UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS

DEPARTAMENTO DE FILOSOFIA

PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

João Lourenço de Araujo Fabiano

Melhoramento humano: heurística evolutiva e riscos existenciais

São Paulo

2014

João Lourenço de Araujo Fabiano

Melhoramento humano: heurística evolutiva e riscos existenciais

Dissertação apresentada ao programa de Pós-Graduação em Filosofia do Departamento de Filosofia da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, para a obtenção do diploma de Mestre em Filosofia, sob a orientação do Prof. Dr. Osvaldo Frota Pessoa Jr.

São Paulo
2014

"Se puderes olhar nas sementes do tempo
E dizer qual semente irá germinar e qual não irá
Fale então comigo."

- William Shakespeare, Macbeth

RESUMO: O objetivo desta pesquisa é explorar a motivação e as potenciais complicações do uso da tecnologia para melhorar fundamentalmente a condição humana. Inicialmente a pesquisa se debruçará sobre alguns pressupostos filosóficos básicos para a discussão deste melhoramento. Para tal será abordado a heurística evolutiva proposta por Anders Sandberg e Nick Bostrom, em seguida será apresentado brevemente alguns traços básicos da condição humana – a saber: cognição, moralidade e ligação afetiva – de acordo com a perspectiva da psicologia evolucionista, um passo importante na heurística evolutiva supramencionada. A seguir o trabalho versará especificamente sobre melhoramentos que tenham como alvo a própria moralidade humana, inicialmente sobre as fortes motivações de realizar tal melhoramento, e ao final sobre os riscos e problemas – tanto filosóficos como técnicos – de tentar realizar tal modificação na moralidade humana. Tentativamente, a análise será original ao (1) aceitar pressupostos dos defensores do melhoramento moral, e sua conclusão de que o mesmo é um imperativo caso conduzido de maneira correta, (2) abandonar alguns dos possíveis contra-argumentos, no entanto, também (3) concluir a existência de severos problemas em potencial no que tange ao melhoramento moral.

ABSTRACT: The intent of this research is to investigate the motivations and potential risks of using technology to alter the human condition. Firstly, it will explore some of the basic philosophical assumptions behind such discussions. Hence, it will evaluate the evolutionary heuristics proposed by Anders Sandberg and Nick Bostrom and its potential for solving many issues arising when considering human enhancement, therefore introducting one basic philosophical ground when arguing for or against these modifications. Thence, it will be given an introduction to some basic traits of the human condition, e.g.: cognition, morality and pair-bonding, from the perspective of Evolutionary Psychology. Such traits will be then considered as targets for human enhancement. These are important steps in, and thus a application of, the aforementioned evolutionary heuristics. Secondly, this dissertation will specifically investigate the risks of using technology to alter human morality. It will focus on the possibility that attempting to improve human moral dispositions – moral enhancement – could in fact yield a future without moral value. This analysis will be tentatively novel in that it will focus on risks that could arise *even if* the claims of moral enhancement advocates are true and some arguments against it unsound.

Sumário

Introdução

O desenvolvimento tecnológico humano vem alcançando patamares cada vez mais elevados e tem submetido ostensivamente a natureza ao controle humano. Que a velocidade com a qual os meios técnicos de controle da natureza avançam sobrepõe a velocidade do desenvolvimento ético humano que permitiria o uso responsável desta técnica tem sido objeto de preocupação de inúmeros teóricos da modernidade (e.g.: ADORNO & HORKHEIMER, 1991). No entanto, ainda pouco exploradas são as consequências éticas de uma categoria particular e nova do desenvolvimento técnico, a saber, a possibilidade de controlarmos e modificarmos a própria natureza humana segundo nossa vontade. Uma vez que a própria natureza humana é vista como objeto direto da manipulação tecnológica, o impacto que a tecnologia tem sob a natureza humana não é mais dado indiretamente pela modificação do meio e da relação do homem com este meio. Quando aspectos básicos humanos como raiva, alegria, tristeza, memória, raciocínio e moralidade são passíveis de alteração tecnológica, insere-se um novo aspecto problemático no uso ético da tecnologia (SAVULESCU et al., 2011). O presente trabalho almeja estudar possíveis encaminhamentos para resolver essa problemática. Quais seriam os usos éticos de tais tecnologias? Quais seriam seus usos catastróficos?

Entende-se como *melhoramento humano (human enhancement)* a modificação da condição humana pelo uso da tecnologia para um aumento do bem estar humano. Estes melhoramentos têm sido alvo de intenso debate dentro de campos como bioética e ética aplicada. Um exemplo recente deste debate é a discussão acerca do uso por estudantes de drogas que aumentam o desempenho cognitivo (KAPNER, 2003). Alguns falam de leis antidoping em testes para garantir uma igualdade maior e chamam a atenção para a ausência de estudos empíricos comprovando a segurança e eficácia de muitas dessas drogas (MEHLMAN, 2004). Alguns melhoramentos cognitivos em potencial serão abordados e os mais recentes estudos empíricos serão expostos. Outro exemplo que será estudado neste trabalho é o uso de melhoramentos morais, tal como o neuro-hormônio ocitocina. Estudos apontam que esta substância eleva os níveis de altruísmo, confiança e generosidade dos participantes. Como veremos, autores consideram o uso dessa droga para operar um necessário melhoramento moral humano

(SAVULESCU & PERSSON, 2012. LIAO & ROACHE, 2011). Como será explorado na dissertação, eles argumentam que sem esse melhoramento moral, existe uma probabilidade considerável de que a humanidade não consiga lidar com os desafios cooperativos enfrentados neste século (i.e.: controle do aquecimento global). Serão exploradas nessa dissertação algumas das principais teorias que tentam explicar o comportamento cooperativo humano e a maneira pela qual tomamos decisões de maneira ética. Serão também abordados os diversos aspectos problemáticos de realizar uma modificação na nossa moralidade, tais como: (1) a fragilidade dos valores humanos, que implica que uma pequena alteração em nossa moralidade possa destruí-la por completo e (2) o problema da natureza auto-reforçadora da moral, que torna qualquer alteração possivelmente irreversível.

O primeiro capítulo será dedicado a introduzir o referencial teórico básico usado no decorrer do trabalho para discorrer sobre o melhoramento humano. Em especial será exposta a heurística evolutiva que norteará boa parte da dissertação. Também será feita uma breve exposição de alguns grupos de melhoramentos em potencial, e algumas posições do debate neste campo serão introduzidas. O segundo capítulo abordará as motivações para e as questões conceituais envoltas no melhoramento moral: a tomada de decisão consequencialista versus deontológica e a cooperação – e falta dela – em dilemas sociais. A seguir, no terceiro capítulo, será feito um levantamento de alguns problemas filosóficos do melhoramento moral que poderiam fazer com que este implicasse o que presupostamente evitaria, uma catástrofe ou a própria extinção da humanidade. Tal levantamento irá duplamente refutar e se beneficiar de críticas já feitas ao melhoramento moral. Um dos problemas levantados – o da complexa interação entre cooperação entre indivíduos e cooperação entre grupos -, bem como outros problemas sobre a natureza da difusão social destes melhoramentos, serão explorados de maneira mais técnica através de uma simulação social realizada no capítulo quatro. Os dois últimos capítulos foram parcialmente baseados no projeto de doutoramento do presente pesquisador, e constituem portanto uma análise tentativa.

Capítulo I

O melhoramento humano: Considerações Gerais

1.1 Introdução ao referencial teórico

1.1.1 O transhumanismo

Inicialmente se fará necessária a apresentação do referencial teórico básico a ser usado na pesquisa. Tal referencial está inscrito na vertente intelectual que preconiza o uso racional da tecnologia para melhorar a condição humana, conhecido como transhumanismo. Esse melhoramento não é obtido apenas através de dispositivos tecnológicos que nos auxiliam na vida cotidiana, mas também por meio de modificações nos níveis mais fundamentais da condição humana (YUDKOWSKY, 2007). Assim como no humanismo, há uma valorização do humano, mas a ênfase está naquilo que podemos nos tornar. O debate acerca da modificação da condição humana se torna cada vez mais necessário, uma vez que ela se torna cada vez mais possível e inevitável. É de grande importância que a eticidade e desejabilidade do desenvolvimento e da aplicação destas tecnologias sejam discutidas e estabelecidas antes da aplicação em larga escala destas técnicas em desenvolvimento eminente. Quase toda a tecnologia até os dias de hoje se voltou, primordialmente, a alterar a natureza e a relação do homem com o meio, alterando indiretamente a condição humana. Os novos melhoramentos vislumbrados pelo transhumanismo, no entanto, se voltam diretamente a aspectos fundamentais da condição humana. Num prefácio a um volume dedicado ao estudo acadêmico desses melhoramentos, os autores afirmam:

> Ainda sim, através das dramáticas reviravoltas da era moderna, as constantes fundamentais da natureza humana – mortalidade humana, um repertório partilhado de emoções e temperamentos, uma gama de percepções e capacidades intelectuais básicas – continuaram relativamente um ponto de referência fixo que podiam ser uma ponte entre diferenças culturais e ideológicas. Mas nas décadas recentes, avanços radicais na genética e nas neurociências, e na computação e em outras formas de tecnologia, levantaram a possibilidade de que estamos à beira de mais uma revolução, dessa vez não na nossa relação com o mundo natural, mas na nossa relação com nós mesmos. Nossos corpos, até mesmo nossos sentimentos, pensamentos, e capacidades intelectuais, estão também gradualmente entrando na esfera do controle e da manipulação científica.

> (...) Parece que em breve nós seremos capazes de melhorar radicalmente as capacidades humanas bem além da normalidade. (SAVULESCU et al., 2011, p. 2)[1].

Existem inúmeros exemplos de aplicações da tecnologia para realizar esse tipo de modificação: aumento do tempo de vida através do uso dos recentes avanços na medicina (DE GRAY, 2007), aumentar nossa capacidade cognitiva com o uso de diversas substâncias (BOSTROM & SANDBERG, 2006), manipular os estados de sono e vigília (SANDBERG & RAVELINGIEN, 2008), a memória (SANDBERG & LIAO, 2008), as emoções (SANDBERG & SAVULESCU, 2008), etc. Cada uma dessas aplicações vem acompanhada de questionamentos éticos: se devemos ou não realizá-las, como devemos implementá-las e se elas devem ser acessíveis a todos. Por exemplo, a manipulação dos estados de vigília e sono – que já é possível até certo grau – pode tanto significar um aumento do tempo de lazer do indivíduo se esse tempo for livre, como uma diminuição do tempo de descanso se esse tempo tiver de ser usado no trabalho (SANDBERG & RAVELINGIEN, 2008).

A primeira instituição declaradamente transhumanista foi fundada pelos filósofos David Pearce e Nick Bostrom, este último, um dos maiores representantes do movimento e diretor do principal polo de pesquisa acadêmica na área. Os trabalhos deste filósofo na área pertinente à presente pesquisa serão a principal fonte norteadora bibliográfica para o desenvolvimento da mesma. Cabe, portanto, breve exposição dos seus trabalhos e formação. Bostrom teve formação acadêmica em diversas áreas relevantes a temática transhumanista. Com graduação em Filosofia ele obteve doutorados nas seguintes áreas: Filosofia, Física e Neurociência Computacional. É diretor do Instituto para o Futuro da Humanidade – uma entidade acadêmica situada em Oxford destinada a pensar questões globais sobre o progresso tecnológico da humanidade. Foi realizada uma visita a este instituto pelo autor da dissertação, os frutos dessa visita serão oportunamente mencionados nos capítulos posteriores. O Instituto para o Futuro da Humanidade reúne diversos pesquisadores de destaque nas suas respectivas áreas. Os trabalhos de Bostrom e seus colaboradores já foram traduzidos para mais de 20 línguas e possuem mais de 100 traduções e reedições. Dentre os temas de suas mais de 200 publicações, incluem-se principalmente: fundamentos da teoria da probabilidade, metodologia científica e racionalidade, melhoramento humano, riscos

[1] Todas as citações presentes na dissertação foram traduzidas pelo autor do original em inglês.

catastróficos globais, filosofia moral e consequências da tecnologia futura (BOSTROM, 2012). Bostrom define o transhumanismo da seguinte forma:

> (1) O movimento cultural e intelectual que afirma a possibilidade e desejabilidade de melhorar fundamentalmente a condição humana através da aplicação da razão, especialmente ao desenvolver e tornar amplamente acessíveis tecnologias que eliminem o envelhecimento e que melhorem substancialmente as capacidades intelectuais, físicas e psicológicas humanas.
>
> (2) O estudo das ramificações, promessas, e perigos em potencial das tecnologias que irão nos permitir superar limitações humanas fundamentais, e o estudo relacionado dos aspectos éticos envolvidos em desenvolver e usar tais tecnologias. (BOSTROM, 2004, p. 4)

O segundo ponto da definição pode ser entendido como a análise dos benefícios e dos riscos das tecnologias que nos permitem ir além das limitações humanas, bem como do fundamento teórico e filosófico que norteia essa análise. Neste campo tem se realizado um grande número de pesquisas acadêmicas em tópicos específicos concernentes às mais diversas tecnologias e suas ramificações éticas. A pesquisa planejada nesse projeto se desenvolverá dentro de um desses tópicos e irá estudar sua fundamentação filosófica, realizando um estudo de caso de alguns dos recentes avanços farmacológicos que permitem um melhoramento da cognição humana.

1.1.2 O melhoramento da condição humana

Introduzida a perspectiva geral do referencial teórico, cabe delimitar dentro dele o tema a ser estudado. Existem muitas possíveis definições do que pode ser considerado ou não um melhoramento da condição humana. Algumas definições partem do pressuposto que qualquer modificação que não seja a cura ou tratamento de alguma doença pode ser considerada melhoramento. Outras de que os melhoramentos consistem em modificações que levem a um nível de funcionamento para além do considerado normal. Para os propósitos do trabalho será usada uma definição de melhoramento que foca no bem estar do individuo: “Definição de melhoramento a partir do bem-estar: Qualquer mudança na biologia ou psicologia de uma pessoa que aumente as chances de ela levar uma boa vida no conjunto relevante de circunstâncias” (SAVULESCU, 2011, p .29)

Além disso, no contexto desta pesquisa, tal melhoramento deve implicar no aumento dessa capacidade dentro ou além do considerado estatisticamente normal. As

modificações da medicina cujo alvo é o tratamento ou a cura de uma doença, para levar o individuo de um funcionamento subnormal para um normal, ficam excluídas da definição de “melhoramento” no presente trabalho. Ainda, esse melhoramento deve ter como alvo certos aspectos considerados fundamentais da condição humana e os seus processos: a cognição, a moralidade e a ligação afetiva. Num primeiro momento, a pesquisa a ser realizada irá estudar alguns pressupostos filosóficos de análise para esses melhoramentos e não os melhoramentos em si, e ao final a metodologia de analise ali desenvolvida será aplicada a alguns melhoramentos específicos. Mais especificamente, será usada uma heurística de análise desenvolvida por Bostrom, que será resumida na próxima seção.

1.1.3 Uma heurística de análise

Uma das principais contribuições teóricas norteadoras do estudo da modificação da cognição humana foi feita no artigo “A sabedoria da natureza: uma heurística evolutiva para o melhoramento humano” (BOSTROM & SANDBERG, 2009a). Este será o principal texto usado durante a pesquisa, juntamente como algumas outras publicações de Bostrom e Sandberg acerca do tema. Inicialmente será exposta a estrutura básica do artigo, e ao final será aplicada a teoria ali desenvolvida a algumas tecnologias já disponíveis. Neste artigo, Bostrom e o neurocientista Anders Sandberg se propõem a elaborar uma heurística que garanta um desenvolvimento e aplicação seguros desse tipo de tecnologia. O primeiro passo dessa heurística se dá a partir do reconhecimento da complexidade do organismo humano e da necessidade do seu entendimento prévio antes que seja possível realizar uma modificação. O principal aspecto desse entendimento deve ser feito no nível funcional: qual o papel original que um determinado aspecto (traço) a ser modificado desempenha?. Essa pergunta necessita ser respondida com uma análise do passado evolutivo do traço e um estudo de como e por que esse determinado design surgiu. Em última análise, esse melhoramento irá constituir numa modificação de um trabalho já realizado pela evolução.

O segundo passo da heurística é uma análise das motivações para modificar o traço em questão. Segundo os autores, as motivações para a realização desse tipo de modificação são inúmeras; citemos alguns exemplos. (1) O processo evolutivo tende a produzir em cada etapa organismos que estão mais bem adaptados a sobreviver no ambiente do que na etapa anterior. Esse processo é extremamente lento e está sempre

inacabado. Assim sendo, é sempre possível acelerá-lo, fazendo uso da tecnologia. (2) Além disso, algumas vezes é o caso que para chegar a um traço evolutivo mais adaptado é necessário passar por um menos adaptado. Nesses casos a evolução dificilmente realiza essa modificação, pois isso significaria passar por uma etapa em que os organismos estão menos adaptados que na etapa anterior. Entretanto, via de regra, os organismos menos adaptados sempre têm uma taxa de sobrevivência menor. Por isso, é muito improvável que a adaptabilidade decresça de uma etapa à outra. Nesse caso a evolução pode ficar presa num ótimo local. Um exemplo é o apêndice vermiforme, localizado no intestino grosso. Se o apêndice não existisse, ele não inflamaria, mas para que ele deixe de existir ele teria de ir diminuindo aos poucos, geração após geração. No entanto um apêndice ligeiramente menor que o nosso tem uma maior incidência de inflamação (KOUTROUBAKIS & VLACHONIKOLIS, 2000) e consequentemente a evolução dificilmente realizaria esse passo desfavorável. (3) Nossos valores podem divergir dos valores indiretamente maximizados pela evolução. Para que o processo evolutivo possa ocorrer, os organismos devem morrer, mas isso não está necessariamente entre os nossos interesses (DE GRAY, 2007). Além disso, a evolução tende a maximizar o sucesso reprodutivo dos genes, sendo que qualquer beneficio ao indivíduo é colateral. Assim, existem inúmeros aspectos que nós enquanto indivíduos consideramos relevantes, mas que não aumentam à adaptabilidade genética. Os autores citam alguns exemplos:

> Bem estar emocional, liberdade de dor severa ou crônica, amizade e amor, memória de longo prazo, habilidade matemática, atenção e consciência, musicalidade, apreciação artística e criatividade, apreciação literária, confiança e auto-estima, prazeres saudáveis, energia mental, habilidade de concentração, pensamento abstrato, longevidade, habilidades sociais. (BOSTROM & SANDBERG, 2009a, p. 395).

