas, possam satisfazer a descrição. Brink apela para uma distinção bastante comum entre *concepções* e *conceitos*. Em filosofia, estamos familiarizados com a distinção entre um determinado conceito abstrato e diferentes abordagens ou concepções de tal conceito. Um bom exemplo é a distinção entre o conceito geral de justiça distributiva e as diferentes concepções substantivas (utilitarista, libertária, igualitária) de tal conceito. Por que as concepções utilitarista, libertária e igualitária são todas concepções de um mesmo conceito? Isso requer uma caracterização abstrata do conceito que permita agrupar essas concepções como sendo concepções de um mesmo conceito (BRINK, 2001, p. 172).

Podemos aplicar o mesmo tipo de raciocínio em relação ao *ponto de vista moral* e às diferentes concepções substantivas de moralidade, sustenta Brink. Ele apela para a ideia contratualista de que "assumir o ponto de vista moral envolve acessar a própria conduta bem como a dos outros em termos de padrões que admitem *justificação interpessoal*" (BRINK, 2001, p. 173). De acordo com essa visão, consideramos ações, pessoas e instituições de acordo com padrões que os outros podem e devem aceitar. Note que esse conceito do *ponto de vista moral* admite muitas concepções distintas, tais como utilitarista, kantiana, egoísta etc.

Neste sentido, enunciar predicados *morais* significa usá-los com a intenção referencial de assumir o ponto de vista moral. E assumir o ponto de vista moral significa considerar as ações das pessoas, inclusive as próprias, seu caráter etc., de acordo com padrões de justificação interpessoal.

> De acordo com essa visão, devemos entender talvez todos os avaliadores morais, e certamente aqueles que introduziram os termos e categorias morais, como usando tais termos e categorias com a *intenção* de captar propriedades de pessoas, ações e instituições – seja lá quais forem essas propriedades – que desempenham um papel importante na justificação interpessoal do caráter dos indivíduos, suas ações e instituições. [...] Essa abordagem nos permite explicar as condições sob as quais é razoável interpretar os juízos de comunidades distintas como *juízos morais*. (BRINK, 2001, p. 174, itálico meu).

Assim, falantes que usam, por exemplo, 'correto' como um predicado genuinamente *moral* o usam com a intenção referencial de expressar aquela propriedade que desempenha um papel importante na justificação interpessoal das ações, pessoas, instituições. Por outro lado, se um grupo de falantes usa o mesmo predicado moral mas sem a intenção referencial de expressar tal propriedade, seu vocabulário não é genuinamente *moral*. Temos aqui a primeira tese importante de Brink: as intensões referenciais determinam quais predicados de uma linguagem são predicados *morais*.

A segunda tese se segue da primeira. Quando falantes de uma linguagem usam termos morais, eles têm a intenção referencial de fixar a propriedade que desempenha um papel importante na justificação interpessoal. Qual é a propriedade que desempenha tal papel? Esse é um problema sobre o qual o metaeticista não precisa se preocupar, pois irá depender de qual abordagem substantiva for aceita como padrão de justificação interpessoal. Mas suponha, por exemplo, que o padrão de conduta aceito como justificável interpessoalmente seja uma forma de consequencialismo de acordo com a qual devemos maximizar o agregado de felicidade. Se os falantes usam predicados como 'correto' com a intenção de captar a propriedade fixada pela justificação interpessoal, então 'correto' expressa a propriedade de *maximizar o agregado de felicidade*. Como afirma Brink, "nessa visão, a referência é fixada por uma intenção original em adotar o ponto de vista moral – ou seja, usar a linguagem moral para captar aquelas propriedades, seja lá quais forem, que tornam o objeto da consideração moral interpessoalmente justificável" (BRINK, 2001, p. 175). Assim, temos a segunda tese: as intensões referenciais do falante fixam a propriedade que constitui o conteúdo semântico dos predicados morais.

O ponto importante dessas duas teses de Brink é que, aparentemente, teríamos como responder ao desafio apresentado pelo ATGM. Tal resposta seria algo como:

> Lembre que o relativismo parecia ser o compromisso da teoria da referência direta na medida em que esta teoria era incapaz de identificar um *significado e referência comuns* sobre os quais os avaliadores da Terra e Terra Gêmea Moral mantinham diferentes crenças. Mas a nossa abordagem da intenção referencial compartilhada para se referir a pessoas, ações e instituições que são interpessoalmente justificáveis, em virtude das quais os juízos dos Terráqueos e dos Terráqueos Morais Gêmeos são ambos juízos morais, *identifica tal significado e referência comuns sobre os quais as duas comunidades mantem um desacordo na crença*. Seu desacordo é sobre quais características das pessoas, ações e instituições as fazem interpessoalmente justificáveis, com os Terráqueos sustentando visões consequencialistas e com os Terráqueos Morais Gêmeos mantendo visões deontológicas. (BRINK, 2001, p.174-175. Itálico meu).

Lembre da primeira tese de Brink. Se correto-*t* e correto-*tg* são predicados genuinamente morais, então falantes das duas comunidades necessariamente os usam com a mesma intenção referencial, isto é, de captar uma propriedade – seja qual for – que desempenha um papel importante na justificação interpessoal de ações pessoas e instituições. Essa comunalidade semântica independe das crenças que os falantes mantem sobre o que realmente está de acordo com padrões interpessoalmente justificáveis. Uma comunidade, T, por exemplo, pode sustentar que '*x* é correto-*t*' enquanto que outra comunidade, TG, pode sustentar que '*x* não é correto-*tg*'. Eles terão crenças diferentes sobre o que é interpessoalmente justificável, mas em última

instância sua discussão é *sobre o que é interpessoalmente justificável*. Portanto, há um significado comum em seus predicados morais que os coloca numa posição de univocidade semântica, sugere Brink. Isso, por sua vez, aniquilaria o ATGM, pois, em primeiro lugar, manteria os compromissos do RMN e, em segundo lugar, forneceria uma explicação para a IUS, isto é, por que faz sentido termos a intuição de que habitantes de T e TG discordam genuinamente, mas sem se comprometer com nenhuma tese expressivista sobre o conteúdo dos termos morais.

Brink argumenta que tal abordagem é claramente um avanço em relação ao tipo de teoria semântica desenvolvida por Boyd. Considere o seguinte. Suponha que habitantes de T sustentam que '*x* é correto-*t*' e habitantes de TG sustenta que '*x* não é correto-*tg*'. De acordo com a teoria da regulação causal de Boyd, a segunda afirmação não está em conflito com a primeira porque os falantes estão predicando propriedades diferentes a respeito de *x* – N e N1, digamos. Portanto, não há desacordo genuíno. Mas de acordo com a teoria das intensões referenciais de Brink, há desacordo genuíno, pois a afirmação dos habitantes de TG nega a afirmação dos habitantes de T porque ambos estão predicando a propriedade N (aquilo que é interpessoalmente justificável) de *x*, embora mantenham crenças diferentes sobre esta propriedade. É neste sentido que Brink afirma que

> Enquanto a abordagem da referência em termos da regulação causal deve permitir que a linguagem moral de falantes diferentes possa ser regulada por – e, portanto, refira a – propriedades diferentes, a abordagem da referência em termos das intenções referenciais não implica que há referências múltiplas quando o uso da linguagem é regulado por propriedades diferentes. Isso nos mostra um modo pelo qual uma interpretação da referência em termos da regulação causal não pode distinguir entre as crenças do falante e o seu objeto. (BRINK, 2001, p. 175).

E essa incapacidade de distinguir entre a referência e as crenças dos falantes é o calcanhar de Aquiles da teoria de Boyd. Diferentemente, a visão da referência em termos das intensões referenciais explica, argumenta Brink, por que T e TG estão em desacordo genuíno sem a necessidade de se aceitar nenhuma tese expressivista.

Podemos denominar o tipo de abordagem desenvolvido por Brink de *Semântica Moral Normativamente Enriquecida* (doravante, SMNE)[42]. Isso porque o que determina o conteúdo semântico dos predicados morais são fatos normativos sobre quais padrões de conduta os agentes *devem* aceitar. No caso de Brink, são fatos sobre o que é interpessoalmente justificável.

[42] Tal expressão é de M. Rubin (2015).

Ele não nos fornece uma teoria completa sobre a justificação interpessoal, mas afirma o seguinte:

> Nessa concepção do conceito de moralidade, o que é distintivo sobre o ponto de vista moral é que consideramos pessoas, ações e instituições de acordo com padrões que os outros podem e *devem* aceitar. (BRINK, 2001, p. 174, itálico meu).

Note a expressão 'devem aceitar'. Trata-se de um uso normativo de 'deve', a partir do qual devemos concluir que há algum tipo de força normativa sobre os agentes a respeito do que é interpessoalmente justificável. É neste sentido que devemos entender a ideia de uma semântica *normativamente enriquecida*.