(4) Finalmente, um dos principais motivos para a realização de modificações é que o ambiente no qual evoluímos no passado não é mais o ambiente em que vivemos. Por milhões de anos a evolução nos moldou para sobrevivermos numa sociedade de caça e coleta de não mais que algumas centenas de habitantes, em que o maior avanço tecnológico era a pedra lascada. No entanto, vivemos hoje numa sociedade de bilhões de seres humanos, altamente interconectada, com tecnologias que têm o potencial de extinguir a humanidade inteira se mal utilizadas. Por exemplo, como a vida no passado era extremamente incerta, nós desenvolvemos por seleção natural uma preferência desproporcional de obter um benefício imediatamente, em oposição a num futuro

próximo (TERHERSON, 1999), mesmo ela sendo injustificada nos dias de hoje. Além disso, segundo Bostrom e Sandberg, essas modificações seriam as mais fáceis de realizar, pois basta identificar uma mudança nas necessidades ambientais para entender que um design passa a ser preferível em relação a outro:

> Se conseguirmos identificar mudanças específicas no nosso ambiente que deslocaram o ponto do custo/benefício ótimo de desideratas de designs competidores para certa direção, talvez consigamos encontrar intervenções relativamente fáceis que poderiam "ressintonizar" o custo/benefício para um ponto mais próximo do ponto ótimo presente. Tais intervenções de ressintonia podem estar entre os frutos mais baixos na árvore dos melhoramentos, frutos ao nosso alcance mesmo na ausência de tecnologias médicas super avançadas. (BOSTROM & SANDBERG, 2009a, p. 381)

A pesquisa irá ainda abordar outras classes de motivações para modificar a condição humana enquanto produto evolutivo. Neste campo, o artigo "Quebrando as correntes da evolução: a promessa de melhorado por design" (POWELL & BUCHANAN, 2011a) traz ampla e aprofundada discussão sobre as razões pelas quais o processo evolutivo é inerentemente imperfeito aos olhos humanos. Neste artigo, aspectos gerais do processo evolutivo biológico são discutidos e limitações inerentes a ele são expostas, tais como a existência de ótimos locais, impossibilidade de macro mutações coordenadas e insensibilidade à qualidade de vida pós-reprodutiva. Com base nesta discussão, a presente pesquisa de mestrado pretenderá defender que, partindo-se de alguns pressupostos gerais sobre o processo evolutivo determinante de certos traços humanos, é possível mostrar a existência de diversas razões para se esperar que certas classes de melhoramentos gerem um relevante benefício em oposição aos seus custos. Uma vez identificada a função, o passado evolutivo e a motivação de por que modificar determinado traço, é necessário realizar uma análise de custo/benefício nessa modificação, baseando-se no atual estado da pesquisa científica. Essa referência à pesquisa científica nem sempre é tão trivial. A análise da ocorrência de um evento no passado foi o principal método que nossa cognição desenvolveu para estimar a probabilidade de o evento ocorrer no futuro. Entretanto, essa ocorrência era sempre algo que aconteceu com alguém conhecido e nunca um valor estatístico emitido por um grupo de pesquisa especializado (BUSS, 2005, pp. 739-740). Todavia, uma pesquisa cientifica tem amostragem muito maior que a nossa experiência imediata, mas mesmo assim a mente humana tende a favorecer a nossa experiência próxima ao invés da experiência científica (POHL, 2005. pp. 61-78). Portanto, é necessário estudar os processos cognitivos que nos levam a lidar mal com as informações cientificas. Tais

processos são conhecidos como vieses cognitivos e serão discutidos no decorrer do trabalho. Esses e outros processos cognitivos também afetam negativamente, de inúmeras maneiras, o nosso julgamento sobre se certa modificação é benéfica ou não. Por isso seu estudo é de fundamental importância, sendo uma parte da pesquisa dedicada a este tema.

As diversas etapas da heurística elaborada por Bostrom e Sandberg podem ser assim resumidas: (1) análise da função evolutiva do traço a ser modificado, (2) estudo da motivação específica para modificar tal função e, finalmente, (3) análise do custo/benefício de se realizar a modificação.

1.2 A cognição, moralidade e ligação afetiva como produtos evolutivos.

A principal hipótese utilizada para entender o funcionamento da cognição, moralidade e ligação afetiva será a de que eles são formada por um conjunto de adaptações que foram soluções à desafios recorrentes no nosso passado evolutivo:

> O resultado adaptativo de cada um desses testes são somados através do tempo, e o resultado é o desfecho da seleção natural. O design resultante, então, é adaptado a um conjunto particular de circunstâncias, a saber, aquelas propriedades do ambiente de teste, estatisticamente somadas, que moldaram o design eventual. (PLATEK, 2009, p. 99)

O cérebro humano é visto como um mecanismo que processa as informações relevantes do ambiente e responde de maneira apropriada (BUSS, 2005, pp. 5-67). Ele é responsável por gerar aquele conjunto de comportamentos que satisfaça minimamente a melhor alocação de recursos finitamente disponíveis no ambiente. Cada situação recorrente no passado evolutivo produziu um módulo cognitivo especializado no cérebro, responsável por processar os dados relevantes àquela situação, e esse processamento é em sua maior parte inconsciente (COSMIDES & TOOBY, 1997). Esse módulo cognitivo irá gerar um conjunto de estratégias que tem como objetivo a boa alocação de energia e tempo com respeito ao sucesso reprodutivo do organismo em relação a uma determinada situação (BUSS, 2005, pp. 68-72). Nesta fase do trabalho será levantado o conjunto básico de explicações evolutivas referentes a vários mecanismos considerados fundamentais da condição humana (e.g.: ligação afetiva, cuidado parental, moralidade, relacionamentos, agressão e sobrevivência), bem como

aqueles que são especificamente alvos de modificações (e.g.: cognição, tomada de decisão e vieses cognitivos). Por exemplo, certos aspectos básicos de nossa cognição moral, tal como a reciprocidade, o altruísmo e a extrema aversão ao dano serão vistos como soluções apresentadas ao problema da cooperação e interação, frequente dentro de agrupamentos humanos no paleolítico (KREBS, 2011). O modo como – e não só o fato de que – nossas estratégias e propensões afetivas na vida adulta são influenciadas por nossa vida afetiva infantil, vida esta que configura níveis de confiança, afetividade e reciprocidade, também tem raízes e explicações evolutivas, pois a infância é vista como um período em que determinados genes e propensões são ativadas ou desativadas em função de estímulos ambientais (BUSS, 2005, pp. 255-445). Esta etapa do trabalho corresponde ao segundo passo da heurística mencionada na seção anterior.

1.3 Algumas classes de razões para melhorar produtos evolutivos

Alguns aspectos gerais do processo evolutivo biológico geram limitações inerentes, tais como a existência de ótimos locais, impossibilidade de macromutações coordenadas e insensibilidade à qualidade de vida pós-reprodutiva. Partindo-se de alguns pressupostos gerais sobre o processo evolutivo determinante de certos traços humanos, é possível mostrar a existência de diversas razões para se esperar que certas classes de melhoramentos gerem um relevante benefício em oposição aos seus custos. Serão abordadas algumas dessas limitações inerentes, compiladas do artigo “Quebrando as correntes da evolução: o prospecto de manipulação genética deliberada em humanos” de Russell Powell e Allen Buchanan (POWELL & BUCHANAN, 2011b, conferir também POWELL & BUCHANAN, 2011a) bem como do já referido artigo “A sabedoria da natureza: uma heurística evolutiva para o melhoramento humano” (BOSTROM & SANDBERG, 2009a).

Que os processos naturais são relativamente amorais e produzem resultados que se distanciam de nossos valores éticos é observado por inúmeros autores. Bostrom (2005) afirma: "Tivera sido a Mãe Natureza um pai ou uma mãe real, ela estaria na cadeia por abuso infantil e assassinato". Richard Dawkins sugere que a evolução é como um relojoeiro cego, e que, portanto não tem qualquer intenção de maximizar nossos valores. No entanto, Powell e Buchanan apontam que a situação pode ser ainda mais grave, para eles a evolução seria como “um funileiro volátil, moralmente cego e firmemente acorrentado” (POWELL & BUCHANAN, 2011a p. 60). Os autores

argumentam que ele é moralmente cego em um duplo sentido: ele não tem o bem-estar humano como um objetivo e não mostra nenhum escrúpulo na sua escolha de meios. Ele é um funileiro porque não produz objetos de acordo com um plano. Ele é volátil porque sempre destrói o seu trabalho eventualmente. Ele é firmemente acorrentado, pois opera sob limitações severas.

Adicionalmente aos já mencionados, podemos listar ainda algumas outras classes de razões que apontam para a existência de melhoramentos exercidos sob traços evolutivos. Tais razões revelam severas limitações do processo evolutivo em produzir traços que valorizamos. Como a evolução primordialmente só seleciona traços com base em seu impacto na reprodutibilidade da indivíduo, (1) existe uma insensibilidade à qualidade de vida pós-reprodutiva, de modo que não existe nenhum padrão claro de seleção para este período da vida e muitas anomalias genéticas costumam surgir conforme os sistemas fisiológicos lentamente degringolam. Dentre os exemplos das disfunções que surgem neste período, temos o câncer e muitos outros processos de envelhecimento. Além disso, (2) a evolução seleciona sempre sob traços que foram os mais adaptados para pressões evolutivas que não existem mais, ela sempre trabalha em cima de um aparato anacrônico, construído em cima de outro aparato e assim por diante. Existe sempre uma ressaca evolutiva. Adicionalmente, (3) como a maior parte da história da vida na terra se deu em Eras muito antigas, grande parte do nosso mecanismo bioquímico mais fundamental só pode ser explicado enquanto respostas evolutivas aos desafios apresentados naquelas Eras. E mesmo os traços que foram selecionados mais recentemente, como o imbricado sistema nervoso central humano, ainda estão em sua maior parte adaptados a períodos anteriores, como o paleolítico – como será discutido na próxima seção, inúmeros vieses cognitivos surgem deste fato. (4) A velocidade com que novas e desejáveis mutações se fixam é extremamente lenta. Caso um grupo de indivíduos com uma nova mutação que os tornassem imunes a AIDS surgisse, poderia levar décadas ou até séculos antes de esta mutação se espalhar pela humanidade. Powell e Buchanan argumentam que seria moralmente errado esperar pela sua fixação e uma alteração genética intencional da humanidade seria nosso dever moral. Ademais, (5) em organismos tão complexos quanto seres humanos, a probabilidade de uma macro mutação ocorrer é nula, uma vez que ela envolveria a coordenação de inúmeras micro mutações de maneira sincronizada – evento extremamente improvável; sendo assim, qualquer melhora fenotípica que implique em

uma mudança estrutural do organismo está fora dos limites da evolução. Como discutido anteriormente, a adaptabilidade dos organismos só pode subir ou manter-se constante, portanto a evolução está sujeita a ficar presa em ótimos locais. Isto é, caso exista uma configuração fisiológica melhor, mas que para surgir envolva uma etapa anterior menos adaptada que a atual, a evolução nunca irá conseguir atingir aquela configuração melhor. Segue-se que (6) a evolução é limitada pela topologia do espaço de trajetórias possíveis entre designs: ela só pode operar caso a trajetória seja ascendente ou plana, caso ela envolva uma trajetória momentaneamente descendente a evolução não poderá tomar tal trajetória. Podemos ilustrar a situação com o seguinte exemplo:

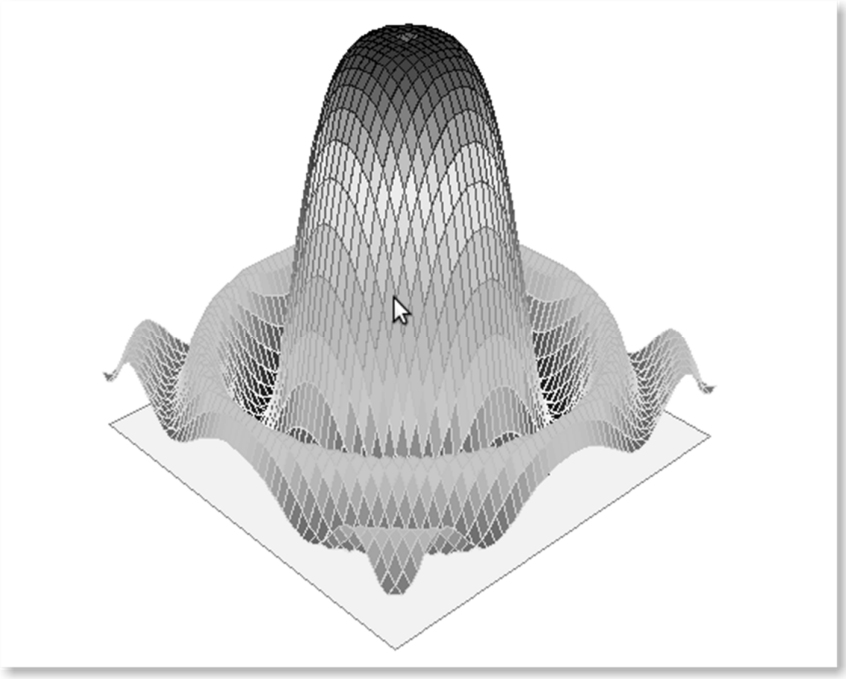

Figura 1 – Exemplo hipotético de uma topologia das trajetórias de design

Se a qualquer momento os processos evolutivos atingirem traços com o grau de adaptabilidade da região azul claro, todos os traços em verde, amarelo e vermelho se tornaram inacessíveis, pois todas as trajetórias até eles envolvem passar por uma região de menor adaptabilidade, ilustrada em azul escuro. Por último, e o mais crucial, está o fato de que (7) a tendência primordial da evolução é selecionar aqueles organismos mais adaptados – com maior taxa de sobrevivência e reprodutibilidade -, ele não é, de modo

algum, a realização dos valores humanos. Apenas tangencialmente o aumento da adaptabilidade envolve a plena realização de alguns valores humanos, e ainda, muitas vezes esse aumento envolve processos que consideremos moralmente indesejáveis. Portanto, uma intervenção intencional humana nas estruturas geradas pela evolução pode melhorar. No entanto, por mais que essas sete razões apresentadas acima forneçam motivações bem gerais para acreditar-se que é possível e factível melhorar traços evolutivos, assim como lembrado por Powell e Buchanan, é importante que tais melhoramentos sejam analisados em sua especificidade técnica e que os processos bioquímicos envolvidos sejam estudados e entendidos.

1.4 Exemplos de traços a serem modificados e os meios já disponíveis

1.4.1 Vieses cognitivos

O estudo dos vieses cognitivos será realizado tanto como exemplo de traço a ser modificado quanto, principalmente, um passo fundamental para evitar os erros sistemáticos cometidos ao absorver informações e analisar os custos e benefícios das novas tecnologias. No passado evolutivo humano, para que determinado algoritmo cognitivo resultasse numa resolução satisfatória de dado problema, não bastava resolvê-lo corretamente, mas resolvê-lo levando em conta inúmeras restrições como, por exemplo, tempo e dispêndio energético. Esse mecanismo de resolução não precisava ser perfeito, apenas bom o suficiente para garantir a sobrevivência do indivíduo:

> Quais pressões seletivas impactaram nos mecanismos de decisão? Antes de tudo é a seleção para fazer a decisão apropriada no domínio dado. Essa pressão específica do domínio não implica na necessidade de realizar a melhor decisão possível, mas em seu lugar uma que é boa o suficiente (uma escolha satisfatória, como Herbert Simon, 1955, colocou) e, na média, melhor do que aquelas dos competidores individuais, dado os custos e benefícios envolvidos (BUSS, 2005 p. 778)

Por isso, nosso cérebro realiza operações que resolvem tarefas de raciocínio através de ‘atalhos’, os quais funcionam bem na maioria dos casos, mas falham em algumas situações. Como os módulos cognitivos que realizam tais tarefas são, segundo a hipótese da psicologia evolucionista, relativamente universais na humanidade, os casos em que nossos ‘atalhos’ falham também apresentam certa regularidade. Essas falhas são conhecidas como vieses cognitivos: desvios sistemáticos que a cognição humana comete da racionalidade. Nem sempre os vieses nos são maléficos, mas em

geral eles nos privam de uma visão correta da realidade. Nesta fase do trabalho será realizada uma análise de alguns vieses cognitivos considerados mais prejudiciais e os considerados cruciais de se evitar na análise dos possíveis riscos e benefícios das modificações da cognição humana.

Um exemplo clássico e bem estabelecido é o viés da conjunção. Num famoso experimento, quando fornecidas com uma descrição de Linda, graduada em filosofia, solteira, inteligente e preocupada com justiça social, 85% dos participantes avaliaram a probabilidade de Linda ser 'caixa de banco' como menor do que ela ser uma 'caixa de banco e ativa no movimento feminista' (KAHNEMAN, 2002, p. 24). Mas, segundo a teoria da probabilidade, isso é necessariamente falso, pois a probabilidade de um evento A acontecer isoladamente é sempre maior ou igual à probabilidade de que o evento A ocorra concomitantemente a outro evento B qualquer (POHL, 2005. p. 23).

Outro exemplo já mencionado é o viés que nos impede de absorver informações estatísticas de certo formato. Verificou-se que quando o individuo é apresentado com a informação de que um evento tem certa probabilidade de ocorrer – i.e.; a mortalidade do câncer de mama é 10% –, ele não consegue atualizar de maneira correta suas crenças relacionadas com base nessa informação. Isso demonstra que, na realidade, ele não consegue absorver corretamente a informação fornecida, ou ao menos essa absorção não é feita de acordo com as regras de inferência probabilística. No entanto, quando a informação vem no formato "a cada mil casos esse evento ocorreu tanto número de vezes", as nossas crenças relacionadas são atualizadas corretamente (POHL, 2005. pp. 61-78). Isso ocorre porque, como já foi dito, no nosso passado evolutivo a forma pela qual tínhamos acesso a informações era por meio de ocorrências diretas e não por probabilidades abstratas, e o segundo formato de apresentação se aproxima muito mais da descrição de uma ocorrência (BUSS, 2005, p. 739-740). É interessante salientar que a maioria dos sujeitos experimentais de tais estudos são alunos universitários que passaram por uma formação básica em metodologia cientifica e estatística. Infelizmente, quase toda a literatura cientifica e de divulgação disponíveis são apresentadas em termos de probabilidades. Isso compromete seriamente a nossa capacidade de julgar a eficácia e os riscos de certos avanços tecnológicos.

Mas talvez o viés que afeta de maneira mais forte a análise das potencialidades da tecnologia, e se torna de extrema relevância para a pesquisa a ser realizada, é o viés do status quo. Esse viés é analisado em detalhe por BOSTROM & ORD (2006). Ele

consiste numa preferência injustificada pela configuração atual das coisas, em detrimento de outras. Diversos experimentos demonstram esse viés: em um deles, os participantes são perguntados sobre onde investir uma hipotética herança. Ambos os grupos recebem as mesmas instruções, exceto que a um grupo é dito que a herança já está investida num fundo de médio risco (o status quo), enquanto que à outra metade essa informação é omitida. Um número muito maior dos participantes da metade a qual foi dada uma descrição do status quo preferiu continuar no fundo de médio risco, do que participantes da outra metade preferiram investir nesse risco (SAMUELSON & ZECKHAUSER, 1988), ainda que os retornos e riscos deste fundo sejam iguais nos dois grupos. Esse viés se manifesta no caso de melhoramentos cognitivos, quando indivíduos preferem não introduzir certo melhoramento, ou até mesmo continuar com um antigo, pois ele acarretaria uma mudança do *status quo*, ainda que esse melhoramento seja benéfico. Uma maneira que Bostrom e Ord propõem de detectar esse viés consiste em: ao se perguntar se uma mudança de certo parâmetro é benéfica, se perguntar também se outras mudanças em direções diferentes do mesmo parâmetro também são benéficas. Caso se decida que todas as possíveis mudanças são maléficas, devemos suspeitar da presença do viés, pois é muito improvável que estejamos com o melhor valor possível desse parâmetro em questão. Nas palavras dos autores: "... se um contínuo de parâmetros admite uma larga gama de valores possíveis, só um pequeno subconjunto destes pode ser um ótimo local, então é implausível *prima facie* que o valor atual deste parâmetro deva acontecer de estar num desses raros ótimos locais" (BOSTROM & ORD, 2006 p. 665) .

1.4.2 Melhoramentos cognitivos

Mesmo praticando nomadismo, o ambiente das savanas para os humanos era relativamente estável comparado ao nosso, e uma vez tendo aprendido as principais técnicas de caça, coleta de alimentos e a linguagem, restavam poucas coisas que deviam permanecer em aprendizado contínuo. Em oposição, nosso ambiente moderno está em contínua mudança, num avanço tecnológico e cultural em constante aceleração (KURZWEIL, 2001). Da hipótese da psicologia evolucionista de que estamos adaptados ao ambiente do paleolítico, segue que falta ao nosso cérebro, entretanto, a capacidade de estar constantemente absorvendo novas informações e de lidar com elas rapidamente, uma habilidade que decai rapidamente ao envelhecer (HERTZOG et al.,, 2003). Em

contra partida, às vezes sobra capacidade de armazenamento de informação no longo prazo, que em poucos anos se torna relativamente inútil. A eficiência com a qual tomamos decisões e executamos tarefas é fortemente influenciada por quão bem nossa cognição funciona, quão bem conseguimos processar as informações relevantes à situação, e lembrar o conteúdo que lhe é pertinente. Deste modo, melhorar nossa cognição tem um impacto dramático tanto no nível individual quanto social. A educação obrigatória por muitos anos de nossas vidas já é um meio pelo qual a sociedade garante um aumento da eficiência da nossa cognição: um ano de educação formal aumenta em média 3 pontos de QI (FALCH & SANDGREN, 2011). No entanto, outros meios farmacológicos veêm se tornando disponíveis e já vêm sendo usados no meio acadêmico. Os benefícios potenciais de tal tecnologia são grande, caso hipoteticamente desenvolvermos no futuro uma droga que aumente nossa eficiência cognitiva em apenas 1% sem efeitos colaterais e possa ser universalmente aplicada – tal como a educação –, ela pode significar um aumento da produção econômica mundial da ordem de bilhões por ano. Porém, toda a nossa capacidade cognitiva foi otimizada para lidar com situações que não enfrentamos mais.