2.1. Problemas para a SMNE

O alcance da SMNE, no entanto, apesar de representar algum avanço em relação a abordagem de Boyd por distinguir entre crença e referência, não está numa posição muito melhor para responder ao ATGM. Como afirmam H&T, o ATGM não é um experimento mental isolado, aplicável a apenas ao tipo de teoria semântica *à lá* Boyd, mas uma *fórmula* para experimentos mentais, de modo que, para qualquer abordagem naturalista sobre o conteúdo dos predicados morais podemos aplicar alguma versão da Terra Gêmea Moral. A proposta de Brink, como pretendo mostrar a partir de agora, implica em alguns custos indesejados e possui compromissos que, isoladamente, são vulneráveis a versões revisadas do ATGM.

2.1.1. O débito da SMNE

Há uma pressuposição fundamental sem a qual a proposta de Brink de uma SMNE não obtém sucesso. Note que Brink argumenta evitar o desafio semântico do ATGM identificando um suposto conteúdo semântico *similar* entre 'correto-*t*' e 'correto-*tg*'. Como vimos, ele sustenta que, embora habitantes de T reivindiquem a propriedade de *maximizar o agregado de felicidade* é o conteúdo real dos predicados morais e os habitantes de TG mantenham que a propriedade de *tratar os outros como fins em si mesmos* é tal conteúdo, eles convergem sobre o seguinte: ambas comunidades estão oferecendo formulações sobre aquilo que é interpessoalmente justificável. Isso supostamente preservaria a ideia de um desacordo genuíno, para Brink. No entanto, sua posição depende do seguinte: deve haver uma *mesma propriedade* que faz com que as ações sejam interpessoalmente justificáveis tanto para habitantes de T

quanto para habitantes de TG se se quiser preservar a univocidade entre ‘correto-*t*’ e ‘correto-*tg*’.

Por que a necessidade de tal padrão de justificação similar? Porque em sua ausência podemos aplicar uma versão do ATGM para aquilo que é interpessoalmente justificável. Considere a Terra Gêmea Moral Discursiva (TGMD).

> *Terra Gêmea Moral Discursiva (TGMD)*: os habitantes de TGMD acabaram adotando uma teoria de primeira ordem D* sobre o que é interpessoalmente justificável. D* estabelece que uma ação é interpessoalmente justificável se for aceita por uma comunidade discursiva regida por normas ideais de justificação. Por outro lado, os habitantes de T acabaram adotando uma teoria de primeira ordem C* sobre o que é interpessoalmente justificável. C* estabelece que uma ação é interpessoalmente justificável se maximiza o agregado de felicidade.

O raciocínio aqui é o mesmo que dos experimentos mentais anteriores. Fossem os habitantes de T e TGMD submetidos a uma discussão a respeito do status moral de uma ação *x* e uns mantivessem que ‘*x* é interpessoalmente justificável’ enquanto outros sustentassem que ‘*x* não é interpessoalmente justificável’, a SMNE implicaria que não há desacordo moral e, portanto, a univocidade semântica sobre ‘interpessoalmente justificável’ estaria perdida. A teoria de Brink, neste sentido, não representaria nenhum avanço na tentativa de assegurar a univocidade entre os predicados morais dos membros de T e TG.

Portanto, como afirma Rubin (RUBIN, 2015, p. 394), para que a SMNE de Brink assegure a univocidade semântica entre ‘correto-*t*’ e ‘correto-*tg*’ devemos assumir, embora não tenhamos razões objetivas para isso, que o padrão de conduta sobre o que é interpessoalmente justificável é o *mesmo* tanto para habitantes de T quanto para habitantes de TG, isto é, devemos pressupor que há um padrão *unitário* de justificação interpessoal para todos os agentes morais possíveis. Isso seria o débito da SMNE de Brink.

2.1.2. O Trilema de Rubin

Imaginemos, agora, que há tal padrão unitário sobre o que é interpessoalmente justificável para *todos os agentes morais possíveis* e que, por conseguinte, a teoria de Brink esteja livre desse primeiro problema. Teríamos razões suficientes para aceitar a SMNE como

uma boa solução contra o ATGM? Acredito que essa abordagem, novamente, não nos traz avanços significativos uma vez que é vulnerável ao que irei chamar de *Trilema de Rubin*. Como veremos, tal trilema reforça a afirmação de H&T de que o ATGM é uma fórmula para experimentos mentais que pode ser adaptado a diferentes teorias naturalistas.

De acordo com a abordagem desenvolvida por Brink, o conteúdo dos predicados morais é fixado por fatos normativos, isto é, fatos sobre quais padrões devem ser aceitos num cenário de justificação interpessoal. O núcleo do *Trilema de Rubin* repousa na seguinte tese: assim como fatos ou juízos *morais*, os fatos e juízos *normativos* também são passíveis dos mesmos problemas e indagações metaéticas (RUBIN, 2015, p. 398). Isso naturalmente pressiona do defensor da SMNE a fornecer uma teoria metaética sobre os fatos normativos ou, no mínimo, reconhecer tal necessidade. E, diante disso, ele terá que escolher entre três alternativas possíveis sendo que nenhuma delas lhe parece ser aceitável.

Quando uma teoria metaética assume que há fatos e propriedades morais (tese realista), a questão colocada logo em seguida é sobre o que são tais fatos e propriedades. Há, basicamente, duas estratégias mais gerais que podem ser desenvolvidas no intuito de explicar sua natureza: (i) o não-naturalismo moral e o (ii) naturalismo moral. Sendo assim, se a SMNE de Brink assume a existência de fatos normativos e tais entidades suscitam o mesmo tipo de indagação metaética que os fatos morais, seu defensor estará diante das mesmas alternativas acima. O *Trilema de Rubin* estabelece que nenhuma dessas alternativas é viável para a SMNE e que, portanto, tal abordagem deve ser rejeitada. Em resumo, o trilema consiste no seguinte. O defensor da SMNE não pode recorrer ao não-naturalismo noral, pois isso representaria a negação da própria tese que quer defender, a saber, o Realismo Moral *Naturalista*. Portanto, lhe restaria apenas a opção (ii). Mas assumir (ii) o compromete ou (a) com alguma versão naturalista de que os fatos normativos são crença-relacionados ou (b) com alguma versão naturalista de que os fatos normativos são desejo-relacionados. Mas a opção (a) é vulnerável ao próprio ATGM e a opção (b) representa o abandono de uma tese central para o RMN, a saber, a tese de que os fatos morais (ou normativos) são independentes da mente. A abordagem de Brink, portanto, não é a melhor estratégia para livrar o defensor do RMN do desafio semântico desenvolvido por H&T. Consideremos o trilema com maiores detalhes.

O não-naturalismo é a tese de que fatos e propriedades morais – nesse caso, fatos e propriedades normativos – não são naturais, mas representam entidades morais autônomas ou *sui generis*. Há várias abordagens positivas sobre o que eles são[43]. Mas o ponto aqui é o

[43] Veja, por exemplo, Moore (1903), Shafer-Landau (2005), Enoch (2011).

seguinte. O não-naturalismo é a negação do naturalismo e essa, obviamente, não é uma opção disponível para o defensor da SMNE uma vez que invalidaria o pressuposto naturalista que se quer preservar.

A única alternativa possível, então, é assumir algum tipo de naturalismo. Há dois caminhos aqui. Alguma forma de naturalismo expressivista ou alguma forma de naturalismo não-expressivista. A melhor abordagem sobre a segunda alternativa é o tipo de teoria desenvolvido pelos *Realistas de Cornell*, a saber, o tipo de RMN que temos considerado no decorrer desse trabalho. A hipótese aqui é que nossa melhor abordagem substantiva sobre o que devemos fazer rastreia o conteúdo dos fatos normativos. No entanto, não haveria avanço nenhum diante do ATGM, pois poderíamos reformulá-lo para qualquer versão substantiva pressuposta, como já vimos anteriormente. A SMNE estaria, assim, diante do mesmo problema que a teoria de Boyd.

Mas as alternativas expressivistas do naturalismo também não são uma opção disponível para o defensor da SMNE. Uma das teses centrais do RMN é a *Tese da Independência*, que sustenta que fatos e propriedades morais (normativos) são independentes de nossas mentes. Isso quer dizer que os padrões morais que determinam esses fatos não são verdadeiros devido a convenções, deliberações ou crenças de agentes atuais ou hipotéticos[44]. Juízos morais (ou normativos) são crenças que descrevem a realidade e são verdadeiros ou falsos. Note que assumir algum tipo de expressivismo naturalista implica na negação da *Tese da Independência*, uma vez que aquilo que é correto, incorreto ou justo depende das *atitudes* dos agentes ou observadores morais.

Como podemos notar, mesmo que Brink nos forneça um padrão *unitário* de justificação interpessoal e dê conta do débito de sua teoria, ainda assim, terá que enfrentar o *Trilema de Rubin*. Neste sentido, a ideia de uma SMNE não parece muito promissor. Agora, à fim de reforçar os problemas apontados como razão para recusarmos a SMNE, irei apresentar e responder a três objeções possíveis ao *Trilema de Rubin*.

2.1.3. Possíveis objeções e réplicas

O defensor do tipo de abordagem desenvolvido por Brink poderia tentar evitar aos problemas acima mencionados escolhendo alguma das seguintes alternativas de réplica disponíveis.

[44] Essa formulação da *Tese da Independência* é de Russ Shafer-Landau (2003, p. 15).

(a) *Independência dos fatos normativos*. Primeiro, tentar objetar àquilo que é a pressuposição fundamental do *Trilema de Rubin*, isto é, argumentando que sua proposta de uma SMNE independe de uma teoria metaética sobre o conteúdo dos fatos normativos. Neste caso, ele não seria pressionado a escolher entre não-naturalismo e naturalismo a respeito dos fatos normativos e, por conseguinte, estaria livre dos problemas apontados por quem optasse por qualquer uma dessas alternativas.

(b) *Hibridismo metaético*. O defensor da SMNE poderia tentar apelar para algum tipo de teoria híbrida sobre o conteúdo dos fatos normativos que combinasse aspectos crença-relacionados com aspectos desejo-relacionados. (Irei considerar com maiores detalhes o que são as abordagens híbridas e como podem ser desenvolvidas na próxima seção. Por enquanto, basta assumirmos que uma teoria híbrida busca fazer algum tipo de compatibilização entre elementos que tradicionalmente são considerados irreconciliáveis, tais como desejos ou atitudes conativas, por um lado, e crenças ou atitudes cognitivas, por outro). Tal abordagem híbrida supostamente preservaria univocidade entre os fatos normativos sobre o que é interpessoalmente justificável (dado o elemento conativo) e, portanto, evitaria formulações alternativas do ATGM e, além disso, preservaria a *Tese da Independência*, honrando, assim, seus compromissos com o realismo.

(c) *SMNE de segunda ordem*. Em terceiro lugar, se poderia tentar evitar o que considerei como sendo o débito da SMNE bem como o *Trilema de Rubin* apelando para o mesmo tipo de SMNE para explicar o conteúdo dos fatos normativos.

Não considero que essas estratégias sejam vias de escape promissoras em favor da SMNE. Em primeiro lugar, manter a independência dos fatos normativos não parece ser uma opção plausível. Como vimos, de acordo com a teoria de Brink fatos normativos sobre o que é interpessoalmente justificável fixam o conteúdo dos predicados morais. Neste sentido, imagine que o niilismo normativo ou o expressivismo normativo sejam verdadeiros. Dado que fatos normativos fixam o conteúdo dos termos morais, então ou os predicados morais não expressariam propriedade nenhuma (niilismo) ou os predicados morais teriam a função semântica de expressar estados mentais conativos (expressivismo). Nos dois casos, a SMNE implicaria a falsidade do RMN, já que esta teoria é incompatível com o niilismo e com o

expressivismo. Portanto, como afirma Rubin, não é uma questão de indiferença para o defensor do RMN qual seria a teoria metaética mais adequada sobre os fatos normativos (RUBIN, 2015, p. 398-399) e o pressuposto do trilema é justificado.

O hibridismo metaético também não parece ser uma opção atrativa. Seu proponente teria que fornecer uma explicação de porquê assumir o hibridismo sobre fatos normativos, mas não sobre fatos morais. Aparentemente, juízos normativos e morais, apesar de terem algumas diferenças superficiais (como, por exemplo, os juízos normativos possuem um escopo mais abrangente do que os juízos morais), compartilham características relevantes. Isso representaria uma dificuldade para a SMNE. E se, por outro lado, mostrasse que o hibridismo é verdadeiro para os fatos normativos, na ausência de um critério que colocasse fatos morais e fatos normativos em categorias claramente distintas, teria mostrado que o hibridismo é verdadeiro, também, para os fatos morais, caso em que acabaria anulando a necessidade de uma SMNE para desviar do ATGM uma vez que, como veremos adiante, as teorias híbridas estão numa boa posição para responder tal desafio semântico.

Finalmente, apelar para uma SMNE de segunda ordem gera um tipo de circularidade, como apontou o próprio Rubin (2015). Se estaria tentando explicar o conteúdo dos predicados morais apelando para uma SMNE. Mas a SMNE pressupõe a existência de fatos normativos e para explicar a natureza de tais fatos normativos se estaria apelando novamente para uma SMNE, que também se comprometeria com fatos normativos de segunda ordem. E assim sucessivamente. Portanto, essa também não parece ser uma opção disponível para o defensor da teoria desenvolvida por Brink.

Sendo assim, acredito que temos boas razões para acreditar que a proposta de uma SMNE como via de escape do ATGM não é promissora. Como vimos, há um débito inerente a tal abordagem que, se não for solucionado, torna a teoria do mesmo modo vulnerável ao ATGM. Para além disso, uma vez que o defensor da SMNE nos deve explicações sobre a natureza dos fatos normativos, fica sujeito ao *Trilema de Rubin*, diante do qual suas alternativas de resposta também parecem ser bastante limitadas.

3. Hibridismo Metaético

A metaética é tradicionalmente marcada por uma oposição aparentemente irreconciliável. Por um lado, temos um grupo de teorias sob a categoria de *realismo moral*. A marca definidora de tais teorias é que o pensamento e a linguagem moral dizem respeito a

aspectos *crença-relacionados*. Por outro lado, temos um grupo de teorias sob a categoria de *antirrealismo moral*. Costuma-se dizer que as teorias antirrealistas tratam o pensamento e a linguagem moral como *desejo-relacionados*. Parte significativa das razões a favor do realismo moral é constituída pela apresentação de uma série de problemas a que o antirrealismo moral suspostamente é vulnerável. O mesmo se aplica em relação à defesa do antirrealismo moral e a discussão parece definitivamente implacável. Mas, e se fosse possível combinar alguma forma de realismo moral com uma forma de antirrealismo moral? Certamente seria uma opção muito atrativa, pois, aparentemente, se poderia unir o melhor dos dois mundos possíveis já que, à primeira vista, tal teoria não seria vulnerável nem aos problemas do realismo e nem do antirrealismo.

Essa é a promessa de uma tendência mais recente da metaética comumente denominada de *teorias híbridas* ou *hibridismo*. A ideia é que o discurso e o pensamento moral combinam aspectos crença-relacionados e aspectos desejo-relacionados. Uma das teorias mais influentes desse tipo de abordagem híbrida é o *Realismo-Expressivista* (doravante, RE) de David Copp (2001). O RE

> […] sustenta que nossas crenças e juízos morais representam estados de coisas morais e podem ser corretos ou incorretos em relação a esses estados de coisas, o que é a tese central realista, mas também sustenta que, ao fazer tais asserções morais, expressamos certas atitudes ou estados motivacionais caracteristicamente conativos, o que é uma visão positiva central do expressivismo. (COPP, 2001, p.1).

Se tal combinação é de fato possível, então, além do defensor do RE poder neutralizar uma série de objeções tanto ao realismo quanto ao expressivismo antirrealista, ele supostamente estará em boa posição para recusar o ATGM. Isso porque, à princípio, ele pode manter a tese característica do RMN de que, por expressarem crenças, o conteúdo semântico dos juízos morais dos habitantes de T e TG é diferente (‘correto-*t*’ e ‘correto-*tg*’), mas sem precisar abrir mão da intuição amplamente compartilhada de que tais indivíduos estão em desacordo moral genuíno, uma vez que suas atitudes conativas seriam incompatíveis.

Sendo assim, nessa seção, pretendo fazer duas coisas: (i) apresentar, brevemente, o RE de Copp como resposta ao ATGM e (ii) considerar algumas objeções especificamente ao modo como tal teoria lida com o desafio semântico sugerindo que o RE, apesar de ter alguns custos, é a melhor estratégia que o defensor do RMN possui para se desviar do ATGM, desde que se aceite a plausibilidade de teorias metaéticas híbridas. Tal resultado promissor, no entanto, não vem sem custos. O RMN teria que aperfeiçoar sua posição original, ou até mesmo se distanciar

dela, já que seria forçado a migrar para uma forma de hibridismo metaético, o que é significativamente diferente de um Realismo de Cornell, por exemplo.

3.1. O Realismo Expressivista

Por ser uma forma de realismo combinada com uma forma de expressivismo, tal teoria tem, obviamente, um lado realista e um lado expressivista que devem ser explicados. O aspecto realista do RE é a combinação do que Copp chama de abordagem baseada-no-padrão (*standard-based account*) e teoria centrada-na-sociedade (*social-centered theory*). A abordagem baseada-no-padrão fornece as condições de verdade das proposições morais. Grosso modo, a ideia é que um juízo moral expressa uma proposição que é verdadeira apenas se há um padrão autoritativo que atesta ou justifica sua verdade. Tal ideia pode ser combinada com várias abordagens sobre o que constitui um padrão autoritativo, mas a opção defendida por Copp é a teoria centrada-na-sociedade, que diz:

> Minha visão sobre isso subscreve-se a um tipo de realismo moral naturalista. Sugiro que um padrão moral é justificado de forma relevante se, grosso modo, a sua ocorrência no código social da sociedade relevante iria melhor contribuir para a satisfação das necessidades – incluindo suas necessidades para a continuidade física, harmonia interna e interação cooperativa e relações pacíficas e cooperativas com seus vizinhos. (COPP, 2001, p. 28).