Já existem drogas que permitem manipular a nossa memória, diminuindo a fixação de eventos indesejáveis (por exemplo, com o uso de porpranolol; ver PITMAN et al., 2002) e aumentando a fixação de informações desejáveis (por exemplo, o aricept; ver GRÖN et al., 2005). Existem ainda outras drogas que aumentam nossa memória de curto prazo, melhorando nossa capacidade de lidar com sistemas informacionais complexos de modo simultâneo (modafinil e ritalina) (TURNER et al., 2008). Apesar destas drogas se mostrarem promissoras, ainda existem poucos estudos sobre seu uso em indivíduos saudáveis no longo prazo, visando os objetivos de manipulação da memória mencionados. De maneira mais pertinente, faltam também estudos que analisem a validade ética de executar tal tipo de manipulação. Ainda que numa hipotética ausência completa de efeitos colaterais, a pergunta de se devemos ou não manipular a memória com este grau de liberdade continua em aberto. O presente mestrado visa suprir minimamente a carência deste campo de analise ética dos melhoramentos. Atualmente, a humanidade já faz uso de uma droga que aumenta nossa habilidade física e cognitiva há milhares de anos: a cafeína. No entanto, este composto apresenta diversos efeitos colaterais, discutidos a seguir, que não estão presentes em

outras drogas recentes. Esta classe de drogas cujo principal efeito é um aumento do nosso desempenho cognitivo é denominada *nootrópicos*.

Cabe realizar uma pequena análise comparativa de dois compostos utilizados para aumentar nosso desempenho cognitivo: a cafeína e o modafinil. A cafeína é dos mais antigos e mais usados nootrópicos. Ela mimetiza o neurotransmissor adenosina e se liga ao seu receptor, inabilitando-o. A adenosina tem um papel inibitório no cérebro. Em quantidades moderadas, a cafeína produz uma vasta gama de efeitos benéficos à saúde (SMITH, 2002). No entanto, a quantidade média de cafeína consumida ultrapassa largamente a recomendada (ILLY & VIVIANI, 1995; NCDT, 2011). Nestas doses, em longo prazo ela aumenta a incidência de infartos (LESON et al., 1988; GREENBERG, J. A. et al., 2007); além disso, após longo período de uso, se desenvolve tolerância e a interrupção causa depressão, irritabilidade, dores e sonolência (JULIANO & GRIFFITHS, 2004).

O modafinil tem seu mecanismo de funcionamento mediado através dos neurotransmissores histamina e dopamina. A histamina regula o estado de vigília. Ele causa um aumento generalizado de outros transmissores de papel predominantemente excitatórios. Existem inúmeros estudos do uso de modafinil em indivíduos saudáveis (BARANSKI et al., 2004. LI et al., 2007. MÜLLER, et al., 2004. TURNER, et al., 2003. MULLER, ET AL., 2012). Existem ao menos 5 estudos do uso de modafinil em indivíduos saudáveis para aumento do desempenho cognitivo (BARANSKI et al., 2004. LI et al., 2007. MÜLLER, et al., 2004. TURNER, et al.,2003, MULLER, et al., 2012). Um deles é um grande estudo com mais de 300 pessoas (TURNER et al., 2003), onde foi relatado um aumento de memória de trabalho verbal e numérica. Houve um aumento do tempo de resposta nos testes, mas também um aumento significativo na taxa de acertos. Outro, um estudo com 60 indivíduos (MULLER, et al., 2012) relatou um aumento da memória de trabalho especial, planejamento e tomada de decisão, assim como reconhecimento de padrões visuais e memória de curto prazo. No entanto, uma pesquisa mais extensa sobre os efeitos de longo prazo do modafinil ainda se faz necessária.

Fazendo-se uso da heurística de Bostrom e Sandberg, é necessário questionar sobre o papel evolutivo do traço a ser modificado. Dentre seus inúmeros efeitos, o modafinil aumenta a memória de trabalho – a memória que armazena no curto prazo os vários fatores sendo considerados num dado processo cognitivo. Estudos em psicologia

evolucionista e cognitiva demonstram que a memória de curto prazo tem um papel crucial na nossa habilidade de detectar correlações no ambiente no qual estamos imersos. No entanto, estes estudos argumentam que um simples aumento da memória de curto prazo ocasionaria um decréscimo na habilidade de detectar correlações (BUSS, 2005, pp. 788-789). Acredita-se que uma memória de curto prazo expandida, sem um respectivo aumento da capacidade de processar essa memória, acarrete em combinatórias complexas demais para serem processadas e tornem as correlações menos evidentes. Em contrapartida, algumas das teorias sobre vieses cognitivos os explicam como sendo atalhos que a cognição humana toma para resolver certas tarefas (POHL, 2005 pp. 10-14). Na maioria dos casos, esses atalhos conseguem chegar ao propósito desejado com um menor poder de processamento do que o usado caso nosso cérebro realizasse o raciocínio ideal. Em alguns casos, no entanto, esses atalhos erram e alguns desses erros podem se revelar catastróficos. Ao aumentar a nossa capacidade cognitiva, os nootrópicos têm um grande potencial de refinar nosso raciocínio, elevando nossa capacidade de processamento e consequentemente evitando que tenhamos que usar atalhos que nem sempre chegam à resposta correta. Este pequeno exemplo de estudo de caso demonstra quanto o campo do debate ético acerca dos melhoramentos cognitivos pode ser enriquecido por uma visão mais integral das várias questões pertinentes, bem como por uma fundamentação ética mais substancial. Essas são tarefas nas quais a discussão filosófica sempre se mostrou frutífera e necessária. Mais uma vez, reforça-se que é de grande importância que a eticidade e desejabilidade do desenvolvimento e da aplicação destes tipos de tecnologias sejam discutidos e estabelecidos antes da aplicação em larga escala destas técnicas em desenvolvimento eminente.

1.4.3 Melhoramentos morais

A nossa eficiência cognitiva só é capaz de aumentar a velocidade com a qual processamos informação segundo um determinado objetivo, e quão rápido conseguimos obter esse objetivo. Ela não diz respeito a quais objetivos devemos perseguir. Um indivíduo com objetivos escusos, cognitivamente melhorado, só significa um risco moral maior. Para que possamos superar inúmeros problemas morais enfrentados pela humanidade, em especial o problema da cooperação, Ingmar Persson e Julian Savulescu sugerem que nossa única alternativa é melhorar tecnologicamente a condição humana (PERSSON & SAVULESCU, 2012). Existem drogas que já se mostraram eficazes em

aumentar nossa empatia, compaixão, altruísmo e confiança (KRUGER et al., 2012. KOSFELD et al., 2005. DOMES, et al., 2007). Os efeitos de uma dessas drogas sob a tomada de decisão moral estão sendo alvo de estudos empíricos. Cabe ressaltar, que no entanto a própria natureza do que consistiria em um melhoramento moral é muito mais controversa do que no caso cognitivo, e o presente mestrado irá explorar as controversas filosóficas envolvidas em capítulos subsequentes (i.e.: incerteza moral). No entanto, dada sua relevância para o escopo geral do trabalho, iremos abordar um aspecto controverso em especial dos valores humanos – a complexidade de valores – na seção a seguir.

1.4.3.1 A complexidade dos valores humanos

É oportuno abordar uma tese que aponta importantes riscos à modificação dos valores morais humanos – bem como colateralmente a qualquer modificação humana em geral. O pesquisador de Inteligência Artificial Eliezer Yudkowsky defende a tese de que os valores humanos, nossas preferências, têm um grau de complexidade extremamente alta, de maneira que uma pequena alteração em apenas um diminuto aspecto de um valor humano poderia alterar por completo nossa moralidade (YUDKOWSKY, 2011). A sua argumentação é traçada primordialmente para defender que criar uma Inteligência Artificial extremamente poderosa é uma tarefa difícil, pois como é difícil fazer com que ela modele corretamente aquilo que valorizamos, ela pode acabar criando um futuro sem o nosso conjunto de valores. No entanto, sua argumentação também demonstra que caso queiramos mudar a natureza humana, em especial nossos valores e nossa moral, também estamos expostos ao risco de criar um futuro que não consideramos moralmente bom. Sem entender corretamente como nossa moralidade funciona, corremos o risco de alterar nosso sistema motivacional e moral de maneira perigosa e irreversível. Uma vez que nossos valores e morais sejam alterados, dificilmente seria possível voltar atrás, pois estes seres humanos com diferentes valores não iriam querer voltar ao sistema de valores antigos – eles não teriam nenhuma razão para tal.

A tese da complexidade de valores afirma que as nossas preferências, aquilo com o que nos importamos, não podem ser resumidas em algumas regras ou preceitos simples. Acompanhada dessa tese vem a da fragilidade dos valores, que afirma que por conta de sua complexidade, a perda de apenas uma fração das regras que compõem

nosso sistema de valores pode levar a resultados que consideraríamos inaceitáveis. Por exemplo, caso seja feito um modelo que leve em conta quase todos os nossos valores de maneira correta, mas falhe em incorporar a aversão ao tédio, ele poderia gerar indivíduos presos em uma única experiência ótima pelo resto da eternidade (YUDKOWSKY, 2011. p. 11) – um cenário que parece longe do que concebemos como desejável. Muitas escolhas humanas podem ser representadas por regras simples – o desejo de sobreviver produz inúmeras ações e objetivos. Mas o desejo de sobreviver não é o único valor humano, saúde, beleza, sabedoria, harmonia, amor, amizade são alguns exemplos de coisas que a humanidade tende a valorizar, mas que não poderiam ser resumidas em um conjunto simples de regras.

Existem inúmeras razões para essa complexidade e fragilidade de nossos valores. Yudkowsky levanta alguns motivos que serão expostos a seguir. Nossos valores foram o produto de inúmeros eventos contingentes e acidentais, com muitos processos aleatórios envolvidos: evolução natural, ótimos locais, a topografia do espaço possível de design e valores, condições locais da terra, a história humana, evolução memética, etc. Uma vez que a mente humana sempre tende a procurar por explicações causais simples, eficientes e elegantes, é muito difícil para ela emular corretamente os processos aleatórios e caóticos envolvidos na criação dos nossos valores. Ademais, uma vez que vários desses valores não são mais adaptados ao nosso ambiente, é difícil predizer suas consequências. Como a maioria dos nossos desejos são subobjetivos evolutivos instrumentais para a reprodução e sobrevivência, mas nós os experienciamos como emoções descontextualizadas em primeira pessoa, é difícil de inferir corretamente a organização hierárquica e funcional desses desejos. Por fim, como nós somos os únicos agentes inteligentes capazes de realizar valores, nós não podemos avaliar quão improváveis nossos valores são no grande espaço total de valores possíveis. Com isso, temos a tendência e antropomorfizar o futuro e falhar em perceber que ele só irá conter o que nós valorizamos se forem tomadas ações para que estado improvável exista.

Desta maneira, é extremamente improvável possuir nosso conjunto específico de valores em vez de uma das outras infinitas permutações no conjunto de todos os valores possíveis e, portanto, melhorar este conjunto de valores enquanto tentamos manter sua configuração básica é uma tarefa extremamente difícil. Planejar o futuro de maneira errada porque nós temos um entendimento raso dos valores humanos é um risco muitas vezes subestimado para a humanidade. Uma vez que um futuro ausente de nossos

valores se torne realidade, não existe mais volta: “Mexa demais na dimensão errada, e a representação física desses valores irá se despedaçar – e não irá voltar, pois não existirá mais nada que queira trazê-la de volta. E os referentes desses valores – um universo desejável – não terá mais nenhuma razão física para existir.” (YUDKOWSKY, 2009).

Não só os melhoramentos morais, como qualquer outro melhoramento, se tornam problemáticos frente à complexidade e fragilidade dos valores humanos. Qualquer melhoramento deverá sempre ser norteado por aquilo que eticamente valorizamos, e se nosso entendimento de nossos valores é insuficiente para declarar que uma pequena modificação em um traço fundamental não iria resultar numa alteração indesejada, então um maior desenvolvimento das áreas como psicologia moral e ética se fazem necessárias antes da aplicação de tais melhoramentos. A presente dissertação é propriamente um esforço neste sentido.

1.4.4 Melhoramentos afetivos

Por último, a nossa ligação afetiva com parceiros de longo prazo consiste num dos mais importantes acontecimentos da vida. A qualidade desta relação tem demonstrado enorme poder preditivo sobre níveis de felicidade, estabilidade financeira, quantidade e qualidade na vida sexual, bem estar e sucesso dos filhos, mortalidade e saúde (NOLLER & FEENEY, 2002), ao ponto que o aumento das taxas de divórcio desde a segunda metade do século XX tem sido comparado a uma crise mundial de saúde, ou uma pandemia. Recentemente tem-se começado a pensar sobre o uso da tecnologia para aumentar a qualidade destes relacionamentos, com o uso de drogas que aumentam nossa estabilidade e ligação afetiva no longo prazo (SANDBERG & SAVULESCU, 2008; EARP at al. 2012). Algumas intervenções tecnológicas têm-se mostrado eficazes em melhorar a resolução de conflito por casais (DITZEN et al., 2009), mas maiores desenvolvimentos nessa área são extremamente necessários.

Capítulo II

Inaptidão moral humana

2.1. Introdução

Neste capítulo serão inicialmente abordados dois processos diferentes de tomada de decisão moral. O primeiro são as perspectivas consequencialistas e deontológicas em dilemas morais, o segundo é o comportamento cooperativo em dilemas sociais. Será mostrado que o comportamento humano nessas situações pode e deve ser melhorado com o uso da tecnológica, traçando a mesma linha argumentativa desenvolvida por Julian Savulescu e Igmar Persson no livro "Inaptos para o futuro?" (SAVULESCU & PERSSON, 2012).

2.2 Éticas Consequencialista e Deontológica

Existem duas principais vertentes dentro da ética contemporânea: a consequencialista e a deontológica. A consequencialista define a moralidade ou bondade de uma ação com base apenas nas suas consequências (SINNOTT-ARMSTRONG, 2011), e as regras morais seriam apenas um meio para se chegar às consequências desejáveis. A perspectiva deontológica mede a moralidade ou corretude de uma ação pelo quanto ela se conforma a determinadas regras e preceitos morais, que devem ser seguidas pelos agentes morais (ALEXANDER & MOORE, 2008).

O consequencialismo pode ser visto como um grande conjunto de teorias éticas que tem em comum medirem a bondade ou retidão de uma ação pela soma de todos os seus desfechos ou consequências. A análise consequencialista pode tomar a forma de uma função de utilidade, na qual a utilidade esperada de uma ação é determinada pela soma da utilidade de cada uma de suas consequências possíveis multiplicada pela probabilidade que a consequência ocorra (FISHBURN, 1970). Uma das formulações mais conhecidas de uma função utilidade é a função de utilidade de von Neumann–Morgenstern:

$$E(A) = p_1 E(D1) + p_2 E(D2) + \cdots + p_n E(Dn)$$

Nexta expressão, a utilidade esperada da ação A (E(A)) é igual a somatória das utilidades esperadas de cada um de seus desfechos (E(Dn)) multiplicadas por suas respectivas probabilidades (p_n). Para que essa equação exista, algumas condições têm de ser preenchidas; a principal delas é que haja uma ordem de preferências segundo a qual o conjunto de todos os desfechos possa ser ordenado. Essa equação é um dos fundamentos da Teoria dos Jogos, que será explorada com mais detalhe numa seção subsequente (VON NEUMANN & MORGESTERN, 1944).

A ação correta na perspectiva consequencialista é sempre aquela que maximiza o valor total das consequências ou desfechos (BENTHAM, 1907), ainda que ela possa infringir alguma regra moral pré-estabelecida. O consequencialismo se baseia numa análise de custos e benefícios das ações e mede sua eticidade a partir do resultado do somatório de todos os bens e males gerados (bens menos males). Desta maneira, ao contrário das teorias deontológicas, não existe ação em si mesma boa ou má, pois elas devem ser medidas por seu resultado.

Cabe mencionar que muitas vezes os termos utilitarismo e consequencialismo são usados de maneira intercambiável – especialmente na pesquisa empírica. No entanto, o utilitarismo é apenas um caso especial de consequencialismo em que as consequências relevantes são aquelas que afetam o bem estar geral. Este equivoco será evitado na medida do possível no decorrer do trabalho, entretanto, ao realizar citações diretas, muitas vezes o termo utilitarismo irá ocorrer significando consequencialismo.

A divisão entre éticas consequencialista e deontológicas, primariamente usada para discernir teorias éticas no interior da filosofia moral, tem sido largamente utilizada – não sem controvérsias (KAHANE & SHACKEL, 2010) – para discernir entre perspectivas morais diferentes na população humana. Deste modo, pessoas consequencialistas tomariam decisões baseadas nas suas consequências, e pessoas deontológicas na adequação ou não às regras. O principal método empírico utilizado para determinar se um indivíduo faz um raciocínio consequencialista ou deontológico são os conhecidos “Problemas do Trem”. Propostos inicialmente por Philippa Foot em 1967, os Problemas do Trem (*trolley problem*), geraram uma vasta gama de dilemas morais de difícil resolução e justificação. Esses dilemas em geral seguem o seguinte esquema: a pessoa A pode tomar uma ação que iria beneficiar muitas pessoas, mas iria contra direitos individuais da pessoa B ou contra certa regra moral X. A resposta a tais problemas pode revelar o tipo de raciocínio moral implícito na escolha feita

(THOMSON, 1976). Nos Problemas do Trem, tomar a ação que viola os direitos individuais de alguém ou alguma regra moral em prol do beneficio geral, revela um raciocínio consequencialista. Por outro lado, não tomar a ação e se ater as regras ou direito individual, revela um raciocínio deontológico.

O exemplo original e mais usado desse dilema pode ser parafraseado do seguinte modo:

"Um vagão está desgovernado em um trilho em direção a cinco trabalhadores que serão mortos se o trem continuar no seu curso presente. Você está do lado do trilho usado pelo vagão, mas você está muito longe dos trabalhadores para avisá-los do desastre iminente.

Do seu lado tem um estranho muito grande que está quieto no seu canto. Se você empurrar essa pessoa nos trilhos na frente do vagão, ele iria parar o vagão e salvar os trabalhadores da morte certa. No entanto, isso quase certamente iria matar o estranho.

Você empurra o homem para salvar a vida dos cinco trabalhadores?"

Neste caso, empurrar tem como consequência o menor número de mortes possível, no entanto infringe a regra deontológica proibindo tirar a vida outro ser humano. Caso o indivíduo empurre o homem, isto revela um raciocínio consequencialista. Caso ele deixe o trem atropelar os trabalhadores, revela um raciocínio deontológico. Existem inúmeras versões destes problemas e as frequências médias de cada tipo de resposta variam conforme o problema é apresentado de uma forma mais ou menos pessoal. Por exemplo, um dilema semelhante ao apresentado há pouco, mas menos pessoal, apresenta quase o dobro de incidência média de respostas utilitárias (BARTELS, 2008):

"Um vagão está desgovernado em um trilho em direção a cinco trabalhadores que serão mortos se o trem continuar no seu curso presente. Você está do lado do trilho usado pelo vagão, mas você está muito longe dos trabalhadores para avisá-los do desastre iminente.