Assim, juízos morais são verdadeiros ou falsos em relação ao conteúdo do código moral ideal estabelecido pela teoria centrada-na-sociedade. A afirmação de que 'o assassinado é incorreto' é verdadeira se a propriedade da *incorreção* instancia a relação de ser proibida pelo código moral ideal da sociedade.

Em relação ao aspecto expressivista do RE, é importante esclarecer o seguinte. Tradicionalmente, compreende-se o expressivismo como uma forma de não-cognitivismo que envolve duas teses constitutivas, a primeira negativa e a segunda positiva: *não-factualismo semântico*: sentenças morais não são verdadeiras ou falsas; e *não-cognitivismo psicológico*: juízos morais expressam estados mentais conativos. Como afirma Copp, o RE

> […] aceita a tese positiva central … [O RE] concorda que, para todo pensamento moral básico M, há um estado conativo ou motivacional C-M, um estado, de alguma forma, similar a um desejo, tal como quando uma pessoa faz um juízo de que M "expressa" o estado C-M. [O RE] sustenta que quando uma pessoa faz um juízo moral M ela expressa a crença moral M e um estado C-M correspondente. C-M" (COPP, 2001, p. 9).

Há várias teorias sobre como melhor desenvolver a tese positiva do expressivismo[45]. Copp, especificamente, adota o *expressivismo de normas* de Allan Gibbard (2003). Assim, o estado mental conativo expresso pelo juízo moral M é o de *endosso* de um sistema de normas que regula M. E, dada a sua versão realista do RE, podemos dizer que se trata da aceitação de um padrão constituído pelo código moral ideal. Desse modo, o RE é a tese de que a asserção de um juízo moral envolve: (i) a expressão de uma crença que possui valor de verdade (estabelecida pela abordagem baseada-no-padrão) e (ii) a expressão de um estado mental conativo (de aceitação de um *padrão* de normas).

No entanto, a parte realmente importante da teoria de Copp é sobre a *relação linguística* entre a asserção de um juízo moral e a expressão de um estado mental conativo. Em que sentido um juízo moral como 'o assassinato é errado', *ao mesmo tempo*, descreve um estado de coisas e, além disso, expressa um tipo de atitude por parte do falante, já que, tradicionalmente, esses dois domínios parecem estar em oposição?

Para explicar tal relação linguística, Copp apela para a noção fregeana de coloração (*coloring*). A coloração é o resultado de convenções linguísticas que desempenham papel importante na determinação o uso de um termo. Tal característica é facilmente vislumbrada em termos pejorativos. Considere o seguinte exemplo.

(1) Este cão uivou a noite toda.
(2) Este vira-lata uivou a noite toda.

Ambas sentenças possuem o mesmo valor de verdade (este cão uivou a noite toda), mas implicam[46] significados diferentes. Em determinados contextos (2) pode expressar desprezo em relação ao cão. Por isso, a palavra 'vira-lata' deve ser evitada para comunicar a proposição de que o animal uivou a noite toda caso o falante não sinta nenhum desprezo pelo cão, já que, dadas as convenções linguísticas sobre o uso de 'vira-lata', o falante pode levar o ouvinte a crer que de fato mantém tal sentimento de repulsa. Esse elemento de interjeição que é parte do significado de 'vira-lata' é o que se denomina de coloração. Há muitos exemplos de coloração no discurso ordinário, mas o ponto importante para Copp aqui é o seguinte: "há uma

[45] Algumas alternativas são os clássicos Ayer (1936) Stevenson (1937) e Hare (1952) e versões mais atuais incluem Blackburn (1984) e Gibbard (2003).
[46] 'Implicar' aqui não diz respeito à relação lógica, mas à propriedade de uma asserção possuir, implicitamente, um conteúdo semântico.

diferença no que os falantes de cada sentença irão implicar sobre seus *estados mentais*, pois há uma diferença nas convenções linguísticas governando os usos dos termos" (COPP, 2001, p. 16, itálico meu). Nesse caso, (1) apenas expressa a crença de que o cão uivou a noite toda, mas (2), para além de expressar tal crença, expressa outro tipo de estado mental, a saber, o sentimento de desprezo do falante pelo cão. Assim, afirma Copp, "temos um exemplo de uma asserção que expressa plausivelmente uma crença e um estado conativo" (COPP, 2001, p. 16)[47].

Copp utiliza a noção de Estrutura Proposicional Múltipla, de Stephen Neale (1999), para melhor organizar as convenções linguísticas, incluindo a coloração, que permeiam o nosso discurso. De acordo com tal estrutura, há várias camadas de proposições sendo implicadas num determinado discurso. Por exemplo, se S expressa a sentença (2), dada a Estrutura Proposicional Múltipla de Neale, S estaria comunicando a seguinte sequência de proposições *Primária*: que o cão uivou a noite toda; *secundária*: que o falante tem repulsa pelo cão. O valor de verdade do que o falante diz depende apenas de se o cão uivou a noite toda, mas seria inapropriado ao falante usar (2) a menos que ele de fato sinta repulsa pelo cão. Paul Grice (1989) usa a expressão 'implicatura convencional' para explicar os mesmos fenômenos da coloração. Se apropriando da nomenclatura de Grice e da Estrutura Proposicional Múltipla de Neale, Copp usa a expressão 'implicação convencional' para se referir a asserções que envolvem coloração. Assim, se S enuncia (2) está convencionalmente implicando, para além da *crença* de que o cão uivou a noite toda, o *estado mental conativo* de que S sente desprezo pelo cão.

A esta altura já deve estar clara a estratégia de Copp. Ele sustenta que que os predicados morais possuem coloração e, por isso, para além do conteúdo proposicional, expressam estados mentais conativos. Mas que tipo de razão Copp nos oferece a favor da tese de que os termos morais possuem coloração? O plano aqui é o seguinte. Primeiro, ele formula quatro testes para determinar se um dado termo possui coloração e, segundo, argumenta que os termos morais passam em tais testes e, por conseguinte, também possuem coloração. Os quatro testes são os seguintes.

Em primeiro lugar, temos o que Copp denomina de *teste da verdade*: "se o uso de um termo T numa dada sentença sugere ou implica que *p*, e se isso é o caso devido à coloração de T, então a crença expressa por uma pessoa que assere a sentença pode ser verdadeira mesmo que *p* seja falso" (COPP, 2021, p. 17). Por exemplo, se S assere (2), sua crença de que o vira-lata uivou a noite toda pode ser verdadeira mesmo que S não sinta nenhum desprezo pelo cão. Em outras palavras, se um termo possui coloração, tal significado convencional não tem

[47] Para mais exemplos de termos que possuem coloração, veja COPP, 2001, p. 16 e 22.

implicações para a condição de verdade da sentença. Em segundo lugar, temos o *teste do mal-uso*, que estabelece que: "seria *inapropriado* usar um termo com coloração num contexto em que tal uso implica que *p* quando o falante sabe ou acredita que *p* não é o caso, mesmo que fazer isso não resulta na falsidade daquilo que o falante diz" (COPP, 2021, p. 17). Neste sentido, 'vira-lata', por exemplo, possui coloração se seria um mal-uso asserir (2) caso o falante não sinta desprezo pelo cão, mesmo que isso não afete a condição de verdade de (2), pois o falante pode estar levando o ouvinte a formar uma crença enganosa. Em terceiro lugar, temos o que Copp chama de *cancelabilidade* (*cancelability*) (COPP, 2021, p. 18-19), que afirma que um termo possui coloração se tal implicatura convencional pode ser *cancelada*. Exemplos incluem contextos em que um termo que possui coloração é empregado em sentenças complexas como no condicional 'se o seu cão é um vira-lata, então você deve vende-lo" ou contextos em que alguém que está aprendendo o português assere (2), mas não porque sinta desprezo em relação ao cão, mas por conhecer apenas 'vira-lata' para denotar os caninos. Em tais contextos, a coloração de 'vira-lata' fora cancelada. Por último, temos o *teste da separação* (*detachability*) (COPP, 2021, p. 19) que estabelece que um dado termo possui coloração se este significado convencional pode ser separado do conteúdo expresso pela crença. No nosso exemplo original, se poderia substituir 'vira-lata' por 'canino' ou 'cachorro' e expressar a crença de que o cão uivou a noite toda. Assim, de acordo com Copp, se um termo possui coloração, então ele deve passar nesses quatro testes.