Do seu lado tem um interruptor para os trilhos que pode mudar a trajetória do vagão. Você poderia colocar o vagão em outro trilho e poupar os cinco trabalhadores da morte certa. No entanto, tem outro trabalhador no outro trilho que com certeza irá morrer se você mudar o vagão de trilho.

Você aciona o interruptor para salvar os cinco trabalhadores?”

Dentre as análises a respeito destas duas perceptivas morais, uma se mostrará particularmente relevante no decorrer do trabalho e será brevemente apresentada agora. Joshua Greene argumenta (GREENE, 2007), por meio de inúmeros estudos empíricos, que a resposta deontológica é permeada por intuições morais biologicamente pré-programadas em nossos cérebros e que a resposta consequencialista é baseada em processos cognitivos racionais de tomada de decisão. Conway e Gawronski (2012) resumem boa parte das evidencias empíricas levantadas por Greene:

> A evidência disponível é consistente com a visão de que julgamentos deontológicos são movidos por processos emocionais, enquanto que julgamentos utilitários [consequencialistas] são movidos por processos cognitivos. Por exemplo, centros emocionais do cérebro demonstraram uma ativação maior quando participantes consideraram dilemas morais pessoais envolvendo contato direto com a vítima (Greene et al., 2001) e quando participantes tomaram decisões deontológicas em dilemas morais difíceis (Greene et al., 2004). Participantes tomaram menos decisões deontológicas quando a distância emocional das vítimas foi aumentada (Petrinovich et al., 1993), depois que um vídeo humorístico (...) (Valdesolo & DeSteno, 2006), ou quando eles sofrem danos a regiões ligadas a emoção do cérebro (Ciaramelli et al., 2007; Koenigset al., 2007; Mendez et al., 2005). Reciprocamente, participantes tomaram mais decisões deontológicas quando imaginando o dano á vítima de maneira vivida e em detalhes (Bartels, 2008; Petrinovich & O’Neill, 1996), quando experienciando estresse psicológico (Starcke, Ludwig, & Brand, 2012) e depois de ouvir uma estória moralmente edificante que evocava sentimentos afetuosos (Strohminger, Lewis, & Meyer, 2011).
>
> Enquanto que julgamentos deontológicos têm sido ligados a centros emocionais do cérebro, regiões cognitivas do cérebro foram mais ativadas quando participantes consideraram dilemas morais impessoais nos quais as vítimas estão distantes (Greene et al., 2001) e quando participantes fizeram julgamentos utilitários em dilemas difíceis (Greene et al., 2004). A facilitação de tomada de decisão racional aumentou os julgamentos utilitários (Bartels, 2008; Nichols & Mallon, 2006), enquanto que introduzir um limite de tempo (Suter & Hertwig, 2011) reduziu decisões utilitárias, e, por fim, a sobrecarga cognitiva debilitou os tempos de reação para julgamentos utilitaristas, mas não para deontológicos (Greene et al., 2008). Participantes com uma memória de trabalho maior tinham uma maior probabilidade de tomar decisões utilitaristas (Moore et al., 2008), assim como participantes com estilo de pensamento deliberativo em oposição a intuitivo (Bartels, 2008). (CONWAY & GAWRONSKI, 2012. p. 3)

Greene argumenta que essa longa lista de evidências revela que escolhas deontológicas são respostas instintivas emocionais enquanto que escolhas consequencialistas são baseadas numa análise cognitiva da situação. Ele argumenta que nortear nossas ações com base em respostas emotivas pré-programadas pode nos tornar

especialmente inaptos a lidar com inúmeros problemas do mundo moderno, como por exemplo, uma ampla desigualdade social entre regiões distantes do mundo:

> Quer dizer, suponha que a única razão para nós dizermos que é errado deixar uma criança se afogar, mas que não há problema em ignorar as necessidades de crianças famintas do outro lado do oceano, é que o primeiro caso mexe com nossas emoções e o segundo não. Suponhamos, além disso, que a única razão para crianças longínquas não mexerem com nossas emoções é que nós evoluímos em um ambiente em que é impossível interagir com indivíduos distantes. Nós poderíamos, ainda assim, permanecer do lado de nossas intuições do senso comum? Nós poderíamos, com a consciência tranquila, dizer "eu vivo uma vida de luxúria enquanto ignoro as necessidades desesperadas de pessoas distantes porque eu, por um acidente da evolução humana, sou emocionalmente insensível ao seu sofrimento. De todo modo, o fato de eu não aliviar o seu sofrimento quando eu poderia muito facilmente fazê-lo é perfeitamente justificado"? Eu não sei sobre você, mas eu não me sinto confortável com essa combinação de afirmações. (GREENE, 2007, p. 76)

Essa linha de raciocínio desenvolvida pelo autor parece apontar para o fato de que a análise consequencialista está muito mais apta a lidar com certos problemas contemporâneos globais – pois pondera o conjunto total de consequências globais de uma ação –, em particular os problemas que serão levantados em seções posteriores do presente capítulo. Em tal momento oportuno, essa análise realizada por Greene será retomada.

Cabe aqui realizar um adendo. A análise feita por Greene explicitamente se distancia da perspectiva filosófica clássica a respeito da divisão entre éticas deontológicas e utilitaristas na medida em que reverte a aproximação padrão entre utilitarismo e hedonismo de um lado e deontologismo e imperativo moral de outro. Longe de associar o utilitarismo apenas à emoção de felicidade buscada por um hedonismo, e o deontologismo a um imperativo moral austero, este autor usa certos resultados empíricos para arguir que decisões deontológicas são na realidade guiadas por uma emoção de repulsa moral e decisões utilitárias pelo raciocínio. Greene quer argumentar contra a posição filosófica moral deontológica a partir de dados empíricos da psicologia moral. Essa transversão filosófica – que vai contra a clássica marcante divisão entre o que é e o que deve – é polemica, mas ela não é um pressuposto deste trabalho. Aqui estamos nos preocupando mais com o processo psicológico pelo qual as pessoas de fato raciocinam a moral, e neste campo os resultados a favor da posição de Greene são de bem aceitos para no pior dos casos discutíveis, mas não controversos.

2.3. Dilemas Sociais

2.3.1 Introdução

O modo como estabelecemos relacionamentos sociais, tanto em relações de namoro e casamento, passando pela organização de pequenos grupos, até a organização em escala global, envolve processos de tomada de decisão. Estes processos muitas vezes envolvem conflitos entre o nível individual e social. Tais situações podem ser chamadas de dilemas sociais. Dilemas sociais vêm sendo objeto de estudo interdisciplinar desde a segunda metade do século XX e tiverem sua concepção inspirada pelos desenvolvimentos da Teoria dos Jogos – área da matemática que procura modelar matematicamente interações entre indivíduos. Eles tem uma vasta gama de aplicações em cenários reais, como por exemplo, a prevenção do aquecimento global, desarmamento nuclear, pagamento de impostos, rodízio de carros, etc.

Os dilemas sociais podem ser entendidos como situações nas quais é tentador para cada indivíduo tomar um curso de ações não cooperativo que é melhor individualmente no curto prazo, mas que caso todos tomem aquele curso de ações, todos estariam pior no longo prazo do que se todos cooperassem (VAN LANGE, 2013). Outra definição mais geral e formal diz que dilemas sociais podem ser formalmente definidos por duas propriedades: (1) cada indivíduo tem uma estratégia racional não cooperativa que rende o melhor resultado em todas as circunstâncias e (2) se todos os indivíduos usarem essa estratégia, o resultado coletivo é pior do que se todos cooperassem (DOWES & MESSICK, 2000). A diferença desta última definição para a dada anteriormente é que a primeira leva em conta aspectos temporais. Muitas vezes é o caso que a decisão não cooperativa só é racional no curto prazo individualmente e, portanto, ela não é a escolha mais racional –, mas ainda sim os indivíduos se sentem tentados a tomar a decisão não cooperativa e envolvem-se em um dilema social.

2.3.2 Teoria dos Jogos

Hoje a área de estudo de dilemas sociais se beneficia de teorias interdisciplinares advindas da Matemática, Economia, Psicologia Social, Teoria da Evolução e Neurociência. Antes que estas teorias sejam abordadas, cabe uma recapitulação da origem deste campo em Teoria dos Jogos e um dos mais clássicos e antigos exemplos de dilema social: o Dilema do Prisioneiro.

Como mencionado, a Teoria dos Jogos busca modelar matematicamente interações entre indivíduos (LEVINE, 2013). Indivíduos são vistos como *agentes racionais*, com um conjunto diverso de ações, um conjunto de *desfechos*/*consequências* preferidos e uma função que escolhe a ação que melhor se adequa às suas preferências – uma *função utilidade*, como previamente discutido na seção sobre consequencialismo. Estes indivíduos interagem de modo que um agente deve antecipar as respostas e ações de outro agente para escolher aquela ação que melhor maximiza suas preferências; esta interação é chamada de um *jogo*. Caso o agente só tenha de considerar suas próprias ações, então a Teoria das Decisões modela melhor o comportamento deste agente e a situação não é mais considerada um jogo. Cada agente em um jogo tem que escolher de um conjunto de algoritmos de ação previamente estabelecidos chamados *estratégias*; cada estratégia é uma lista exaustiva de respostas para cada situação possível que possa ser encontrada pelo agente durante o jogo (ROSS, 2012).

2.3.3 Alguns exemplos

O exemplo mais famoso de um jogo é o Dilema do Prisioneiro, adaptado de Ross, 2012: “Suponha que a policia prendeu duas pessoas que eles sabem que cometeram um assalto à mão armada juntos. Infelizmente, eles não têm evidencias o suficiente para condená-los. Eles tem, no entanto, evidência o suficiente para mandar cada um dos ladrões para a prisão por dois anos por conta do crime que eles comentaram para conseguir o veículo de fuga. O inspetor chefe faz a seguinte proposta para cada ladrão: “Se você confessar o assalto à mão armada, implicando o seu parceiro, e ele não confessar também, você sairá livre e ele pegará dez anos.”” Se ambos confessarem, cada um pegará dez anos. Se nenhum dos dois confessar, cada um irá pegar dois anos pelo roubo do carro. Sejam estes os valores de *retorno* para cada desfecho, de modo que desfechos com maior grau de preferência tem retornos maiores:

- Liberdade: 4
- Dois anos na prisão: 3
- Cinco anos na prisão: 2
- Dez anos na prisão: 0

Se ambos confessarem, cada um tem um retorno de 2. Se nenhum confessar, cada um tem um retorno de 3. Se só um deles confessar, mas não o outro, um tem um retorno de 4 e o outro de 0. Como eles são agentes racionais ideais, o ladrão I avalia suas ações pelas consequências levando em conta cada uma das ações possíveis do ladrão II, e o mesmo faz o ladrão II. Para o ladrão I, se o ladrão II confessar então o ladrão I tem um retorno de 2 ao confessar e um retorno de 0 ao recusar confessar. Se o ladrão II recusar, então o ladrão I tem um retorno de 4 ao confessar e um retorno de 3 ao recusar. Portanto, o ladrão I sempre tem um retorno melhor se confessar independente do que o ladrão II faça. O mesmo raciocínio é válido para o ladrão II, ele estará sempre melhor confessando. Alguém poderia argumentar que se ambos os ladrões se recusarem a confessar, cada um teria um retorno de 3 e estariam melhor. No entanto, é necessário analisar cada consequência possível dada cada ação possível do outro ladrão. Se o ladrão I recusar e o ladrão II confessar, o ladrão I tem um retorno de 0. Por conseguinte, de todas as estratégias possíveis, na média, sempre confessar é aquela com melhores desfechos e maiores retornos. Essa estratégia *domina* todas as outras estratégias possíveis, e é uma solução para o jogo. Neste estado, o jogo é considerado em um *equilíbrio , Nash equilibria ou Equilíbrio de Nash*. Podemos resumir a situação descrita acima na seguinte *Matriz de Retornos*:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	2,2	4,0
	Recusa	0,4	3,3

*O par ordenado 4,0 indica, por exemplo, que o Ladrão I teve um retorno de 4 e o Ladrão II de 0.

Tabela 2.3.3.1: Matriz de Retornos para o Dilema do Prisioneiro

Como veremos adiante, a solução racional de uma perspectiva individual nem sempre será a melhor solução possível. No caso do Dilema do Prisioneiro, se ambos conseguissem fazer um acordo pré-estabelecido de cooperar e não confessar, eles conseguiriam obter o melhor desfecho possível (3,3). Evidências demonstram que seres

humanos tendem a fazer a escolha de cooperar muito mais frequentemente do que a análise individual acima indicaria (WEDEKIND & MILINSKI, 1996). Sendo assim, essa situação é um dilema social, pois se cada indivíduo tomar o curso de ações com o maior retorno individual, eles terão retornos coletivos menores do que se cada indivíduo cooperar.

Existem ainda inúmeros exemplos clássicos de situações com conflitos similares. Cabe discorrer sobre os mais famosos. Um deles é o da Caça ao Cervo (*Stag Hunt*), descrito primeiramente por Jean-Jacques Rousseau no *Discurso da Desigualdade*. Na Caça ao Cervo um grupo de indivíduos se organizou para caçar um cervo e estão escondidos em um arbusto esperando o cervo aparecer. Passa-se um longo tempo e o cervo não aparece, no entanto pequenas lebres estão presentes. Um cervo seria capaz de alimentar bem todos os membros do grupo, e uma lebre, ainda que consumida por um só individuo, o alimentaria mal. Caso um dos indivíduos escolha atacar uma lebre, ele irá com certeza espantar qualquer cervo que possa vir a aparecer. Individualmente, cada um dos membros teria um risco menor se atacasse e comesse a lebre em vez de esperar por um cervo que pode ou não aparecer. No entanto, todos os indivíduos estariam melhor caso esperassem pelo cervo. A Caça ao Cervo é ligeiramente diferente do Dilema do Prisioneiro. Neste último a diferença entre o cenário que o outro não coopera e você coopera não são tão altas, por isso esse jogo tem apenas dois equilíbrios. A Caça ao Cervo tem, no entanto, dois equilíbrios diferentes. Em um, os riscos são minimizados, e a estratégia dominante é pagar a lebre, em outro os retornos são maximizados e a estratégia dominante é esperar o cervo. Seja uma caçada com apenas dois indivíduos em que os retornos são assim definidos:

Cervo: 2

Lebre: 1

Nada: 0

A matriz de retornos será:

	Indivíduo II		
Indivíduo I		Cervo	Lebre
	Cervo	2,2	0,1
	Lebre	1,0	1,1

Tabela 2.3.3.2: Matriz de Retornos para o Dilema da Caça ao Cervo

Como podemos ver, caso o indivíduo queira minimizar a probabilidade de ficar com 0 (risco), ele deve optar pela Lebre, enquanto que se quiser maximizar seus retornos, deve optar pelo Cervo. Mais a frente, veremos como existem outros critérios que podem ser usados além dos clássicos minimização de riscos e maximização de retornos – critérios que extrapolam a teoria dos jogos clássica apresentada nessa seção.

No Jogo do Covarde (*Chicken Game*), ambos os indivíduos estão dirigindo em direções contrárias em uma estrada de mão única. Eles podem escolher seguir adiante ou desviar para o acostamento. Caso ambos sigam em frente, eles irão bater. Caso um desvie e o outro não, ele seria considerado o 'covarde'. Atribua-se os seguintes valores de retorno:

Desviar se o outro desviou: 0

Desviar se o outro foi reto: -1

Ir reto se o outro desviou: +1

Bater: -10

A matriz de retornos fica assim construída:

	Indivíduo II		
Indivíduo I		Desvia	Segue
	Desvia	0,0	-1,+1
	Segue	+1,-1	-10,-10

Tabela 2.3.3.3: Matriz de Retornos para o Dilema do Jogo do Covarde

Neste caso, o equilíbrio de Nash depende da estratégia que um indivíduo acha que o outro irá usar. Caso ele ache que o outro irá desviar, a melhor estratégia é seguir em frente, caso ele ache que o outro irá seguir em frente, a melhor estratégia é desviar. Diferentemente dos outros dois dilemas apresentados, neste caso o melhor resultado acontece quando os indivíduos usam estratégias diferentes. Caso ambos escolham seguir, o dano seria máximo. Por conta deste aspecto, acredita-se que o Jogo do Covarde espelhe alguns aspectos de cenários reais que os outros dilemas falham em reproduzir: a saber, a interdependência das ações de cada indivíduo, o fato de que a escolha da melhor estratégia depende de que crença o indivíduo tem a respeito das escolhas dos outros indivíduos.

Existem ainda mais dois tipos de dilemas que se tornaram proeminentes no estudo da Psicologia Social, dada suas altas relevâncias em cenários reais. Os dilemas ‘pegue um pouco’ (*take some*) envolvem situações em que cada indivíduo deve retirar uma pequena parcela de um recurso comum e conter-se de exageros. Em tais situações uma ação que leva a um retorno positivo para o individuo, pode, se exacerbada, levar a um retorno negativo para o coletivo. Exemplos podem ser retirados do uso de recursos naturais. O uso exacerbado de combustíveis fósseis pode gerar inúmeros impactos ambientais de consequências globais, bem como o seu completo esgotamento. Um uso cauteloso desses recursos, no entanto, pode gerar benefícios que compensam um reduzido impacto ambiental. Em contrapartida, existem situações nas quais cada indivíduo deve incorrer um pequeno custo para o bem do coletivo, estas situações são conhecidas como dilemas ‘dê um pouco’ (*give some*). Exemplo de tal situação seria, por exemplo, impostos para criar um sistema público de saúde, onde cada indivíduo incorre um pequeno custo para que todos tenham acesso a saúde de qualidade. Situações nas quais cada indivíduo deve contribuir com sua parcela para a realização de um bem comum são uma importante subcategoria desses dilemas conhecida como ‘Dilemas do Bem Público’ (*Public Goods Dilemmas*). Dentro dos dilemas ‘pegue um pouco’ existe uma importante subcategoria com respeito à partilha de recursos comuns conhecida como ‘Tragédia dos Comuns’ (*Tragedy of the Commons*). Na Tragédia dos Comuns, recursos partilhados por uma comunidade são exauridos, porque os indivíduos não conseguem se refrear de usar este recurso em demasia. Este dilema será de especial importância no presente trabalho, pois será um dos objetos do estudo empírico a ser

realizado com o neuro-hormônio ocitocina. Ele será, portanto, trabalhado com maior detalhe numa seção exclusiva.

Como mencionado, o comportamento dos indivíduos em muitos dos dilemas expostos até aqui depende fortemente de qual curso de ações acredita-se que a maioria dos indivíduos irá tomar. Em dilemas do tipo 'dê um pouco', quanto maior o grau de confiança que um indivíduo tem de que os outros indivíduos irão cooperar e contribuir, maior as chances de que ele também contribua. No entanto, num grupo suficientemente grande e sem mecanismos punitivos, caso o indivíduo saiba que todos irão contribuir, ele pode optar por parasitar o grupo e não contribuir, uma vez que a sua falha em contribuir não terá um efeito perceptível e não será punida. Ele também terá seu comportamento determinado pelas suas propensões intrínsecas à cooperação, individualismo, altruísmo, competição, igualdade e agressividade. Estes seis traços são consideradas as quatro dimensões intrínsecas básicas que afetam a escolha de cada indivíduo e serão discutidas em mais detalhes a seguir.