O passo importante agora é sustentar que os *termos morais* possuem coloração. Copp faz isso na seguinte passagem:

> Para testar se o termo "moralmente errado" possui coloração, podemos aplicar os quatro testes para a coloração. [...] Para aplicar o teste da verdade, considere um amoralista: alguém que não tem nenhuma posição moral e, portanto, não se subscreve a nenhuma norma moral. Suponha que ele diga que "xingar alguém é moralmente errado". Mesmo que ele não se subscreva a uma norma que proíba o xingamento, não concluiríamos, com base nisso, que a proposição que ele asseriu é falsa. Em segundo lugar, o teste da separação (*detachability*). Me parece que as aspas podem ser usadas para descolorir termos que são normalmente coloridos. Posso dizer, por exemplo, que tal e tal comportamento seria "pouco elegante" indicando através de ênfase ou gesto que coloco tal termo entre aspas. Alternadamente, posso falar em "O xingamento é 'moralmente errado'", colocando "moralmente errado" entre aspas ou alguém poderia dizer que "O xingamento é moralmente errado, como dizem"; qualquer um dos métodos teria separado a implicação de que tenho algum tipo de posição contra o xingamento. Em terceiro lugar, considere o teste da cancelabilidade. Um amoralista poderia cancelar explicitamente a implicação de que ele subscreve a um padrão que proíbe o xingamento ao dizer "Condordo que o xingamento é moralmente errado, mas certamente não tenho nenhuma inclinação para evitar o xingamento". Finalmente, temos o teste do mal-uso. Me parece que na maioria dos contextos o amoralista teria usado mal o termo "moralmente errado" ao dizer "O xingamento é moralmente errado", já que ele não subscreve a nenhuma proibição ao xingamento. (COPP, 2001, p. 35-36).

Se isso é o caso, então sentenças morais, além de asserirem proposições que podem ser verdadeiras ou falsas, expressam um estado mental conativo. Aplicando a Estrutura Proposicional Múltipla, podemos dizer que quando alguém forma um juízo moral há duas camadas proposicionais envolvidas: *primária*: a crença de que M; *secundária*: o estado mental conativo C-M.

Ora, se tal abordagem híbrida realmente obtêm sucesso, parece que temos um avanço significativo não apenas no cenário geral das teorias metaéticas, mas especificamente em relação ao tipo de resposta que o defensor do RMN pode fornecer para o desafio semântico proposto pelo ATGM. No nível mais geral, o RE, argumentativamente, pode reivindicar para si tanto as vantagens o realismo quanto do expressivismo moral e, ao mesmo tempo, evitar os pontos fracos das duas abordagens. Mas, consideremos o que é nosso principal foco aqui, a saber, o ATGM.

Lembre que, por um lado, 'correto-*t*' e 'correto-*tg*' expressam propriedades distintas e, por conseguinte, possuem referências distintas, considerando a explicação fornecida pelo RMN. Isso implica que não há univocidade semântica entre tais predicados morais, uma vez que num cenário de putativa controvérsia entre habitantes de T e TG não haveria desacordo moral genuíno. Por outro lado, a intuição de que há desacordo moral genuíno e que, portanto, há univocidade semântica, parece ser não-negociável. Portanto, o RMN se encontra em sérios problemas. No que a dificuldade para o RMN é a não compatibilidade entre seus pressupostos *crença*-relacionados e a *Intuição da Univocidade Semântica* (IUS), já que esta aponta para aspectos *desejo*-relacionados. Mas, com a adoção do hibridismo na forma do RE, o RMN está livre desse problema. Ou seja, o defensor do RMN consegue preservar, agora, a ideia de que os predicados morais dos habitantes de T e TG possuem conteúdos semânticos diferentes (aspecto *crença*-relacionado) mas sem abrir mão da IUS (aspecto *desejo*-relacionado), já que parte do conteúdo de tais predicados é conativo.

Temos, finalmente, a melhor resposta possível para o ATGM e o RMN estaria, agora, livre do desafio semântico? Como veremos, há alguns custos que o RE possui, especificamente em relação ao ATGM, e, além disso, a aceitação do hibridismo representa uma grande mudança teórica para o RMN. Em outras palavras, uma resposta ao ATGM *à lá* teoria híbrida não livra o *Realismo de Cornell*, por exemplo, desse problema semântico e o defensor do RMN teria que migrar para uma posição que, tradicionalmente, não é ortodoxa nas formas de realismo moral. No entanto, se compararmos o RE com outras respostas ao ATGM, como a SMNE de Brink ou

as abordagens vistas nos capítulos anteriores, o RE representa superioridade teórica e é capaz de fornecer uma resposta plausível para o desafio semântico em questão.

3.2. Os custos do RE

Em *The Promise and Perils of Hybrid Moral Semantics for Naturalistic Moral Realism*, M. Rubin rejeita o RE como saída para o ATGM argumentando que, uma vez que o realista concede que há atitudes conativas de subscrição a normas morais, ele enfraquece o papel desempenhado pelos fatos e proposições morais pelo menos a respeito de três características específicas: desacordos morais, investigação moral e atribuição de pensamento moral e isso seria um custo demasiado alto para o RE. Rubin aponta para problemas centrais que uma proposta híbrida que tem por objetivo evitar o ATGM teria que enfrentar. Por isso, é importante que consideremos tais problemas.

3.2.1. Desacordo e investigação moral

Realismo e expressivismo moral fornecem explicações bem diferentes para o fenômeno do desacordo e investigação moral. Consideremos, em primeiro lugar, apenas a questão do desacordo. Comecemos com o realismo. Há, obviamente, diversas teorias realistas e elas podem divergir amplamente sobre características específicas da moralidade, como sobre o desacordo. Mas há uma base mínima de convergência entre tais teorias a que podemos denominar de *realismo mínimo*. Mark van Roojem (2015) fornece se seguinte caracterização do que constitui o realismo mínimo sobre um predicado moral 'P':

1. '**P**' representa uma propriedade (por conveniência, irei usar ***P*** para representar esta propriedade).
2. A sentença '**X** é **P**' representa '**X**' como tendo a propriedade ***P***.
3. O conteúdo de '**X** é **P**' é a proposição de que **X** é ***P***.
4. A asserção '**X** é **P**' expressa a crença de que **X** é ***P***.
5. A crença de que **X** é ***P*** representa um estado de coisas do mesmo modo que a sentença de que '**X** é **P**'; ambos representam X como tendo a propriedade ***P*** e ambos representam a proposição de que **X** é ***P*** como sendo verdadeira.
6. A crença de que **X** é ***P***, a sentença '**X** é **P**' e a proposição de que **X** é ***P*** são verdadeiras quando **X** de fato tem a propriedade ***P***.
7. Uma sentença da forma '**B** acredita que **X** é **P**' atribui a atitude de crença de que **X** é ***P*** a **B**.
8. Pelo menos algumas sentenças indicativas da forma '**X** é **P**' de fato representam o mundo corretamente e essas sentenças são verdadeiras em virtude de representar as coisas corretamente. (VAN ROOJEN, 2015, p. 13).

Suponha, agora, que S sustente que '*x* é errado' e S_1 defenda que '*x* não é errado'. Temos um caso de desacordo e o realista mínimo tem uma forma clara de explicar o que está acontecendo aqui. S está dizendo que *x* tem determinada propriedade, digamos *P*, enquanto S_1 está dizendo que *x* não tem tal propriedade *P*. Obviamente, *x* não pode ter e não ter uma propriedade ao mesmo tempo, de modo que o conteúdo das afirmações, isto é, as proposições '*x* é errado' e '*x* não é errado', não podem ser ambas verdadeiras. Por isso, as crenças expressas pelas afirmações de S e S_1 são incompatíveis e tais indivíduos estão em desacordo. Dadas as afirmações 4–7 de Van Roojen, podemos notar que a explicação do realista mínimo para o desacordo é fundamentalmente sobre *crenças*. Um falante crê no conteúdo de uma proposição enquanto outro crê no conteúdo da negação de tal proposição.

Por outro lado, um expressivista fornece outro tipo de explicação para o desacordo moral. Expressivistas negam que predicados morais expressam propriedades. Por conseguinte, uma sentença moral do tipo '**X** é **P**' não representa **X** como tendo ***P*** e, portanto, não pode ser verdadeira ou falsa. Além disso, a sentença '**X** é **P**' não expressa uma crença, mas um estado mental conativo. Aqui diferentes teorias expressivistas fornecem explicações diferentes sobre o que é tal estado mental conativo. Charles Stevenson (1944), por exemplo, defende que se trata de *atitudes de aprovação* ou *desaprovação*. Então, quando S sustenta que '*x* é errado' e S_1 que '*x* não é errado' não há um conflito entre duas proposições para o expressivista, mas um *conflito de atitudes*. S desaprova a ação *x* enquanto S_1 aprova a ação *x*. Desacordos morais são, portanto, conflitos de atitudes conativas.

Consideremos, agora, como o defensor do RE deve lidar com a ocorrência do desacordo moral entre habitantes de T e TG. Claramente há desacordo entre os falantes e o proponente da teoria híbrida quer fornecer uma explicação sem descartar o elemento realista de sua teoria. No entanto, quando pensamos especificamente no cenário do ATGM devemos aceitar que o aspecto realista do RE não possui eficácia explanatória sobre o desacordo. Pois, se as sentenças morais expressam crenças, a crença dos membros de T de que 'x é correto-t' e dos membros de TG de que 'x não é correto-tg' não são incompatíveis e ambas podem ser verdadeiras ao mesmo tempo. Isso porque os falantes atribuem propriedades diferentes a x, a saber, a propriedade de maximizar o agregado de felicidade e de tratar os outros como fins em si mesmos. Portanto, o aspecto realista do RE não dá conta do desacordo. Neste sentido, defensor do RE precisa apelar para o elemento puramente expressivista. Com isso, ele

certamente acomoda o fato do desacordo entre T e TG: embora as crenças dos membros das duas comunidades não sejam incompatíveis, há um *conflito de atitudes conativas*.

Portanto, o RE é capaz de dar conta do desacordo moral do ATGM, mas é forçado a fazer uma concessão importante: no nível mais profundo, desacordo moral é desacordo de atitudes conativas e não de crenças.

Passemos agora à questão da investigação moral. Aqui, o mesmo raciocínio se aplica. Realistas e expressivistas tem abordagens significativamente diferentes a respeito de tal característica da moralidade. Considere, por exemplo, a resposta de R. Boyd. Ao ser confrontado com a questão sobre o que desempenha, na moralidade, o papel que a observação desempenha nas ciências naturais, ele responde o seguinte:

> Proponho a resposta: "Observação". De acordo com a concepção do consequencialismo homeostático sobre a moralidade (na verdade, de acordo com qualquer concepção naturalística) bondade é uma propriedade ordinária natural e seria estranho se a observação não desempenhasse o mesmo papel no estudo desta propriedade como ela desempenha no estudo de todas as outras. De acordo com a concepção do consequencialismo homeostático, bondade é uma propriedade muitos similar às outras propriedades estudadas pelos psicólogos, historiadores e cientistas sociais e, *na investigação moral*, as observações irão desempenhar o mesmo papel que desempenham nos tipos de investigação empírica sobre as pessoas. (BOYD, 1989, p. 332, itálico meu).

Essa é a característica central do realismo acerca da investigação moral. Para o realista, não há construção de fatos morais, mas a investigação moral é guiada à *descoberta* de fatos e propriedades morais.

O antirrealismo expressivista, por outro lado, possui uma abordagem diferente. Considere a formulação de James Lenman (2007).

> Investigação moral é, em larga medida, a tentativa de membros de uma comunidade – ou do que pretende ser uma – chegarem, por co-deliberação, ao acordo sobre o que pode ser um conjunto aceitável de padrões morais para a conduta de tal comunidade. É uma tentativa de determinar quais normas morais podemos concordar em endossar como uma base para governar nossas vidas em comunidade, mas onde por 'determinar' não quero dizer descobrir tanto quanto acordar. Para colocar de uma forma um pouco provocativa, a investigação moral é política. (LENMAN, 2007, p. 75-6).

Para o expressivista, a investigação moral busca coordenar o comportamento dos indivíduos em comunidade por meio do endosso conjunto de normas. Claramente, esse procedimento é incompatível com o pressuposto realista de descoberta de fatos morais.

Mas qual a implicação disso para o RE? A implicação diz respeito à resolução de desacordos morais, especificamente no cenário do ATGM. Dado que, em última instância, o que explica o desacordo moral no ATGM são os conflitos de atitudes e não crenças morais, apelar para a observação, como sugere o realista, não parece uma alternativa promissora. A observação irá nos diz apenas que diferentes propriedades constituem a referência dos predicados morais, e nada para além disso. Por outro lado, a coordenação de atitudes parece ser a alternativa restante. Essa seria outra concessão importante que o defensor do RE teria que fazer para o expressivismo antirrealista: em última instância, o que guiaria a resolução do conflito moral no cenário do ATGM, dado o fato do desacordo, seriam normas sobre a coordenação das atitudes dos falantes.

3.2.2. Atribuição de pensamento moral

Outro ponto em que o defensor do RE deve fazer uma concessão importante ao antirrealismo expressivista diz respeito à atribuição de pensamento moral em cenários de comunidades hipotéticas. No capítulo anterior, apresentei uma versão diferente do ATGM a que denominei de *Argumento Invertido da Terra Gêmea Moral* (AITGM). Irei relembrá-la em termos gerais para que possamos notar o suposto déficit do RE.

Grosso modo, o AITGM simplesmente inverte os pontos de similaridade e diferença entre T e TG da seguinte forma. Ambos, T e TG adotam teorias de primeira ordem similares sobre a propriedade *N* que regula causalmente o uso dos termos morais de seus habitantes. Digamos que tanto em T quanto em TG trata-se de uma forma de consequencialismo segundo a qual o correto é aquilo que *maximiza o agregado de felicidade*. No entanto, os membros das duas comunidades usam predicados como ‘correto’ e ‘bom’ de forma bastante diferente. Enquanto em T usa-se ‘correto’ para aprovar ações, pessoas, instituições, os falantes normalmente agem de acordo com juízos sobre o que é correto, usam tal predicado em discussões relacionadas ao bem-estar etc., em TG usa-se ‘correto’ para propósitos bem diferentes. Entre eles, o uso de ‘correto’ não diz respeito ao bem-estar, os habitantes de TG não usam tal termo para aprovar ações, pessoas, instituições, não agem de acordo com juízos sobre o que é correto etc.

Como vimos anteriormente, não atribuiríamos linguagem e pensamento genuinamente morais aos habitantes de TG. Ou seja, julgaríamos que tal comunidade está fazendo algo bem diferente do que moralidade com o uso de predicados como ‘correto’. No entanto, note que se

assumíssemos que o que determina o conteúdo semântico dos predicados morais é apenas a propriedade rastreada pela teoria de primeira ordem – no presente caso, a teoria segundo a qual correto é aquilo que maximiza o agregado de felicidade – teríamos que aceitar que os membros de TG possuem pensamento e linguagem moral, pois ambos adotam a mesma teoria. Mas atribuir pensamento e linguagem moral aos habitantes de TG parece inaceitável.

Isso mostra que o que determina a nossa atribuição de pensamento e linguagem moral a comunidades hipotéticas tal como no ATGM não é o conteúdo das crenças de tais indivíduos (o que representaria o lado realista), mas sua disposição conativa (o que representaria o lado expressivista). Aqui teríamos outra importante concessão que o defensor do RE teria que fazer ao expressivismo antirrealista. Que, em última instância, crenças apenas não são suficientes para atribuir pensamento moral a uma comunidade, mas é necessário o conteúdo expressivista.

3.2.3. Devemos rejeitar o RE?

Como vimos, a conjunção do realismo com o expressivismo não parece lidar muito bem com certas características da moralidade, especialmente no que diz respeito ao contexto do ATGM. O realista acaba tendo que conceder vários pontos ao expressivista e isso acaba colocando o alcance do realismo em segundo plano. Assim, diante desses problemas levantados por Rubin contra o RE, o ponto deve ser o seguinte: temos razão para rejeitar o RE? A resposta de Rubin é de que sim. Como ele afirma, "à luz desses problemas, realistas devem resistir a adotar uma semântica moral híbrida" (RUBIN, 2015, p. 709).

No entanto, para finalizar esse capítulo, gostaria de propor uma conclusão mais *moderada*. Os problemas do RE a respeito do desacordo, investigação moral e atribuição de pensamento moral, não são logicamente inconsistentes com o aspecto realista da teoria de Copp. São custos da teoria e podem subtrair alguns pontos metaéticos, por assim dizer. Mas o defensor do RE pode aceitar tais custos. Vejamos.

O RE busca compatibilizar duas tradições metaéticas, a saber, realismo e expressivismo, e a grande virtude dessa abordagem é a possibilidade de dar conta de problemas de ambos os lados. Assim, certamente haverá problemas cujas repostas ficarão à cargo do aspecto realista e problemas cujas respostas ficarão à cargo do aspecto expressivista. Dizer que o lado realista do RE lida melhor com a explicação da verdade e objetividade dos juízos morais não nos dá razão suficiente para rejeitar o lado expressivista. Da mesma forma, dizer que o lado expressivista do RE lida melhor com o ATGM não é razão suficiente para rejeitar o aspecto

realista. O defensor da teoria híbrida busca uma abordagem mista justamente pelo seu potencial de resolver conflitos dos dois lados. Por isso, acredito que os custos apontados por Rubin podem fazer com que o RE perca alguns pontos a seu favor, mas isso não significa que tal teoria deve ser definitivamente recusada.

Neste sentido, para voltar a questão que coloquei acima sobre se temos a melhor resposta possível para o ATGM e o RMN estaria livre de tal desafio semântico, podemos dizer que, em comparação com as abordagens que temos visto no decorrer do presente trabalho, o RE é a estratégia mais promissora. No entanto, dizer isso não significa que o RMN está livre de problemas. Isso porque a adoção de um hibridismo metaético representaria uma mudança ampla para defensores do RMN, tais como os *Realistas de Cornell*, por exemplo, cuja abordagem é unicamente realista. Aceitar uma teoria híbrida como o RE, implicaria em reconhecer justamente o ponto positivo que o ATGM pretende evidenciar, a saber, que há um tipo de conteúdo conativo nos juízos morais que é indispensável.