2.3.4 Teoria da Interdependência

A teoria da interdependência nasce diretamente da análise da Teoria dos Jogos apresentada na seção anterior. Ela parte do princípio de que as ações dos agentes são uma função da estrutura interdependente (a matriz de retornos), os agentes que interagem e uma dinâmica de interação (as estratégias). O que esta teoria traz de inovador é que ela afirma que ocorre uma transformação na matriz de retornos objetivos quando são considerados aspectos sociais, temporais e subjetivos, resultando em uma matriz de retornos subjetivos onde certos resultados podem ser preferidos, ainda que tenham retornos objetivos menores (RUSBULT & VAN LANGE, 2003). Outros autores defendem que as transformações podem ser entendidas como diferentes pesos que as pessoas atribuem a certas configurações específicas de retornos para ele mesmo e para os outros (e.g., VAN LANGE, 1999). Em consonância com essa perspectiva, muitos autores têm usado a perspectiva da Orientação de Valores Sociais (Social Value Orientations – SVO), na qual, dependendo da orientação de cada indivíduo, ele tenderá a valorizar mais certas configurações em detrimento de outras (MURPHY & ACKERMANN & HANDGRAAF, 2011). Podemos resumir as seis orientações mais usadas com a seguinte tabela:

Orientação	**Tipo de desfecho buscado**
Altruísmo	Maximização dos retornos do outro
Cooperação	Maximização da soma dos retornos
Individual	Maximiza o próprio retorno
Igualdade	Minimização da diferença entre os retornos
Competição	Maximização da diferença entre os retornos
Agressão	Minimização dos retornos do outro

Tabela 2.3.4.1: As seis orientações sócio-valorativas

Deste modo podemos transformar a matriz de retornos para o dilema do prisioneiro segundo as mencionadas orientações sócio-valorativas **para o Ladrão I**, com os desfechos preferidos em negrito e com fonte maior:

Altruísmo:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	2,2	4,0
	Recusa	**0,4**	3,3

Tabela 2.3.4.2: Transformação na matriz de retornos pela perspectiva altruísta.

Cooperação:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	2,2	4,0
	Recusa	0,4	**3,3**

Tabela 3.3.4.3: Transformação na matriz de retornos pela perspectiva cooperativa.

Individual:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	2,2	**4,0**
	Recusa	0,4	3,3

Tabela 2.3.4.3: Transformação na matriz de retornos pela perspectiva individual.

Igualdade:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	**2,2**	4,0
	Recusa	0,4	**3,3**

Tabela 2.3.4.5: Transformação na matriz de retornos pela perspectiva igualitária

Competição:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	2,2	**4,0**
	Recusa	0,4	3,3

Tabela 2.3.4.6: Transformação na matriz de retornos pela perspectiva competitiva.

Agressão:

	Ladrão II		
Ladrão I		Confessa	Recusa
	Confessa	2,2	**4,0**
	Recusa	0,4	3,3

Tabela 2.3.4.7: Transformação na matriz de retornos pela perspectiva agressiva.

Neste caso a diferença entre agressão, competição e individual não se faz clara, no entanto, como função de exemplo, caso existisse um cenário no qual ambos ficassem com -10, este seria preferido pelo agressor, mas não pelo competitivo e nem pelo individualista. Já em um cenário no qual ambos ficassem com +10 seria preferido pelo individualista, mas não pelo competitivo e nem pelo agressor. Infelizmente, apenas uma bateria de testes mais complexa e extensa é capaz de plenamente distinguir todas as seis dimensões. Tais testes serão usados no estudo empírico descrito no capítulo V, onde seu papel é mais bem discutido. Os testes se encontram ao final do Anexo D.

Existem ainda duas orientações que recentemente tem levantado interesse, que podem ser chamadas de transformações temporais (VAN LANGE & JOIRMAN, 2008). São elas:

Orientação	**Tipo de desfecho buscado**
Longo prazo	Maximiza retornos futuros
Curto prazo	Maximiza retornos imediatos

Tabela 3.3.4.8: Orientações temporais

Muitas vezes a orientação de longo prazo leva a estratégias mais cooperativas, pois os resultados da cooperação costumam vir a longo prazo. No entanto, tem-se dado atenção a casos de efeitos paradoxais destas orientações (VAN LANGE, 2008). Por exemplo, pode ser o caso que os benefícios de cooperar sejam apenas imediatos, neste

caso a orientação de longo prazo promoveria a não cooperação. A orientação individualista pode promover cooperação caso o resultado da cooperação seja benéfico para o indivíduo. Por outro lado, a orientação cooperativa pode produzir comportamentos agressivos ou competitivos com relação a membros de outro grupo. A orientação agressiva, por sua vez, pode promover a cooperação ao funcionar como mecanismo punitivo para aqueles que não cooperam; está questão será elaborada em maior detalhe na seção 3.2.6.

2.3.5 Perspectiva da adequação

Contrariamente à teoria da interdependência esboçada anteriormente, esta teoria argumenta que nenhum indivíduo racionalmente calcula os retornos numa matriz e que a sua escolha é na verdade determinada pelo o que alguém parecido com ele, na situação dele, iria fazer. A pergunta que o indivíduo se faz é "O que uma pessoa como eu (identidade) faz (regras/heurística) numa situação como essa (reconhecimento) dada essa cultura (grupo)?" (WEBER, 2004). Congruentemente com a teoria da interdependência, a perspectiva da adequação também afirma que as escolhas são determinadas pelas orientações de valores sociais, no entanto não ocorre uma transformação sobre a matriz de retornos clássica e sim apenas um recurso a como o indivíduo se vê normalmente (cooperador, altruísta, competidor, igualitário, agressivo), como ele se vê naquela situação e como ele vê as expectativas do grupo em relação a ele; dados estes aspectos, ele toma sua decisão.

2.3.6 Teoria Evolutiva

Três principais teorias evolutivas são usadas para explicar altruísmo e cooperação em dilemas sociais. A seleção de parentesco explica o altruísmo dirigido a parentes consanguíneos através do aumento da adaptabilidade no nível genético proporcionado por aqueles genes que geram um comportamento cooperativo a indivíduos com genes semelhantes (HAMILTON,1964). O altruísmo recíproco explica o altruísmo dirigido a indivíduos cooperativos, mostrando que indivíduos que cooperam entre si têm maior adaptabilidade evolutiva e, portanto são selecionados (TRIVERS, 1971). A teoria da sinalização custosa almeja explicar porque indivíduos são altruístas mesmo em casos onde a reciprocidade é incerta. Ela afirma que um indivíduo tem muito

a ganhar ao entrar numa relação cooperativa com outro e muito a perder caso fique isolado. Como a reputação de cooperador de um indivíduo é um forte determinante da disposição de outros indivíduos a cooperarem com ele, cada indivíduo tem fortes pressões seletivas de sinalizar cooperação mesmo que isso seja custoso. Os indivíduos engajam no que ficou conhecido como altruísmo competitivo, em que cada um tenta sinalizar níveis de cooperação maiores que o dos outros para que ele seja escolhido como parceiro de cooperações (GILBERT, 1998). Este comportamento foi extensamente estudado e exemplificado em pássaros, que competem para tirar parasitas de outro pássaro – sinalizando assim comportamento cooperativo.

A Teoria da Evolução também ajuda a explicar o surgimento de comportamentos agressivos. Resultados matemáticos em Teoria dos Jogos e Teoria da Evolução mostram que níveis estáveis e duradouros de cooperação só podem emergir caso existam mecanismos punitivos através dos quais o indivíduo trapaceiro é alvo de comportamento agressivo. Podemos definir os agentes trapaceiros e cooperativos desta maneira: os cooperativos partilham seus retornos com o grupo e depois permitem a divisão igualitária destes recursos entre todos os membros; os trapaceiros, apenas retiram uma parcela dos recursos compartilhados sem contribuir. Caso só existam estes dois modelos de agente, mesmo que apareça uma pressão evolutiva para grupos cooperativos, demonstra-se que os retornos de trapaceiros sempre serão maiores que de cooperadores, pois é provado que os retornos de ambos dependem exclusivamente da quantidade compartilhada pelos cooperadores (BOYD, 2003) e não da frequência de trapaceiros. Sendo assim, a adaptabilidade evolutiva do trapaceiro será sempre maior e ele tenderá a dominar o pool fenótipo e a cooperação nunca irá surgir. Porém, caso exista um agente punidor que dirige comportamento agressivo aos trapaceiros, a adaptabilidade e os retornos dos trapaceiros serão inversamente proporcionais ao número de agentes punidores, multiplicado pelo fator da intensidade da punição. Sendo assim, toda vez que existe uma pressão seletiva para grupos cooperativos, ela faz surgir punidores, uma vez que só assim é possível garantir a cooperação. Fica claro, portanto, como o comportamento agressivo pode ser fundamental para estabelecer níveis de cooperação.

2.3.7 A Tragédia dos Comuns

Um dilema em especial será de grande importância neste trabalho, por ser objeto de estudo empírico sendo conduzido neste mestrado: a Tragédia dos Comuns. Como já apresentado, ele aparece em situações nas quais indivíduos partilham de um recurso limitado e podem ou cooperar e usar o recurso de maneira, cautelosa ou esgotá-lo completamente.

Um exemplo clássico é o de pescadores, que têm de pescar até certo limite, pois caso contrário os peixes se esgotariam por completo e não poderiam reproduzir. Caso todos os pescadores pescassem de modo que o total de peixes pescados ficasse abaixo desse limiar, os peixes se reproduziriam e sempre haveria peixes para a próxima temporada. No entanto, se a pescaria exceder o limiar; os peixes não conseguirão se reproduzir e irão se esgotar por completo e ninguém mais poderá pescar.

Julian Savulescu e Ingmar Persson no artigo "Inaptos para o futuro? Natureza humana, progresso cientifico e a necessidade de melhoramento moral" (SAVULESCU & PERSSON, 2011) dedicam uma atenção especial a esse dilema, dada a sua relevância em muitos problemas globais atualmente enfrentados pela humanidade. Eles relembram que nessa situação, caso o grupo de agentes envolvidos for pequeno, cada agente tem uma motivação individualista para cooperar, pois caso ele não coopere os recursos serão esgotados e ele sairá perdendo. No entanto, em grupos maiores, ainda que um único indivíduo não coopere, a cooperação dos restantes provavelmente consegue garantir o não esgotamento dos recursos. Neste caso, assim como no caso de altruísmo competitivo descrito anteriormente, um indivíduo pode evitar trapacear, pois sua não contribuição poderia ser observada pelos outros, afetando negativamente sua reputação e a probabilidade de conseguir parceiros cooperativos no futuro. No entanto, os autores argumentam, em grupos suficientemente grandes a não contribuição de um único membro não seria sequer notada, portanto um indivíduo em particular tem poucas razões para contribuir. Eles observam que seres humanos possuem um forte senso de justiça que os fariam cooperar em situações como essas, caso eles acreditem que todos os outros irão cooperar também. No entanto, em situações de incerteza existe um grande risco de cada indivíduo escolher pela opção não cooperativa. Nesse caso, o problema ultrapassa a contribuição negligencíavel de cada individuo, pois se cada indivíduo raciocinar deste modo, todos irão falhar em cooperar. Os autores mencionam inúmeros problemas globais do mundo moderno que se assemelham ao cenário descrito aqui:

“Problemas ambientais provavelmente fornecem o exemplo mais preocupante das limitações de nossas disposições cooperativas no mundo contemporâneo. Problemas ambientais importantes incluem: mudança climática global, da qual nossas emissões de gases de efeito estufa, como dióxido de carbono e metano, contribuem substancialmente; perda da biodiversidade pela destruição de florestas, zonas úmidas e recifes de corais; desperdício de fontes de energia não renováveis.” (SAVULESCU & PERSSON, 2011 p. 491)

2.4 Inaptos para o futuro?

Ponto crucial para a pesquisa desenvolvida na presente dissertação é a tese central desenvolvida tanto no mencionado artigo de Savulescu e Persson quanto no volume *Inaptos para o futuro: A necessidade de melhoramento moral* (2012) acerca do imperativo de se realizar um melhoramento biotecnológico da moralidade humana. Os autores esboçam no artigo, e desenvolvem extensamente no livro, a tese de que o melhoramento moral humano é uma necessidade inescapável. Eles mostram que as ferramentas atualmente a nossa disposição para resolver o problema da cooperação são ineficientes e que os desafios que enfrentamos nestes campos podem, caso ignorados, levar a nossa extinção. Os autores não defendem que o melhoramento da moralidade humana por meio da tecnologia seja possível, apenas que com as ferramentas disponíveis atualmente, seria impossível lidar com certos problemas. Os autores citam principalmente riscos ligados à ameaça nuclear e ao aquecimento global; estas e outras ameaças são discutidas com maiores detalhes no volume *Riscos Catastróficos Globais* (CIRKOVIC & REES, 2008), no qual diversos especialistas analisam inúmeros riscos à humanidade que poderiam eliminar milhões de vidas humanas ou extinguir por completo a humanidade. A respeito da ameaça nuclear, é relembrado o fato de que os sistemas de detecção de ameaças e retaliação continuam com o mesmo nível de sensibilidade desde a Guerra Fria, e que um erro de detecção ou um ataque terrorista poderia, mesmo nos dias de hoje, desencadear um holocausto nuclear. Uma recalibração dos sistemas de detecção e um desarmamento nuclear é um problema que envolve justamente situações colaborativas permeadas pelos dilemas sociais explorados nas seções anteriores. O aquecimento global também pode se tornar um risco catastrófico na medida em que é possível um inesperado e abrupto aumento da temperatura média muito maior do que temos testemunhado. Ele é também uma situação que envolve uma

coordenação global que em muitos casos se assemelha a situação da Tragédia dos Comuns, onde cada nação deve refrear o seu uso de combustíveis fósseis. Existem ainda muitos outros riscos explorados neste livro, como pandemias e riscos advindos de novas tecnologias como nanotecnologia, inteligência artificial e biotecnologia cujas prevenções e soluções também envolvem problemas de cooperação e dilemas sociais.

Uma pesquisa com especialistas na área revelou que na média eles atribuem uma probabilidade de 19% para a humanidade ser extinta durante o século XXI. Conseguintemente, não é ponto controverso na argumentação de Savulescu e Persson que a não resolução de problemas cooperativos apresenta enorme risco para a humanidade. Mais controversa é a premissa de que a democracia moderna não apresenta os instrumentos para resolver estes problemas cooperativos. Os autores fundamentam esta premissa no fato de que um dos pilares da democracia moderna são a liberdade e privacidade individual. Num mundo onde a tecnologia propicia um poder individual de causar danos cada vez maiores e onde a não cooperação de apenas um pequeno grupo de indivíduos pode pôr em risco a humanidade inteira, o direito à liberdade e privacidade pode tornar impossível atingir os níveis de cooperação necessários para enfrentar os problemas aqui apresentados. Não cabe a essa dissertação entrar no mérito desta segunda premissa, é suficiente apenas que a relevância de se encontrar meios para solucionar o problema da cooperação seja salientada; sejam esses meios parte da democracia moderna ou advindos do melhoramento tecnológico humano.

Um dos meios tecnológicos que os autores consideram que deva ser explorado é o uso do neuro-hormônio ocitocina. Ainda que reconheçam que ele é apenas um primeiro passo na busca por uma solução, eles apontam para as recentes descobertas sobre como a ocitocina influencia níveis de confiança e ligação social entre indivíduos:

> Mas nós pensamos que a enormidade de problemas morais que enfrentamos torna razoável a exploração das possibilidades de tais técnicas (...). Melhoramentos morais biomédicos, caso sejam factíveis, seriam o tipo mais importante de melhoramento biomédico (p. 498)

Ainda sobre a ocitocina, Liao e Roache (2011) afirmam: “a ocitocina parece melhorar a capacidade de ler o estado emocional de outras pessoas, o que é importante para a empatia. Isso sugere que administrar estas substancias em indivíduos poderia nos ajudar a agir conjuntamente para resolver problemas importantes” (p. 247) Como os estudos sobre os efeitos da ocitocina em pessoais saudáveis ainda são escassos, é de

grande importância que a influencia dessa droga nos níveis de cooperação e decisões morais sejam explorados. Quse todos nossos atos são afetados pelos nossos raciocínios a respeito do que devemos ou não fazer, nossa cognição moral. Uma leve alteração generalizada nesse tipo de cognição teria um vasto impacto sobre nossas ações (BOSTROM, 2009. p. 357). Além disso, certos traços da moralidade são considerados universais humanos e parte da identidade de nossa espécie, e sua modificação poderia ter fortes consequências sobre a condição humana (BOSTROM, 2011).

Há algum tempo, um dos únicos usos da ocitocina intranasal em humanos era para induzir a lactação. No entanto, estudos recentes consideram o uso de ocitocina para o tratamento de autismo (ANDARI, 2010). Outros consideram a sua aplicação em indivíduos saudáveis, para melhorar nossas relações afetivas (SANDBERG, 2008) e alguns estudos até demonstraram um papel benéfico da administração de ocitocina na resolução de conflitos entre casais (DITIZEN et al., 2009). Apesar disso, pouco se sabe sobre a vasta gama de efeitos cognitivos da ocitocina.

A ação da ocitocina é mediada pelos receptores dessa droga (os metabotrópicos) que se encontram tanto na periferia do sistema nervoso quanto dentro do encéfalo. A ocitocina é sintetizada nos neurônios magnocelulares dos núcleos paraventricular e supraóticos do hipotálamo e é armazenada em terminais axônicos na neuro-hipófise (SQUIRE et al., 2008). As ações conhecidas da ocitocina se encontram em processos como: lactação, contração uterina, comportamento social, excitação sexual, modulação do eixo Hipotalâmico Pituitário Adrenal (HPA), generosidade, confiança, ligação, afetiva e empatia.

Os estudos prévios realizados com administração exógena de ocitocina intranasal demonstram sua vasta e variada gama de efeitos cognitivos. Inicialmente cabe ressaltar que a administração intranasal de neuropeptídios mostrou-se um método válido de elevar a concentração dos mesmos dentro do encéfalo. Foi encontrado um aumento da concentração de neuropeptídios no fluído cérebro-espinhal após a sua administração intranasal dentro de um período de 80 minutos. A concentração começou a aumentar após 10 minutos e o pico da concentração de ocitocina foi atingido 30 minutos após a administração e foi mantido relativamente estável até 80 min (BORN et al., 2002). Em todos os estudos levantados a seguir, não foram encontrados nenhum efeito colateral quando comparado ao grupo placebo com, portanto sua administração tem-se mostrado

extremamente segura. Além disso, o amplo uso da ocitocina intranasal em mães para estimular a lactação tem-se mostrado seguro.

Em 15 homens saudáveis, um fMRI demonstrou que a ativação da amígdala – região do encéfalo ligada à modulação do medo – era menor em resposta a estímulos visuais amedrontadores no grupo ao qual se havia administrado 27 UI de ocitocina intranasal em oposição ao placebo (KIRSCH et al., 2005). Num estudo com 30 homens saudáveis, 24 UI de ocitocina intranasal aumentou a desempenho em relação ao placebo no teste Reading the Mind in the Eyes Test, um teste frequentemente usado no diagnóstico de autismo em que o indivíduo tem de estimar o estado emocional com base em fotos de olhares. (DOMES et al., 2007). Num estudo com 194 participantes, a administração intranasal de 24 UI de ocitocina gerou um aumento da confiança nos outros participantes durante investimentos monetários (KOSFELD et al., 2005). Numa tarefa na qual os participantes tinham que doar dinheiro para outros, os que foram administrados com 40 UI de ocitocina intranasal se demonstraram 80% mais generosos que o placebo (ZAKET al., 2007).

Existem ao menos mais seis outros estudos relacionando a administração intranasal de ocitocina com outros comportamentos sociais tais como: inveja (SHAMAY-TSOORY et al.,2009), empatia com vítimas (KRUGER et al., 2012), facilidade de se lembrar de rostos felizes (GUASTELLA et al., 2007), facilidade de se lembrar de palavras positivas (UNKELBACH et al., 2008), resolução de conflito em relacionamentos (DITZEN et al., 2009) e precisão ao detectar relacionamentos entre desconhecidos (FISCHER-SHOFTY et al., 2012).

Capítulo III

Problemas em aberto do Melhoramento Moral

3.1 Introdução

Este capítulo irá especificamente versar sobre os riscos de usar a tecnologia para alterar a moral humana, i.e. irei abordar o melhoramento moral, e em particular, os seus riscos. Parte de minha análise será inovadora, pois irei focar nos riscos que podem surgir *mesmo* que os pressupostos admitidos pelos defensores do melhoramento moral sejam verdadeiros e alguns dos contra-argumentos conhecidos sejam inválidos.