4. Conclusão

Neste capítulo, abordei uma estratégia de réplica ao ATGM que busca desenvolver abordagens semânticas alternativas ao tipo de teoria apresentada por Boyd. Tentei argumentar que o tipo de teoria semântica tal como sugerida por Brink, a proposta de uma SMNE, possui vários custos e não representa grandes avanços para o RMN, uma vez que é vulnerável a reformulações do ATGM. A ideia de uma semântica bi-dimensional, como o RE híbrido de Copp é muito mais promissora. O hibridismo tem seus custos no placar metaético, mas tais custos não parecem ser fatais. Portanto, se a ideia de combinar realismo e expressivismo for plausível (e essa parte do condicional não considero aqui) temos uma boa resposta ao ATGM. Estaria, então, o RMN livre de problemas? Certamente, a maioria das formas de RMN não, já que são exclusivamente realistas. Esse tipo de conclusão sugere que, para escapar ao desafio semântico proposto pelo ATGM, o defensor do RMN precisa fazer uma mudança significativa em favor do hibridismo.

CONCLUSÃO

Como vimos no decorrer desta investigação, grande parte da literatura a respeito do *Argumento da Terra Gêmea Moral* (ATGM) converge no sentido de que tal desafio semântico não implica em problemas para o *Realismo Moral Naturalista* (RMN). No entanto, com exceção de alguns artigos, não havia nenhum trabalho que se dedicasse especificamente à análise generalizada dessa literatura. Nesta tese, tentei preencher essa lacuna. O objetivo fora defender, de forma contrária à essa aparente convergência, que o ATGM resiste à maior parte dos ataques de seus críticos e que, portanto, há um obstáculo relevante à aceitação do RMN.

Há uma conclusão de ordem mais geral e uma série de conclusões particulares que fundamentam a conclusão geral do presente trabalho. As conclusões particulares são as que seguem:

(a) O ATGM possui autonomia e independência em relação ao *Argumento da Terra Gêmea*. Portanto, a estratégia de recusar o ATGM apelando para sua suposta falha em preservar a analogia com tal argumento (Margolis et. al, 1999; Geirsson, 2014) deve ser recusada. (Capítulo 2).

(b) As três propostas (Copp, 2001; Merli, 2002; Plunkett e Sundell, 2013) a favor da tese de que desacordos morais genuínos não requerem identidade extensional dos predicados usados pelos falantes falham. Portanto, até então não temos uma abordagem que mostre a falsidade da segunda premissa do ATGM. (Capítulo 3).

(c) Contrariamente ao que sugerem as três principais propostas que visam recusar a terceira premissa do ATGM (Levy, 2010; Viggiano, 2007; Sonderholm, 2012), temos razão para manter a *Intuição da Univocidade Semântica*. (Capítulo 4).

(d) Das duas abordagens que propõem metassemânticas alternativas em favor do RMN (Brink, 2001; Copp, 2001), a primeira deve ser recusada e a segunda, apesar de ser o melhor tipo de réplica ao ATGM, sugere um distanciamento relevante dos pressupostos mais ortodoxos do RMN. (Capítulo 5).

Essas quatro conclusões indicam que devemos recusar as quatro principais linhas de réplica ao ATGM. Isso sugere duas conclusões de ordem mais geral:

(1) O ATGM é sólido.

Se (1), então

(2) Temos razão para recusar o RMN e manter alguma forma de não-cognitivismo.

Como podemos notar, (2) é a principal contribuição de ordem geral do presente trabalho. E, para finalizar, dois esclarecimentos são necessários.

Primeiro, se o ATGM é a objeção central ao RMN e busco mostrar que tal argumento sobrevive aos mais variados tipos de ataque, porque não concluo que temos razão suficiente para rejeitar, de uma vez por todas, qualquer proposta realista naturalista? Segundo, se temos razão para manter alguma forma de não-cognitivismo, qual? E por que não desenvolvi isso com mais detalhes neste trabalho?

É demasiado exigente sugerir que há *knockouts* definitivos na metaética. Para voltarmos ao ponto de David Enoch (2011), a atitude preferível é encarar o terreno da metaética como um jogo de soma de pontos. Se uma teoria é vulnerável a determinada objeção, ela perde pontos significativos. Se, por outro lado, faz importantes progressos explanatórios, soma pontos. As melhores teorias serão aquelas que somarem mais pontos, serão aquelas que apresentarem mais benefícios do que custos. E a divisão do trabalho requer do metaeticista medir e pesar esse placar metaético. Por isso, prefiro a conclusão mais modesta de que o ATGM representa uma significativa perda teórica para o RMN. Em outras palavras, a contribuição do presente trabalho é mostrar que as abordagens realistas naturalistas perdem importantes pontos no placar metaético.

Além disso, há dois modos de se defender uma teoria metaética. O primeiro é positivo e consiste em enunciar as principais teses de determinada teoria e mostrar como lida com problemas metaéticos específicos. O segundo é negativo e consiste em fornecer razões para se recusar teorias alternativas. O papel desempenhado pelo ATGM abrange esses dois aspectos. Ele sugere que temos evidência para rejeitar o RMN, por um lado, e que temos evidências para aceitar o não-cognitivismo, por outro, uma vez que a melhor explicação para o juízo intuitivo de tal argumento é dada pela abordagem conativa. Minha contribuição aqui é para com a parte negativa da defesa do não-cognitivismo, apenas. Ou seja, tentei fornecer razões contra um forte adversário do não-cognitivismo, a saber, o RMN.

REFERÊNCIAS BIBLIOGRÁFICAS

AYER, A. J. *Language, Truth and Logic*. Harmondsworth: Penguin Books, 1971.

BALDWIN, T. *G. E. Moore*. Routledge, New York, 1990.

BALL, D. Relativism, Metasemantics, and the Future. *Inquiry*, 2020.

BARKER, C. The dynamics of vagueness. *Linguistics and Philosophy*, 25, p. 1–36, 2002.

BLACKBURN, S. *Spreading the World*. New York: Oxford University Press, 1984.

BOLINGER, R. J. Metalinguistic Negotiations in Moral Disagreement. *Inquiry*, 2020.

BOYD, R. How to be a Moral Realist. In SAYRE-MCCORD, G. (ed) *Essays on Moral Realism*, Ithaca: Cornell University Press, 1988.

BRANDT. R. *Ethical Theory: The Problems of Normative and Critical Ethics*. Prentice Hall, 1959.

BRINK, D. O. Moral Realism and Skeptical Arguments for Disagreement and Queerness. *Australian Journal of Philosophy*, 62:2, p. 111-125, 1984.

BRINK, D. O. *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press, 1989.

BRINK, D. O. Realism, Naturalism, and Moral Semantics. *Social Philosophy and Policy*, 18 (2), p. 154-176, 2001.

CAPELLEN, H. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press, 2018.

COPP, D. Milk, Honey, and the Good Life on Moral Twin Earth. *Synthese*, Volume 124, N. 1, p. 113-137, 2000.

COPP, D. Realist-Expressivism: A Neglected Option for Moral Realism. *Social Philosophy and Policy*. 18, (2), p. 1-43, 2001.

COPP, D. Why Naturalism? *Ethical Theory and Moral Practice*, Vol. 6, p. 179-200, 2003.

DALL'AGNOL, D. *Valor intrínseco: metaética, ética normativa e ética prática em G. E. Moore.* 2º Ed: Florianópolis, Editora da UFSC, 2014.

DALL'AGNOL, D. (Org.) *Metaética: algumas tendências*. Florianópolis: Ed. da UFSC, 2013.

DARWALL, S., GIBBARD, A. & RAILTON, P. *Moral discourse & Practice: Some Philosophical Approaches*. New York, Oxford: Oxford University Press, 1997.

DAVIS, W. Implicature. *The Stanford Encyclopedia of Philosophy*. ZALTA, E. N. (Ed), URL: <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=implicature>. Acesso em: 08/08/2022.

DREIER, J. Internalism and Speaker Relativism. *Ethics*, Vol. 101, n. 1, p. 6-26, 1990.

DONELLAN, K. Reference and Definite Descriptions. *Philosophical Review*, 75 (3), p. 281-304, 1966.

DUNAWAY, B. McPHERSON, T. Reference Magnetism as a Solution to the Moral Twin Earth Problem. *Ergo*. Vol. 3, n. 25, 2016.

ENOCH, D. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford University Press, 2011.

FELDMAN, F. Introductory ethics. Upper Saddle River, NJ: Prentice Hall, 1978.

FELDMAN, F. The Naturalistic Fallacy: What it is and What it isn't. In.: SINCLAIR, N. *The Naturalistic Fallacy*. Cambridge University Press: Cambridge, 2018.

FINLAY, S. Four Faces of Moral Realism. *Philosophy Compass*, 2, (6), p. 820-849, 2007.

FOOT, P. Moral Arguments. *Mind* 67:502-513, 1958.

GAMPEL, E. H. Ethics, Reference and Natural Kinds. *Philosophical Papers* 26: 2, p. 147-163, 1997.