Cabe definir mais especificamente o conceito de melhoramento moral dentro do debate geral sobre melhoramento humano discutido no primeiro capítulo deste trabalho. Segundo Douglas (2014), melhoramento moral será definido como qualquer intervenção cuja expectativa é melhorar as tendências morais de um ser humano, em particular aquelas destinadas a melhorar as nossas capacidades de cooperação e de raciocínio moral. Excluirei da definição algumas melhorias convencionais na moralidade ou no raciocínio moral. A educação e um ambiente benéfico estão correlacionados com um melhor raciocínio moral, no entanto – presumivelmente – estes não resolveriam os problemas morais discutidos, nem apresentam sequer riscos graves dignos de investigação filosófica[2].

Cabe relembrar brevemente a discussão do capítulo precedente. Ao longo do século passado, a incapacidade da humanidade para cooperação numa escala internacional tornou-se uma grande preocupação, especialmente quando os problemas do aquecimento global e do desarmamento nuclear surgiram[3]. Persson e Savulescu argumentaram que não estamos equipados com o conjunto certo de traços morais para

[2] Se o melhoramento moral convencional pudesse resolver os problemas de cooperação, então o debate deveria ser sobre educação ou outros métodos conhecidos. Se o melhoramento moral convencional pudesse produzir riscos graves, também o debate deveria ser sobre eles uma vez que já estão em funcionamento. Presumivelmente, um debate sobre o melhoramento moral implica que cada um destes dois antecedentes são falsos.

[3] Ser capaz criar tais tecnologias requer também grandes niveís de cooperação dentro de um grande grupo (por exemplo, um país). Contudo, o seu desenvolvimento é também motivado pela falta de cooperação entre grandes grupos que competem entre si por vezes.

resolver este problema. Podemos cooperar em pequenos grupos (OSTROM, 1994), uma vez que a interação social em pequena escala foi um desafio importante em termos evolutivos. A generosidade, o altruísmo, o senso de justiça e o desejo de punir trapaceiros são expressados facilmente por muitos, permitindo que cooperemos. No entanto, nós não possuímos a habilidade de cooperar corretamente em grupos extremamente grandes, espalhados por países e territórios, e de etnias e origens diferentes (GREENE, 2013, pp. 1-12). Além disso, testemunhamos um aumento do poder destrutivo e a tecnologia está rapidamente sendo globalizada, de modo que a probabilidade de um indivíduo particular ter o poder de destruir toda a humanidade tem aumentado (POSNER, 2004, pp. 71-78). Por isso, os dois autores concluem: temos um imperativo moral para prosseguir o melhoramento moral, não fazê-lo irá expor a humanidade a grandes riscos de catástrofes ou extinção – a que os autores chamam de *mal supremo*.

Não obstante os argumentos acima, poder-se-ia concluir que o melhoramento moral – embora possa ser obrigatório se feito de forma correta – é susceptível de ser feito de forma incorreta implicando ele mesmo o mal supremo. Se for este o caso, poder-se-ia conceder as maiorias das alegações de Persson e Savulescu – por exemplo, que o melhoramento moral, se feito de forma correta, poderia evitar danos extremos no futuro – sem conceder a sua conclusão principal: a de que nós temos a obrigação de perseguir o melhoramento moral.

Como discorrido na introdução desse trabalho, o uso da tecnologia de forma a mudar o nosso ambiente tem sido uma força importante por trás do progresso da humanidade. A tecnologia tem reformulado completamente a nossa relação com o meio ambiente e com nós mesmos. A história da tecnologia revela casos passados nos quais a maioria iria concordar que falhamos em fazer o uso correto e ético dos avanços tecnológicos – por exemplo: guerra nuclear e química. Persson e Savulescu discutiram que muitos desses erros éticos surgiram, em parte, devido às falhas na nossa moralidade (PERSSON & SAVULESCU, 2012). Ademais, recentes melhorias tecnológicas permitem-nos modificar aspectos da nossa própria biologia, incluindo os nossos processos cognitivos e morais. A pesquisa filosófica e científica (GREELY et al., 2008) neste campo emergente, conhecido como melhoramento humano, prosperou. Uma de minhas preocupações é a seguinte: se não estamos adequadamente preparados para lidar com avanços passados, então poderemos estar menos preparados para lidar com

tecnologia (1) com grande potencial de serem destrutivas ou (2) cujo alvo pretendido é a natureza humana (SAVULESCU, MEULEN, KAHANE, 2011, p. 2). Adicionalmente, tecnologias mais recentes estão sendo investigadas como soluções potenciais para esses desafios morais. Essas tecnologias, denominadas melhoramentos morais (DOUGLAS, 2008), podem, por exemplo, aperfeiçoar o nosso raciocínio moral, melhorar a cooperação internacional ou aumentar a nossa empatia para com os mais necessitados. A pesquisa nesta área tem sido também uma resposta de Persson, Savulescu e outros (e.g., PACHOLCZYK, 2011) para a preocupação expressa acima: dado a nossa inaptidão moral para resolver os problemas cooperativos que surgem do desenvolvimento de tecnologias poderosas, podemos apenas sobreviver à introdução de tecnologias ainda mais poderosas e complexas se melhorarmos as nossas disposições morais e capacidade de cooperar (PERSSON & SAVULESCU, 2012).

Se a humanidade desenvolver e distribuir uma droga que dramaticamente aumente a cooperação entre indivíduos, poderá haver dois problemas. O primeiro problema ocorre devido à complexidade das nossas inclinações morais; uma única modificação aparentemente inofensiva pode levar a mudanças graves de forma inesperada. Por exemplo, eu irei discutir a incerteza de que a mencionada droga melhoraria a cooperação entre grupos; em vez disso, poderá melhorar a cooperação dentro do grupo, a um custo de cooperação entre grupos. O segundo problema ocorre devido a mecanismos de autorreforço, já que estes indivíduos são motivados a serem cooperativos, eles poderiam, de forma iterativa, envolverem-se em futuras modificações produzindo um desrespeito crescente por todos os outros aspectos possíveis da moralidade. Uma única iteração pode ser suficiente para se conseguir o resultado desejado, porém esta iteração poderia implicar necessariamente modificações indesejadas no futuro.

O presente capítulo, enquanto admitindo a conveniência do melhoramento moral (PERSSON & SAVULESCU, 2012) e considerando mal sucedidas algumas críticas a ele, tem como objetivo investigar seus riscos a longo prazo que ainda não foram explorados. Se assumirmos que resolver a inaptidão do melhoramento moral na humanidade é obrigatório, que essa modificação é factível, e negarmos que haja algo necessariamente errado em relação a mudar a moralidade humana; então, poderia ainda haver razões para se ser cauteloso no que diz respeito ao melhoramento moral? Este capítulo irá argumentar de forma afirmativa. Assim, eu irei procurar enriquecer o

debate, explorando desafios que devem ser enfrentados até mesmo por quem aceita as suposições dos defensores do melhoramento moral e rejeita algumas de suas críticas.

3.2 Problemas epistêmicos

É possível argumentar que há razão para defender uma forte regra deontológica contra qualquer mudança na natureza humana. Neste caso o debate será definido por posições ou flexíveis ou bastante rígidas da natureza humana. Com esse raciocínio, tem-se então argumentado que a nossa genética (SANDEL, 2002) é uma parte fundamental na natureza humana e, portanto, a seleção de embriões deveria ser proibida. Mas definições da natureza humana tão estritas, que não permitem graus moderados de oscilação genética, iriam excluir grandes populações humanas da própria definição de humanidade (DANIELS, 2009, pp. 25-42). Posições deônticas a respeito da permissibilidade ou obrigação de se fazer o bem terão consequências sobre se certos melhoramentos humanos deveriam ser mandatórios (DOUGLAS & DEVOLDER, 2013). Se apenas os seres que são conscientes podem instanciar valor moral, então caso o melhoramento humano radical produzisse algum futuro vazio de consciência[4], ainda cheio de civilizações complexas e imbricadas, ele consequentemente teria um valor moral esperado igual ao da extinção completa da humanidade. Além disso, quaisquer meta-éticas que descrevem qualquer relação entre tendências morais humanas e éticas normativas[5], são aptas de complicação pelo melhoramento moral, já que o que é normativamente correto teria de acompanhar, até certo ponto, uma mudança descritiva na natureza daquilo que os humanos consideram correto.

Mais importante, há uma preocupação particular que é crucial: o valor moral do futuro. Por exemplo, caso estejamos primordialmente preocupados com os interesses das pessoas vivas hoje e não com os interesses de gerações distantes – como alguns filósofos com visões da ética em termos de afetar as pessoas (*person-affecting views*)[6] –,

[4] "Pode haver uma abundância da riqueza econômica e da capacidade tecnológica num mundo como este, contudo isso seria de nenhum proveito porque não haveria ninguém que beneficiasse disso" (BOSTROM, 2004, p. 344). Este problema pode ser solucionado por mudanças graduais organizacionalmente invariantes, de modo que a consciência seria preservada ao se preservar os mesmo padrões de organização causal (cf. CHALMERS, 2010, pp. 9-10).

[5] Por exemplo, filósofos que vêm o valor como dependente na natureza humana e preocupações como as Thomas Hurka ou Bernard Williams

[6] Uma das mais recentes defesas dessa posição é encontrada na Saturated Harm Mimizing, contrada em MEACHAM (2012).

então, um melhoramento moral com um valor positivo imediato mas com grande valor negativo num futuro longínquo poderia ser considerado como desejado; ao contrário, se considerarmos alguém no futuro como igualmente importante a alguém no presente, devemos focar apenas num futuro longínquo quando avaliando esta tecnologia. Excluindo visões em termos de afetar as pessoas, alguém poderia se comprometer com desconto marginal simples (cf. BECKSTEAD, 2013, pp. 159-62) ou limites temporais (cf. TEMKIN, 2012, pp. 238-62) no futuro. Contudo, irei assumir, concordando com Parfit (1984), Bostrom (2013) e Beckstead (2013), que as posições defensivas em que o valor do futuro não domina a estimativa são difíceis de encontrar. Além disso, qualquer tecnologia que pode ameaçar a nossa existência a logo prazo, aumentando a probabilidade de extinção, merece uma cuidadosa avaliação ética e prática.

3.3. Problemas estruturais: A moralidade é frágil

3.3.1 Os três argumentos de Agar contra o melhoramento moral

Nestas seção, irei primeiramente criticar os argumentos apresentados por Nicholas Agar contra o melhoramento moral. Escolhi os seus argumentos porque percebi que eram uma crítica razoável do melhoramento moral e muito próximos daquela que gostaria de propor.

Os argumentos de Agar são três. Primeiro, a sua tese fundamental é que as experiências humanas, *espécie especifica,* são necessárias para o valor moral, uma posição que ele denomina de relativismo-especista. Ele argumenta que se eliminarmos a possibilidade dessas experiências nos tornando pós-humanos, nós iríamos perder uma condição necessária para o valor moral. No entanto, a concepção de Agar que espécie é um enorme grupo de indivíduos que pode acasalar, reproduzir entre si – creditada a Ernst Mayr – ainda que bastante usada na biologia, é reconhecidamente mal sucedida em casos limítrofes ou artificiais (HANAGE, 2013).

O melhoramento humano é precisamente um caso tanto limítrofe como artificial, por isso subscrever a uma visão moral em que o valor é firmemente amarrado a um conceito tão rígido parece conduzir aos mesmos problemas que este conceito enfrenta na biologia. Além disso, o relativismo-especista significa que o ato de um indivíduo racional e consciente deixar de ser humano devido a valorizar outras experiências seria sempre considerado imoral (RODUIT, 2011). Contudo, parece pelo menos plausível

que para algumas pessoas isto seria uma escolha moralmente válida. Dados esses dois problemas, o relativismo-especista é inadequado para lidar com o melhoramento moral.

Em segundo, Agar afirma que o melhoramento moral irá criar um futuro com duas diferentes classes de pessoas: o mero humano que existe hoje e uma classe de pós-pessoas, cuja capacidade elevada de raciocínio moral irá conferir a eles uma prioridade moral sobre as meras pessoas, que seriam desta forma sacrificadas. Agar considera *a criação* de um cenário onde tal sacrifício é desejável – e não o sacrifício em si – como tendo um valor moral extremamente negativo; e assim conclui que aumentar a capacidade moral, ou o status da moral, não seria desejável. Este argumento, apesar de extremamente relevante para a matéria, parece estar já sendo investigado (AGAR, 2014 e DOUGLAS, 2013).

Em terceiro, ele mantém que a tendência da moral humana está sobre uma delicada balança que o melhoramento moral irá provavelmente perturbar, destruindo a nossa moralidade, a colocando numa configuração indesejada, ao invés de melhorá-la (AGAR, 2013). Ele acredita que a moralidade é baseada tanto em cognições racionais como em intuições emocionais – um equilíbrio ao qual ele denomina de *normalidade moral*. Caso mudemos as nossas tendências inatas para qualquer direção, teremos o risco de um desequilíbrio catastrófico. Este argumento será analisado com mais cuidado na seção seguinte.

3.3.2 Revisitando Agar: a normalidade e fragilidade da moral

Valendo-me da tese da complexidade de valores exposta no primeiro capítulo, defendo que a moralidade humana é suportada por um frágil equilíbrio não das duas tendências que Agar menciona, mas por uma super-abundância de tendências. De acordo com a tese da complexidade de valores (BOSTROM, 2014, pp. 380-384 e YUDKOWSKY, 2011, p. 388-393)[7]: (1) as tendências morais humanas são extremamente complexas e entender o que nós valorizamos e desejamos enfrenta grandes dificuldades epistêmicas; e (2) qualquer pequena mudança num pequeno

[7] Outros eticistas que acreditam que há uma pluralidade de valores incomensuráveis, datando de Isaiah Berlin até Thomas Nagel e Larry Temkin, iriam, talvez, também defender uma tese similar. No entanto, a tese da complexidade de valores foi particularmente construída para tratar de questões que surgiram das novas tecnologias.

subconjunto da nossa moralidade poderia significar numa mudança drástica do *todo* em formas inaceitáveis.[8]

Irei apresentar duas maneiras pelas quais essa fragilidade se evidencia. A primeira será esboçada abaixo como o problema da natureza auto-reforçadora e irreversível de uma modificação na moralidade, valores morais sendo vistos como intimamente ligados à motivação. A segunda será estudada num capítulo a parte, o capitulo IV, e lida com o fato de que a cooperação – um dos alvos mais cobiçados do melhoramento moral – é complexa e frágil.

3.3.2.1. Melhoramento moral: Auto-reforçador e irreversível

Podemos realizar qualquer quantidade ilimitada de melhoramentos sem tornar a humanidade algo completamente irreconhecível? Poderia o melhoramento moral radical, eventualmente, levar à extinção da humanidade? Se, inadvertidamente, mudássemos os nossos valores de maneiras indesejadas, poderíamos reverter esse erro?

A primeira questão a ser investigada para responder a essas perguntas é se o grau da herdabilidade de valores ou moralidade, enquanto realizando múltiplas iterações de melhoramento moral, é aceitável.

Indiscutivelmente, se a melhoria for suficientemente moderada, então o resultado de uma única iteração será moralmente desejado. Adotando o utilitarismo como a teoria moral correta, supnhamos que mapeemos corretamente as estruturas neurológicas e as projeções neuroquímicas relacionadas com o comportamento utilitarista e que desenvolvamos uma droga que aumente a motivação utilitária. Com estas suposições, a aplicação generalizada desta droga uma única vez levará provavelmente ao melhoramento moral. O resultado disso seria a produção de uma humanidade com um conjunto diferentes de objetivos, valores e motivações. Se fôssemos mais utilitaristas, estaríamos mais susceptíveis a estarmos propensos a querer desenvolver e a tomar novas drogas que nos tornassem mais utilitaristas ainda, o que, por sua vez, iria nos fazer desejar esses valores e ter esses objetivos num grau ainda maior. Eventualmente, um grande número de iterações iria produzir indivíduos que

[8] Por exemplo, a tentativa de aumentar as tendências morais no sentido de aumentar a felicidade dos outros pode fazer com que alguém procure formas artificiais de induzir felicidade, como com a simulação direta do cérebro ou o uso de drogas, ignorando outras dimensões potencialmente importantes como a veracidade e a variedade de experiências.

seriam considerados moralmente indesejados por aqueles primeiros envolvidos no processo de melhoramento; os indivíduos iniciais poderiam até se sentir moralmente enojados em relação a eles, vendo-os como alienígenas e criaturas imorais, produtos de demasiadas modificações radicais. No entanto, se é provável que a operação do melhoramento irá transformar os indivíduos de forma a fazer com que eles queiram e valorizem coisas diferentes, a partir daí querendo aumentar mais esses valores até finalmente tornarem-se imorais da perspectiva dos indivíduos iniciais, então porque seria moralmente desejável embarcar na primeira iteração de aprimoramento? Além disso, o resultado mais desejável poderia ser aquele que se encontraria numa etapa intermediária, mas não seríamos capazes de parar nela, uma vez que ela implicaria mais iterações.

Contrariamente, tal raciocínio poderia ser um exemplo falacioso de um Argumento Derrapante (*Slippery Slope Argument)*. Tais argumentos costumam concluir que um curso presente de ação – considerado agora desejado – é errado, pois é susceptível de produzir uma linha inevitável de efeitos que levam a um futuro com consequências indesejadas e inesperadas. Este raciocínio é frequentemente considerando falacioso (DOUGLAS, 2010 e VOLOKH, 2003). No entanto, deve-se notar que a minha principal preocupação é em relação ao grau de valor da herdabilidade, não apenas dos resultados finais; portanto, não constitui, necessariamente, um caso em que se poderia aplicar um Argumento Derrapante. Além disso, nem todos os casos de Argumentos Derrapantes são considerados falaciosos; devido aos fortes aspectos motivacionais de auto-reforço em jogo no melhoramento moral, o uso de tal argumento poderia ser válido.

Ademais, o melhoramento moral é provavelmente irreversível. Em várias teorias de ética descritiva, a moralidade humana corresponde aos desejos de segunda ordem que moldam a motivação (SMITH, LEWIS & JOHNSTON, 1989), que por sua vez gera este comportamento; em outras teorias a moralidade é, pelo menos, fortemente ligada à motivação. Como discutido na seção 1.4.3.1 do primeiro capítulo, se conduzirmos esta estrutura motivacional num certo caminho, isto pode criar uma cadeia de auto-reforço de motivação que iria funcionar para manter tal estrutura constante, fazendo com que fosse irreversível. A nova estrutura de valor – atribuindo menos valor à estrutura anterior – iria, naturalmente, opor a sua reversão.

Concluindo-se, uma solução trivial seria evitar a modificação da moralidade, ou usar formas convencionais de melhoramento moral em consonância com a educação. São necessárias ainda investigações maiores para que possamos concluir se poderia haver uma solução entre a trivialidade e uma moralidade humana recursivamente instável.

Capítulo IV

Modelando a difusão dos melhoramentos morais: estratégias sociais vendidas no balcão

4.1 Introdução

Neste capítulo irei defender que alterar fundamentalmente as preferências sociais inatas, ainda que numa tentativa de melhorá-las, pode resultar em algumas consequências catastróficas. Alterar o modo com o qual cooperamos e formamos relações é visto como um imperativo moral, ou ao menos, moralmente desejável. Pesquisas recentes na área da neurociência da moralidade, cooperação e ligação afetiva indicam que isso pode ser factível num futuro próximo. Portanto, é razoável assumir que estas tecnologias podem vir a ser desenvolvidas e usadas. No entanto, não existe nenhum modelo de difusão para nos auxiliar na previsão dos impactos sociais de tais tecnologias. Neste capítulo, pretende-se construir um primeiro modelo de difusão e, ao fazê-lo, severos riscos de se modificar ingenuamente as preferências sociais humanas ficarão aparentes. Suponha que desenvolvamos drogas que de fato modifiquem fundamentalmente estas preferências. Num mercado livre, indivíduos poderiam comprar uma "pílula da cooperação", uma "pílula do individualismo'', uma "pílula do altruísmo" e assim por diante. Eles estariam escolhendo estratégias sociais inatas da mesma maneira que agora nós compramos outros bens valiosos, como um carro ou casa. Evidentemente, uma pílula individualista teria efeitos sociais indesejáveis e deveria ser regulada. No entanto, ao modelar a difusão de tal pílula, será mostrado que a situação é surpreendentemente mais severa. Caso exista qualquer tipo de atraso entre o desenvolvimento e o banimento de tais tecnologias, então ela irá difundir-se pela população, produzindo efeitos socialmente disruptivos. Se o modelo desenvolvido aqui é confiável, o momento correto para banir a pílula individualista é agora: antes que ela seja desenvolvida. Ademais, ainda que possa aparentar que uma droga cooperativa seja um melhoramento moral claro, aumentar os níveis cooperativos individuais pode não necessariamente levar a um aumento da cooperação de larga escala. Existe uma

interconexão complexa entre as camadas sociais, e um aumento da cooperação *entre indivíduos* poderia paradoxalmente emergir como uma capacidade diminuída de cooperar *entre grupos*. Também será discutido como a existência de uma moralidade melhor e passível de ser produzida com intervenção da tecnologia não necessariamente implica que é seguro nos melhorarmos naquela direção, dado a existência de ótimos locais.

4.2. Comprando preferências sociais: um primeiro modelo de difusão

4.2.1 Pressupostos básicos

Alguns pressupostos básicos relativamente intuitivos serão assumidos. Eles advêm da ciência da difusão de inovações (1 e 2; ROGERS, 1983), teoria evolutiva da cooperação (3 e 4; WEST, GRIFFIN & GARDNER, 2007 e BOYD, 2003) e da psicologia social experimental (5; VAN LANGE, 2013).

1. **Inovadores**: Dada a introdução de uma nova tecnologia potencialmente benéfica, ao menos um diminuto grupo de tomadores de risco e inovadores irão fazer uso da mesma.

2. **Imitação**: A probabilidade de um agente comprar a "pílula para estratégia social x" será uma função de quão lucrativa é a estratégia social x, que por sua vez é uma função de quão bem outros agentes com a estratégia social x atuam, que por fim é uma função das frequências de cada estratégia.[9]

3. **Punição para não cooperadores**: Adeptos da pílula individualista ou competitiva serão eventualmente punidos como trapaceiros. O Governo – ou outra estrutura cooperativa – tentará banir drogas individualistas ou competitivas, e fomentar as cooperativas ou altruístas. A sua habilidade de fazer

[9] Em modelos de difusão simples, o fator de imitação é o fator decisivo que determina a difusão. No nosso modelo, as estruturas de punição e propriedades de alto nível serão também muito importantes. No entanto, uma avaliação correta do fator imitação ainda é extremamente relevante. Isto será feito através de estudos empíricos, profundamente necessários sobre, como as melhorias são percebidas. Sem dúvida, e com razão, alguns afirmam que a perceção negativa de melhoramentos cognitivos podem levar à experiência de ostracismo, um fator análogo que pode ser aplicado para o nosso caso e que iria reduzir drasticamente a propagação através da diminuição do fator de imitação. Além disso, o adotante pode ser infligido com custos adicionais de natureza psicológica, devido a ser percebido como um trapaceiro, ou de possuir a necessidade usar um melhoramento por não ter capacidade o suficiente ou como tendo almejando deixar parcialmente d condição humana (FAULMÜLLER et al., 2014). Devem ser igualmente tidos em conta, com possíveis punições ou possíveis incentivos; na medida em que, as percepções potencialmente positivas, também devem ser.

qualquer um dos dois é uma função de sua força. Punição é o aspecto fundamental garantindo que agentes não-cooperativos estejam em desvantagem. Uma vez que os cooperadores não possam mais punir não-cooperadores, eles sempre terão um retorno menor.

4. **Punição requer cooperação**: A força do governo, ou estrutura análoga, será uma função das frequências das estratégias. Alta cooperação e/ou baixo individualismo fazem o governo mais forte. Baixa cooperação e/ou alto individualismo fazem o governo mais fraco.

5. **Orientações sócio-valorativas**: Preferências sociais ou estratégias correspondem a *tendências individuais* a uma certa configuração específica de desfechos, conforme já discutido. Aqui será assumido que essas preferências podem ser corretamente mapeadas da seguinte maneira computacionalmente tratável:

Orientação	**Tipo de desfecho buscado**
Altruísmo	Maximização dos retornos do outro (MaxOther)
Cooperação	Maximização da soma dos retornos (MaxJoint)
Individual	Maximiza o próprio retorno (MaxOwn)
Igualdade	Minimização da diferença entre os retornos (MinDiff)
Competição	Maximização da diferença entre os retornos (MaxDiff)
Agressão	Minimização dos retornos do outro (MinOther)

Tabela 4.2.1

Cada agente irá ter uma orientação global igual à soma ponderada de todas as preferências, levando em consideração o peso que o agente fornece a cada uma delas.

4.2.2 Como escolher como se escolhe?

Ao contrário da difusão de inovações, os nossos agentes não têm preferências estáveis; precisamente, são eles que escolhem essas preferências. Diferente da evolução da cooperação, a força de seleção não advém da lenta, milenar e estável pressão evolutiva. Finalmente, embora o modelo de orientações sócio-valorativas da psicologia experimental social muitas vezes possua uma certa flexibilidade, estamos – *arguendo* – supondo que as drogas aumentem, em grande parte, a liberdade de escolher voluntariamente essas orientações. Aqui temos um problema incomum, de liberdade sobre o próprio processo de decisão. Precisamos de uma solução matemática para escolher como os agentes escolherão estratégias; estratégias, que, por sua vez, determinam como eles vão fazer escolhas no futuro.

Nós podemos resolver isto como se segue. Primeiro, observamos que uma orientação específica pode ser descrita como uma função de sub-utilidade, com base no resultado desejado. Em segundo lugar, cada agente individual terá pesos diferentes para cada orientação. Por exemplo, um agente 50% cooperativo e 50% individualista terá uma utilidade total igual à média entre *MaxOther* e *MaxOwn*. Ao invés de simplesmente igualar a utilidade total do agente com o retorno individual, a utilidade será a soma ponderada de cada sub-função, e cada sub-função será, por sua vez, uma função do retorno individual e do retorno dos outros agentes.

Tomemos o_α como o pagamento do agente em questão e o_x o pagamento do agente x; $\bar{o}$ representa o n-tupla $(o_1, o_2, .., o_n)$ com todos os pagamentos de todos os outros agentes e $\bar{p}$ o conjunto completo de pagamentos $(o_\alpha, \bar{o})$, todos definidos nos números reais. $\mathbf{U}_\alpha(A)$ será a função utilidade completa do agente em questão, dado o resultado A. Finalmente, teremos $z\mathbf{Alt}(\bar{p})$ correspondente à função de utilidade sub-altruísta para pagamentos $\bar{p}$ ponderada por *z*. As demais sub-utilidades serão nomeadas de acordo, e todos as utilidades pertencem ao intervalo aberto]-1,1[.

Temos então que:

$$\mathbf{U}_\alpha(A) = z\mathbf{Alt}(\bar{o}) + x\mathbf{Coo}(\bar{p}) + y\mathbf{Ind}(o_\alpha) + w\mathbf{Equ}(\bar{p}) + q\mathbf{Com}(\bar{p}) + j\mathbf{Agg}(\bar{o})$$

As sub-funções de utilidade podem ser construídas do seguinte modo, modelando matematicamente o que foi intuitivamente exposto:

$$\lim_{\frac{\sum_n \overline{o}}{n} \to \infty} Alt(\overline{o}) = 1 \therefore \lim_{\frac{\sum_n \overline{o}}{n} \to -\infty} Alt(\overline{o}) = -1$$

$$\lim_{\frac{\sum_n \overline{p}}{n+1} \to \infty} Coop(\overline{p}) = 1 \therefore \lim_{\frac{\sum_n \overline{p}}{n+1} \to -\infty} Coo(\overline{p}) = -1$$

$$\lim_{o_\alpha \to \infty} Ind(o_\alpha) = 1 \therefore \lim_{o_\alpha \to -\infty} Ind(o_\alpha) = -1$$

$$\lim_{(\sum_n \overline{o}) - (n \times o_\alpha) \to 0} Equ(\overline{p}) = 1 \therefore \lim_{|(\sum_n \overline{o}) - (n \times o_\alpha)| \to \infty} Equ(\overline{p}) = -1$$

$$\lim_{(n \times o_\alpha) - (\sum_n \overline{o}) \to \infty} Com(\overline{p}) = 1 \therefore \lim_{(\sum_n \overline{o}) - (n \times o_\alpha) \to \infty} Com(\overline{p}) = -1$$

$$\lim_{\frac{\sum_n \overline{o}}{n} \to -\infty} Agg(\overline{o}) = 1 \therefore \lim_{\frac{\sum_n \overline{o}}{n} \to \infty} Agg(\overline{o}) = -1$$

Equações 4.2.2: Sub-utilidades de cada orientação sócio-valorativa

Em seguida, pode-se definir que os agentes irão encontrar jogos gerados aleatoriamente e receberão pontuações com base em sua própria estimativa de utilidade, de acordo com as funções acima, para os resultados do jogo. Em seguida cada agente vai atualizar os pesos em cada sub-função utilidade com base na sua utilidade e na de outros agentes. Uma maneira simples de atualização tenta imitar o agente nas proximidades com a pontuação mais alta, comportamento comum na teoria dos jogos evolucionários. No entanto, isso seria negligenciar o fato de que os agentes possuem funções utilidade diferentes e portanto imitar o agente com a pontuação mais alta não necessariamente levara esse último a ter uma pontuação elevada segundo sua própria função. Uma maneira melhor para atualizar é se um determinado agente imita o agente nas proximidades, que segundo a função utilidade deste determinado agente, teve a melhor pontuação. Provavelmente, situações reais vão conter misturas de ambas as soluções. Seres humanos muitas vezes avaliam os resultados de acordo com os seus próprios padrões, mas eles também almejam sucesso isoladamente, sem levar em consideração o critério deste sucesso.

4.3. Simulação: Resultados

Usando-se os princípios estabelecidos nas seções anteriores podemos contruir uma simulação rudimentar para modelar a difusão das variadas drogas. O Anexo A deste trabalho contém o código fonte da simulação feita com o programa MatLab 2013 para modelar o comportamento destes agentes. Ele foi escrito pelo Dr. Anders

Sadnberg, que prestou assistencia técnica para realização deste trabalho. Tal código tenta modelar computacionalmente os aspectos teóricos discutidos na seção anterior. Uma vez que se trata de uma simulação ainda rudimentar, a qual pretendo melhorar durante meu trabalho de doutorado, alguns aspectos não serão modelados. Não estará presente a punição para não-cooperadores e incentivo a cooperadores. A imitação também será simples, meramente copiando-se o comportamento do agente de maior sucesso.

No inicio da simulação temos 150 agentes iniciais que possuem orientações sociais aleatórias, i.e. os pesos que sua função utilidade fornece a cada sub-função são aleotórios. Estes agentes então entram em 100 rodadas de jogos ou interações aleatórias entre si, nas quais escolhem de acordo com suas orientações. A cada rodada um agente irá mudar os pesos da cada orientação conforme o sucesso do agente vizinho, tentando copiar inclinações que tiveram alto sucesso.

O gráfico a seguir ilustra uma população no inicio da simulação. No eixo vertical temos a porcentagem de peso que um agente dá a cada uma das orientações, no eixo horizontal temos as orientações sendo:

1: Altruísta

2: Cooperativo

3. Individualista

4: Igualitário

5: Competitivo

6: Agressivo

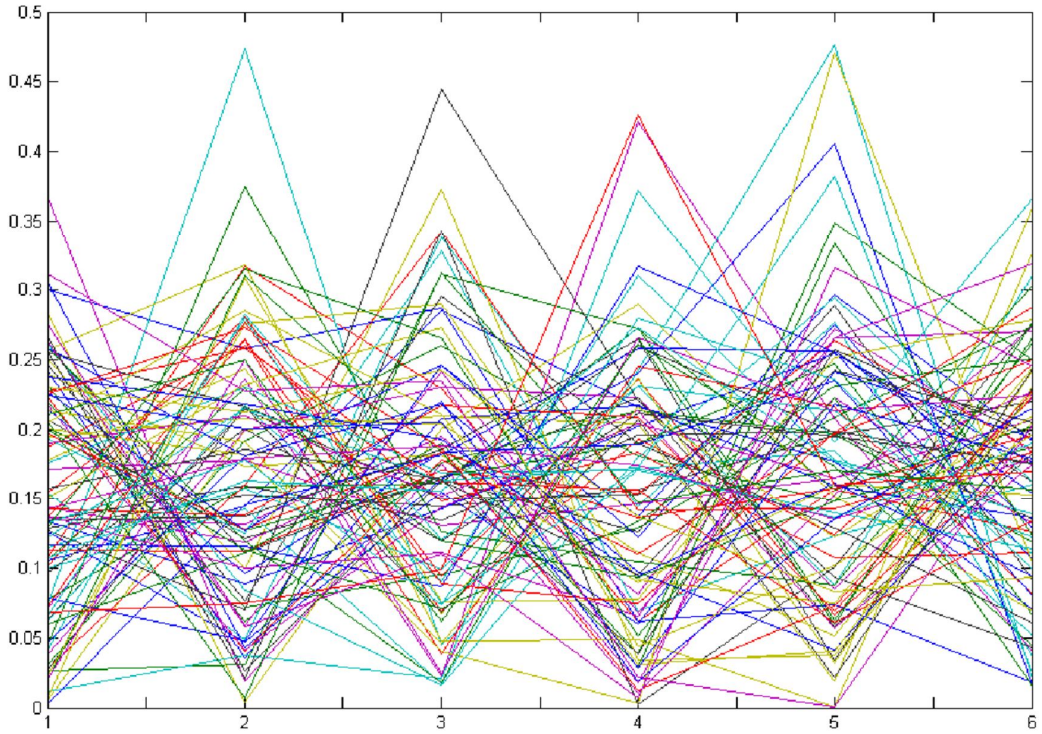

Figura 4.3.1: Exemplo de distribuição inicial das pesagens de cada orientação.

Cada linha do gráfico representa uma sub-população que possui uma determinada pesagem entre as várias funções, os diferentes grupos cores sendo meramente ilustrativos. Como podemos ver, as pesagens estão distribuídas aleatoriamente. Com o prosseguimento da simulação foi possível notar que todas com o tempo tendiam a um atrator de estabilidade em que todas as sub-populações de agentes tinham mais ou menos a mesma pesagem, o que acredita-se ser fruto do alto fator de imitação presente na simulação. A pesagem específica que todos os agentes tendiam a ter variou de simulação para simulação, mas uma configuração muito comum foi a abaixo:

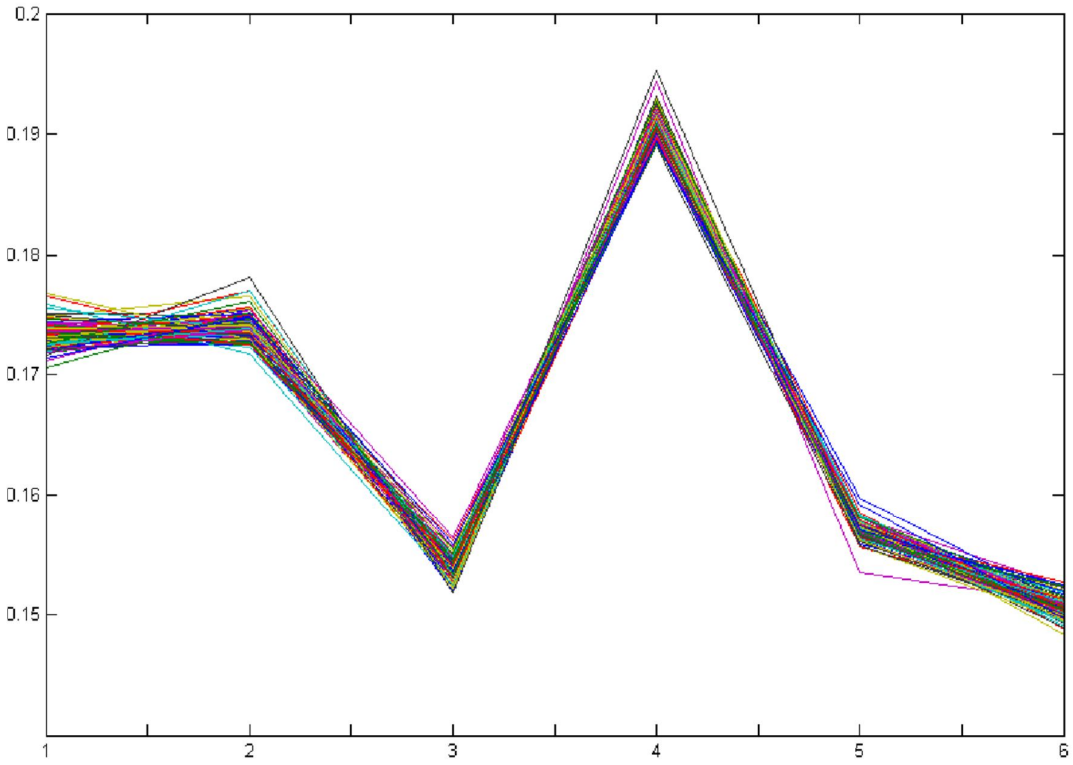

Figura 4.3.2: Exemplo de distribuição final das pesagens de cada orientação.

Isto é, ao final da simulação os agentes possuíam em sua maioria uma distribuição de orientações muito próximas de: 17% Altruísta, 17% Cooperativo, 15% Individualista, 20% Igualitário, 16% Competitivo e 15% Agressiva. Aqui temos o exemplo de outra configuração final mais incomum:

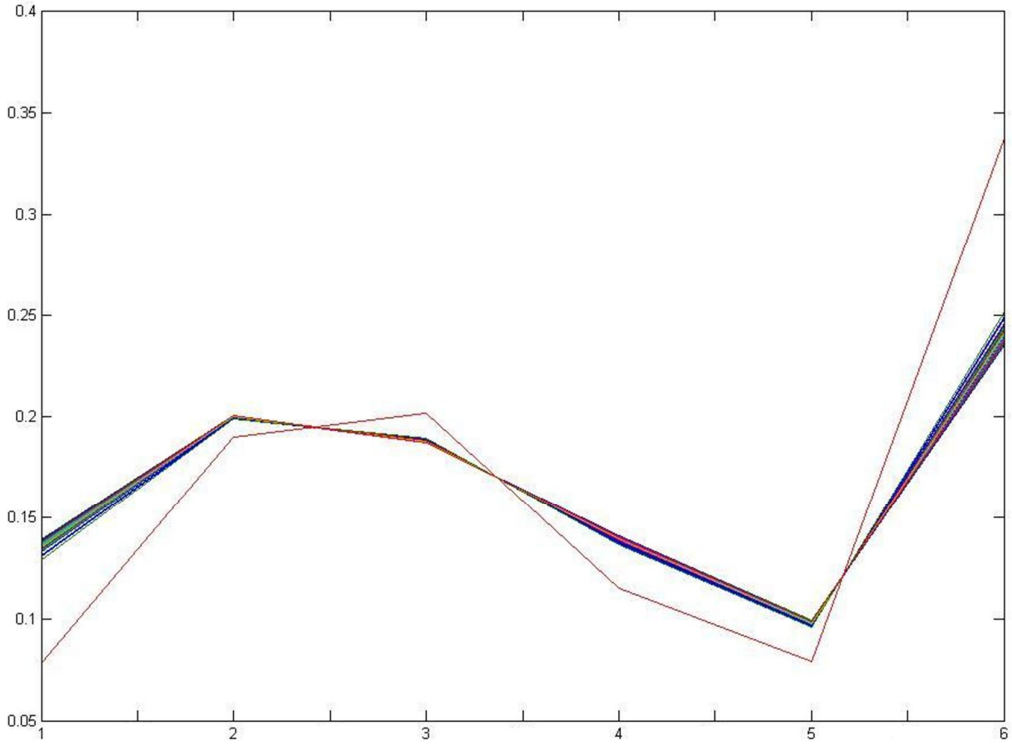

Figura 4.3.3: Exemplo de distribuição final das pesagens de cada orientação

Neste caso os agentes possuíam em sua maioria uma distribuição de orientações muito próximas de: 15% Altruísta, 20% Cooperativo, 20% Individualista, 15% Igualitário, 10% Competitivo e 25% Agressiva. Aqui a população como um todo é comparativamente menos cooperativa e muito mais agressiva que a anterior. Note-se que neste caso temos uma sub-população ligeiramente desviante que alcança níveis próximos a 35% de agressividade e de apenas 7% de Altruísmo.