GEACH, P. Assertion. *Philosophical Review*. 74, (4), 449-475, 1965.

GEIRSSON, H. Moral Twin Earth: The Intuitive Argument. *Southwest Philosophy Review*, 19 (1): p. 115-124, 2003.

GEIRSSON, H. Moral Twin-Earth and Semantic Realism. *Erkenntnis*, 62, p. 353-378, 2005.

GIBBARD, A. *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press, 1990.

GIBBARD, A. *Thinking how to live*. Cambridge: Harvard University Press, 2003.

GOWANS, C. Moral Relativism, *The Stanford Encyclopedia of Philosophy.* EDWARD, N. Z. (Ed). URL: <https://plato.stanford.edu/archives/sum2019/entries/moral-relativism>. Acesso em: 08/08/2022.

GRICE, P. *Studies in the Way of Words*. Cambridge: Harvard University Press, 1989.

HARE. R. M. *The Language of Morals*, New York: Oxford University Press, 1952.

HARMAN, G. *The Nature of Morality*. New York: Oxford University Press, 1977.

HARMAN, G. Moral explanations and natural facts: can moral claims be tested against moral reality? *Southern Journal of Philosophy*, 24, p. 57-68, 1986.

HENNING, T. Moral Realism and Two-Dimensional Semantics. *Ethics*, Vol. 121, n. 4, p. 717-748, 2011.

HORGAN, T. & TIMMONS, M. New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research*, Volume 16, p. 447-465, 1991.

HORGAN, T. & TIMMONS, M. Troubles on Moral Twin Earth: Moral Queerness Revived. *Synthese*, 92, p. 221-260., 1992a.

HORGAN, T. & TIMMONS, M. Troubles for New Wave Moral Semantics: The Open Question Argument Revived. *Philosophical Papers*, Volume 21, No. 3, p. 153-175, 1992b.

HORGAN, T. & TIMMONS, M. Copping Out On Moral Twin Earth. *Synthese* 124, p. 113-137, 2000.

HORGAN, T. & TIMMONS, M. Analytical Moral Functionalism Meets Moral Twin Earth. In: RAVENSCROFT, I. (org). *Minds, Ethics, and Conditionals: Themes From The Philosophy of Frank Jackson*. Oxford: Oxford University Press, 2009.

JACKSON, F. & PETTIT, P. Moral Functionalism and Moral Motivation. *Philosophical Quarterly* 45: p. 20-40, 1995.

JACKSON, F. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press, 1988.

KAGAN, S. Normative ethics. Boulder, CO: Westview Press, 1998.

KAVETSKI, S. *Realismo, Naturalismo e Semântica Moral*. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro de Filosofia e Ciências Humanas, Programa de Pós-Graduação em Filosofia, 2017.

KEYZER, J. *Twin Earth and the Normativity of Meaning*. (PhD Dissertation) – University of Otago, Dunedin, New Zeland, 2016.

KRAEMER, E. R. On The Moral Twin Earth Challenge to New Wave Moral Realism. *Journal of Philosophical Research* 16: p. 467-472, 1991.

KRIPKE, S. *Naming and Necessity*. Cambridge, MA: Harvard University Press, 1980.

LAURENCE, S., MARGOLIS, E. & DAWSON, A. Moral Realism and Twin Earth. *Facta Philosophica* 1, p. 135-165, 1999.

LENMAN, J. The Externalist and the Amoralist. *Philosophia*, 27, 441–457, 1999.

LENMAN, J. What is Moral Inquiry. Proceedings of the Aristotelian Society, Supplementary Volumes, 81, p. 63-81, 2007.

LEVY, N. Moore on Moral Twin Earth. *Erkenntnis*, (75), p. 137-146, 2011.

LYCAN, W. *Philosophy of Language: a Contemporary Introduction*. New York: Routledge, 2000.

MACFARLANE, J. *Assessment Sensitivity: Relative Truth and its Aplications*. Oxford: Oxford University Press. 2014.

MACKIE, J. *Ethics: Inventing Right and Wrong*. New York: Penguin, 1977.

MCPHERSON, T. e DUNAWAY, B. Reference Magnetism as a Solution to the Moral Twin Earth Problem. *Ergo: An Open Access Journal of Philosophy*, 3, p. 639-679, 2016.

MERLI, D. (2002). Return to Moral Twin Earth. *Canadian Journal of Philosphy* 32: 2, p. 207-240.

MILLER, A. *An Introduction to Contemporary Metaethics.* Cambridge: Polity Press, 2011.

MOORE. G. E. *Principia Ethica*. (Revised Edition). Cambridge: Cambridge University Press, 1993.

NEALE, S. Coloring and Composition. In: MARASUGI, K e STAINTON, R (eds). *Philosophy and Linguistics*. Westview Press, p. 35-82, 1999.

OGDEN, C. K. & RICHARDS, I. A. *Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt, Brace & World: New York, 1946.

PLUNKETT, D & SUNDELL, T. Disagreement and the Semantic of Normative and Evaluative Terms. *Philosopher's Imprint*, vol. 13, n. 23, 2013.

PLUNKETT, D & SUNDELL, T. Metalinguistic Negotiation and Speaker Error. *Inquiry*, 2019.

PUTNAM, H. The Meaning of "Meaning". In *Mind, Language and Reality*. Philosophical Papers, Volume 2, Cambridge University Press, p. 215-271, 1975.

RAILTON, P. Naturalistic Realism in Metaethics. In.: McPHERSON, T. & PLUNKETT (eds), D. *The Routledge Handbook of Metaethics*. Routledge, New York, 2018.

RIDGE, M. Moral Non-Naturalism. *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed). URL: <https://plato.stanford.edu/archives/fall2019/entries/moral-non-naturalism>. Acesso em: 08/08/2022.

RIDGE, M. *Impassioned Belief*. Oxford University Press: Oxford, 2014.

Ross, W. D. *The right and the good*. Indianapolis: Hackett Publishing Company, 1930.

RUBIN, M. Sound Intuitions on Moral Twin Earth. *Philosophical Studies* 139 (3): p. 307-327, 2008.

RUBIN, M. Biting the Bullet on Moral Twin Earth. *Philosophical Papers* 43 (2): 285-309, 2014a.

RUBIN, M. On Two Responses to Moral Twin Earth. *Theoria* 80 (1): 26-43, 2014b.

RUBIN, M. The Promise and Perils of Hybrid Moral Semantics for Naturalistic Moral Realism. *Philosophical Studies* 172 (3): 691-710, 2015a.

RUBIN, M. Normatively Enriched Moral Meta-Semantics. *Philosophy and Phenomenological Research* 91 (2): 386-410, 2015b.

SHAFER-LANDAU, R. *Moral Realism: A Defense*. Clarendon Press: Oxford, 2003.

SIDGWICK, H. The methods of ethics. Indianapolis: Hackett, 1981.

SINCLAIR N. *The Naturalistic Fallacy*. Cambridge University Press: Cambridge, 2018.

SMART, J. J. C. Ethics and science. Philosophy, 56(218), 449–465, 1981.

SONDERHOLM, J. Unreliable Intuitions: A New Reply to the Moral Twin-Earth Argument. *Theoria*, (79), p. 76-88, 2013.

SCHROEDER, M. *Being For: Evaluating the Semantic Program for Expressivism*. Oxford University Press: Oxford, 2008.

SCHROEDER, M. *Noncognitivism in Ethics*. Routledge, New York, 2010.

SCHROETER, L. & SCHROETER, F. Normative Realism: Co-Reference Without Convergence? *Philosophers's Imprint*, Vol. 13, n. 13, 2013.

STEVENSON, C. L. *Ethics and Language*. New Heaven, Yale University Press, 1958.

STEVENSON, C. L. *Facts and Values: Studies in Ethical Analysis*. New Heaven and London, Yale University Press, 1963.

STURGEON, N. Moral Explanations. In David Copp and David Zimmerman (eds.), *Morality, Reason and Truth*. Totowa, NJ: Rowman & Littlefield, 1985a.

STURGEON, N. Moore on Ethical Naturalism. *Ethics*, 113 (3), p. 528-556, 2003.

STURGEON, N. Ethical Naturalism. In COPP, D. (ed) *The Oxford Handbook of Ethical Theory*. Oxford: Oxford University Press, 2006.

SZABÓ, Z. G. The Distinction Between Semantics and Pragmatics. In: LEPORE, E. e SMITH, B. (eds) *The Oxford Handbook of Philosophy of Language*. Oxford, Oxford University Press, 2009.

TERSMAN, F. *Moral Disagreement*. Cambridge University Press: Cambridge, 2006.

VAN ROOJEN, M. Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument. In: SHAFER-LANDAU, R. (Org.) *Oxford Studies in Metaethics*. Volume 1. Oxford: Clarendon Press, p. 162-193, 2006.

VAN ROOJEN, M. *Metaethics: A Contemporary Introduction*. New York: Routledge, 2015.

VIGGIANO, A. Ethical Naturalism and Moral Twin Earth. *Ethical Theory and Moral Practice*, Vol. 11, n. 22, p. 213-224, 2008.