No entanto, a informação mais relevante para os propósitos deste capitulo se encontra no gráfico a seguir.

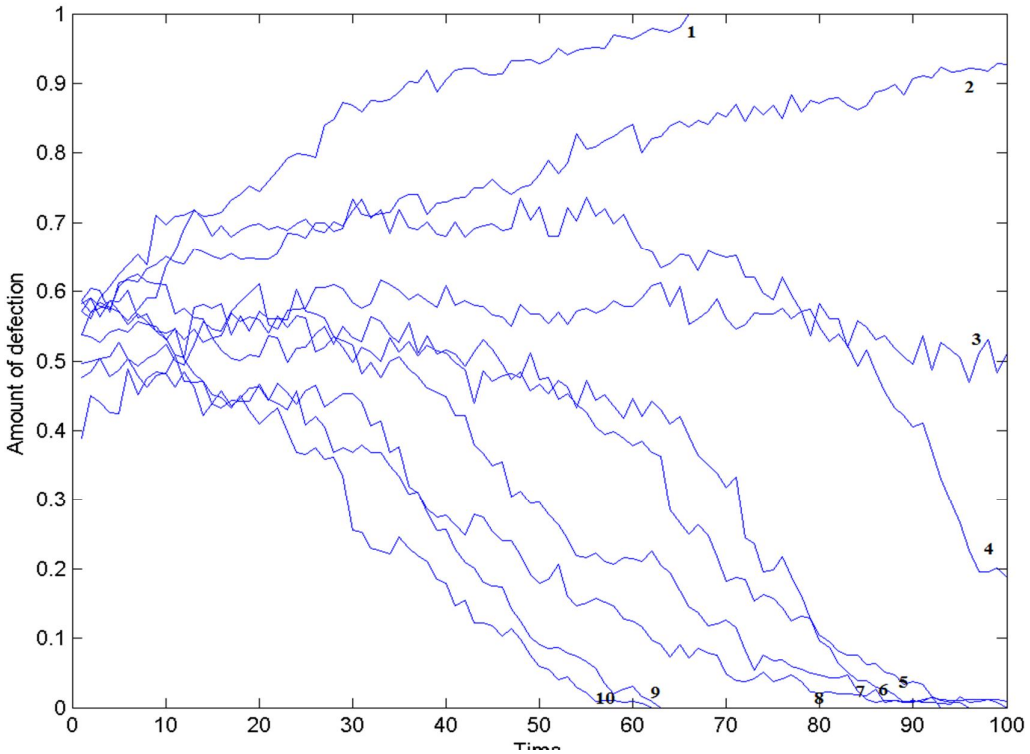

Figura 4.3.4: Distribuição das interações não-cooperativas ao curso de 100 interações, em 10 simulações diferentes.

Aqui temos no eixo vertical a porcentagem total de interações não-cooperativas entre os agentes na população e no eixo horizontal temos o número de rodadas da simulação. Cada uma das linhas numeradas de 1 a 10 constituem simulações diferentes, cada uma delas com 100 rodadas. Estas simulações possuem algumas particularidades, diferente de uma simulação mais geral em que os agentes podem escolher – "comprar" – todas as estrategias disponíveis. Nas simulações 1 a 2 os agentes poderiam apenas escolher se tornar mais individualistas – comprar a pilula individualista – ou não se modificar. No caso das simulações 3 a 8 os agentes poderiam escolher livremente suas orientações. Nas simulações 9 e 10 os agentes poderiam apenas escolher se tornarem mais cooperativos ou não se modificarem. Podemos observar portanto que disponibilizar a opção de tornar-se mais individualista aumentou de maneira geral o individualismo e diminui a cooperação, em alguns casos chegando-se a 100% de não cooperadores (Simulação 1). É possível notar que quando eles puderem escolher entre quaisquer uma das estratégias em alguns casos a cooperação diminuiu, mas na maioria

dos casos aumentou. Quando os agentes poderiam se tornar apenas mais cooperativos o nível de não cooperadores tendeu a zero.

Caso os agentes possam escolher apenas entre altruísmo e individualismo, a simulação tende a se manter estável na sua configuração inicial, como podemos observar no gráfico das simulações a seguir:

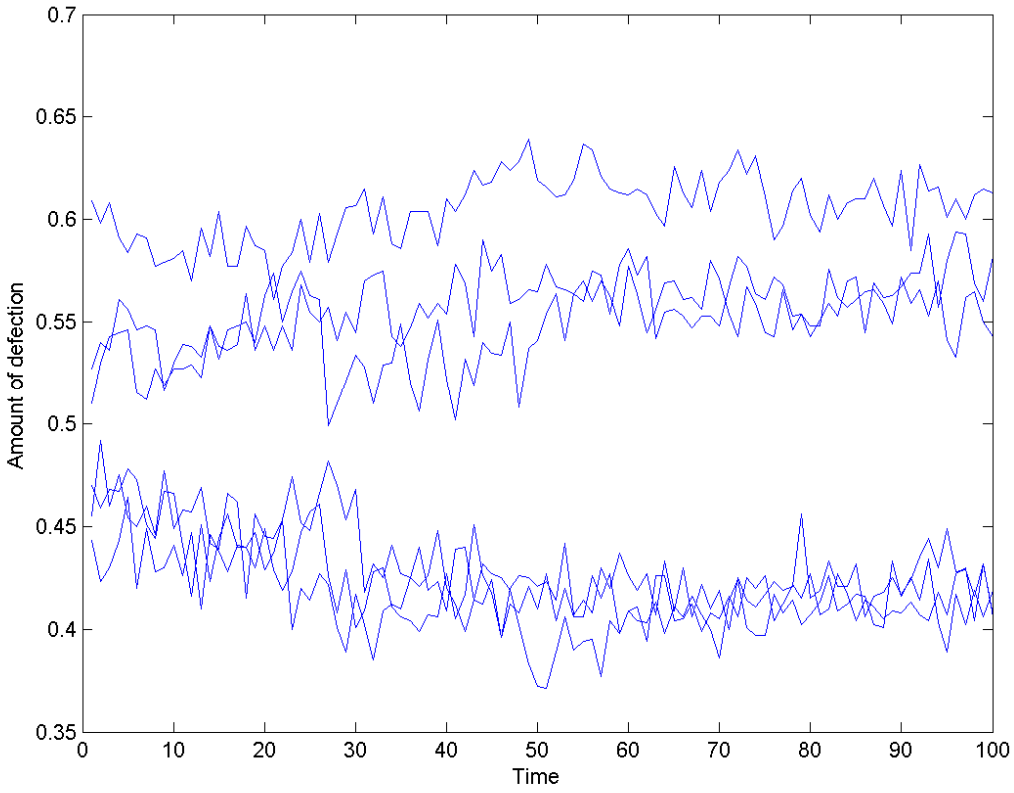

Figura 4.3.5: Distribuição das interações não-cooperativas ao curso de 100 iterações, em 10 simulações diferentes.

Disto podemos tirar três conclusões:

1) Caso se disponibilize apenas uma pilula individualista, a sociedade se tornará rapidamente menos cooperativa, podendo se chegar a níveis de 100% de individualismo.

2) Caso se disponibilize apenas uma pílula cooperativa a sociedade tende, na maioria parte dos casos, a se tornar mais cooperativa.

3) Caso se disponibilize todas as opções, a situações podem variar.

No caso 1 teríamos supostamente um cenário de caos e desmantelamento das instituições sociais, uma vez que seria impossível garantir a cooperação continuada dos indivíduos da sociedade. Uma vez que na Simulação 1 tivemos um aumento grande e rápido do individualismo, e como após o desmantelamento das instituições sociais seria assumidamente impossível restabelecer estruturas punitivas para não-cooperadores, parece prudente assumir que o desenvolvimento de uma pílula individualista deve ser barrado *ex ante*, antes mesmo de ser iniciado, uma vez que seus efeitos poderiam ser rápidos, irreversíveis e catastróficos.

Ficaríamos tentados a concluir que cenários em que a cooperação geral aumentou fortemente seriam desejáveis, e que portanto deveríamos – ao contrário do caso da pílula individualista – incentivar fortemente o desenvolvimento de uma pílula cooperativa. A seção a seguir irá apresentar argumentos contra essa conclusão intuitiva.

4.4. Problemas no paraíso do melhoramento moral, um demônio chamado emergência

Embora um paraíso cooperativo possa parecer sugestivo, não está claro que um aumento nos níveis de cooperação dos agentes individuais vá aumentar a cooperação global na sociedade. Ainda que se possa esperar que um aumento na nossa *tendência individual* para a cooperação *entre os indivíduos* implicaria em uma maior cooperação *entre os grupos*, deve ficar claro que o último é o mais desejável e o objetivo do melhoramento moral. Assim, a pergunta é: aumentos de disposições morais favoráveis à cooperação *entre os indivíduos* irão necessariamente promover a cooperação *entre grupos*? Irei argumentar que a resposta é não. Não só não há conexão inerente como existem mecanismos plausíveis pelos quais a cooperação entre os indivíduos na realidade diminui a cooperação entre os grupos.

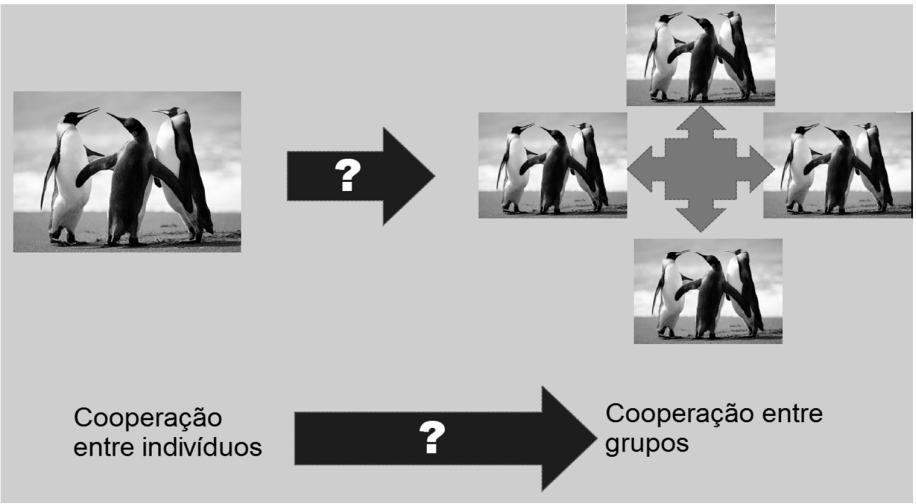

Figura 4.3.6: Relação entre cooperação entre individuos e cooperação entre grupos.

Se tivermos em mente muitos exemplos bem compreendidos de propriedades emergentes, tal resposta negativa deve vir sem nenhuma surpresa. Propriedades emergentes são padrões ou organizações que surgem de propriedades ou interações mais simples. Um exemplo clássico são os flocos de neve com padrões simétricos decorrentes de gotículas de água arrefecidas. A organização emergente muitas vezes pode possuir características inesperadas. Isso é verdade mesmo para os processos físicos extremamente simples. Faça um prato com água ser aquecida por baixo. Um fluxo natural de aquecimento por convecção irá ocorrer e o processo que rege a condução de calor em nível microscópio seguirá um movimento aleatório desordenado. No entanto, em certas configurações relativamente simples, estruturas macroscópicas hexagonais, ordenadas e estáveis se tornam visiveis na superfice. Caso diminuamos a altura da panela os hexágonos transformam-se em espirais; se aumentarmos muito a temperatura, os padrões desaparecem no caos. Este fênomeno é conhecido como células de Rayleigh-Bérnard.

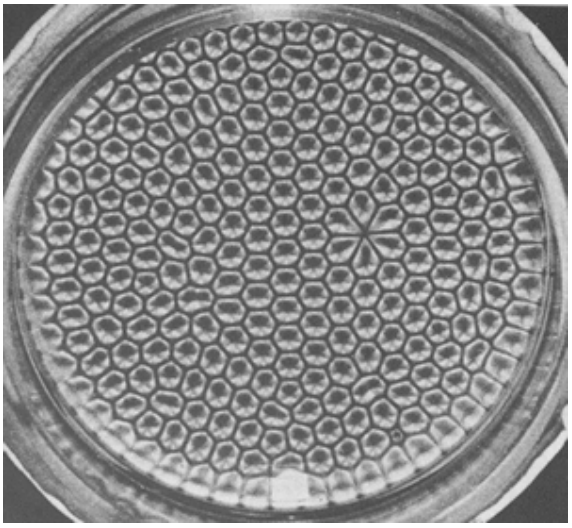

Figura 4.3.7: Fenomeno físico de emergência, células de Rayleigh-Bénard

Podemos perceber que as formigas teriam que exibir individualmente um conjunto de normas sociais complexas para que a organização grande e complicada de suas colônias fosse possível. No entanto, as formigas individuais na verdade apresentam um comportamento quase aleatório e caótico. Caso alguém conhecesse apenas uma formiga individual, tal criatura boba e caótica pareceria incapaz de qualquer organização complexa. No entanto, quando colocamos muitas dessas formigas juntas, padrões emergem e ocorre uma sincronização complexa que dá origem à colônia de formigas, que pode ser vista quase como uma única criatura macroscópica. Pertinentemente, se mexermos com o comportamento aparentemente assistemático da formiga, o padrão macroscópico da colônia pode mudar de forma inesperada e dramática.

A seguir, vou descrever três mecanismos plausíveis, em que o aumento da cooperação indivídual poderia levar a efeitos inesperados na organização do grupo, ou seja, uma diminuição da cooperação entre os grupos.

A tabela de orientações sócio-valorativas tenta justamente encapsular padrões sociais geralmente emergentes no nível individual, no entanto, ele pode não fazê-lo suficientemente bem, permitindo a existência de casos nos quais o comportamento individual vai em uma direção e o comportamento de um grupo em outro. Passo agora a explorar várias maneiras plausíveis em que aumentos no nível mais baixo, o de tendências individuais para a cooperação entre os indivíduos, poderiam emergir como

uma diminuição no nível mais elevado, o de tendências de grupo para a cooperação entre grupos.

1. **Paroquialismo**: O exemplo mais clássico são os casos em que uma maior cooperação dentro de um grupo leva à diminuição da cooperação e até mesmo agressividade entre os grupos. Cada indivíduo valoriza o seu próprio grupo, de tal forma que, às vezes, ele pode prejudicar a si mesmo e os outros se tiver a crença de que ele vai beneficiar o seu próprio grupo ao fazer uma ação. Isto é particularmente preocupante, uma vez que, um aspirante para o aprimoramento moral, a oxitocina, é conhecida por produzir tais efeitos, levando a etno-centrismo e paroquialismo (DE DREU et al., 2011a, 2011b).

2. **Mal necessário**: A forma como uma econômia moderna está organizada depende muito de agentes individuais moderadamente auto-interessados. Modelos e políticas macroeconômicas muitas vezes dependem de tais pressupostos, e construimos nossas estruturas cooperacionais de nível superior sobre um nível inferior individualista. Sociedades capitalistas ocidentais são particularmente conhecidas por depender do individualismo. Poderia ser o caso que conseguiríamos uma maior cooperação se todos fossem completamente cooperativos. Ainda sim, se o caminho que conduz a estes níveis mais elevados de cooperação implica diminuir o individualismo numa sociedade dependente dele, então talvez seja o caso que não podemos alcançá-los sem correr o risco de desestabilização grave.

3. **Líderes**: A forma como a política é organizada também depende do individualismo. Uma característica comum das sociedades cooperativas coesas é a presença de líderes. Muitas vezes, a única forma viável de restringir desejos, valores, opiniões e posições em um grupo cooperativo é delegar responsabilidade a alguns poucos indivíduos. Mas ao chegarmos próximo de extinguir o individualismo, pode ser o caso que ninguém nunca vai querer se destacar e se tornar um líder. Mais uma vez, é plausível que tais sociedades altamente cooperativas não precisam de um líder. Ainda sim, se houver apenas uma única iteração no processo de aumento cooperativo que leva a uma sociedade suficientemente cooperativa para extinguir os grandes líderes, mas ainda individualista o suficiente para tornar a cooperação sem líderes inviável, a não-cooperação viria à tona.

4. **Polarização**: Os defensores do melhoramento moral poderiam almejar melhorar a nossa moralidade primitiva tentando apagar o pensamento político polarizado. Alguns veem certos debates políticos como os exemplos mais evidentes de irracionalidade. Muitos vieses cognitivos surgem quando defendemos posições inflamadas. Os seres humanos buscam evidências de maneira seletiva, procurando apenas evidência favorecendo a sua posição e negligenciando a evidência em contrário. Como mencionado, podemos até tomar ações para prejudicar outras partes e nós mesmos ao defender o nosso partido. Parece que isso seria um caso simples, onde o melhoramento moral seria desejável ao eliminar irracionalidade política cega. No entanto, pode ser o caso de que a organização entre milhares ou milhões só pode surgir se os indivíduos se comprometerem a algumas escolhas ou partidos políticos binários. Se as posições dos indivíduos puderem ir além das posições políticas opostas, escolhendo apenas as partes que julgassem verdadeiras e coerentes de cada um – como a razão ditaria, então o caos poderia surgir. Da mesma forma, se as pessoas pudessem apreciar o futebol apenas porsuas por boas jogadas, em vez de torcer pelo seu time preferido, isso resultaria no desaparecimento do futebol como uma instituição social rentável.

Note-se que essas são considerações apenas exploratórias, cuja plausibilidade *prima facie* refuta uma implicação necessária entre cooperação individual e a cooperação entre grupos. No entanto, essas considerações não desmerecem a alegação de que, todos os fatos considerados, os aumentos na cooperação individual levam a aumentos na cooperação entre grupos. Se confiarmos numas das principais posições sobre a semântica de modalidade, conceptibilidade primária não-ideal (o fato de que alguém pode conceber algo como verdadeiro sobre o nosso mundo) implica possibilidade epistêmica (algo que pode ser verdade no nosso mundo). Portanto, meramente conceber cenários coerentes onde existem aumentos na cooperação individual mas sem um aumento da cooperação de grupo significa que o anterior poderia ser verdade (epistemicamente possível), enquanto o último é falso; isto é, o primeiro não garante ou implica o segundo (CHALMERS, 2002). Estritamente falando, nossos casos hipotéticos mostram que a cooperação entre grupos não supervém necessariamente à cooperação entre indivíduos. Por isso, estes casos nos dão motivos contra a ingênua crença que o melhoramento moral poderia resolver problemas de

cooperação entre grupos ao aumentar a cooperação entre indivíduos. Pelo contrário, aumentar a cooperação individual poderia até apresentar graves riscos ao diminuir a cooperação entre grupos. Por outro lado, estas considerações não podem ser usadas para concluir que não há mecanismos plausíveis para aumentar a cooperação entre os grupos aumentando-se a cooperação entre os indivíduos, ou que esses mecanismos não podem vastamente superar as cinco considerações acima apresentadas. Isso só seria estabelecido se pudéssemos provar que aumentos de cooperação do grupo são inconcebíveis se aumentarmos a cooperação individual. Evidentemente, ao contrário dos nossos argumentos relativos à pílula individualista, estes argumentos não devem ser tomados como uma prescrição para uma proibição *ex ante* da pílula cooperativa.

Em vez disso, eles devem ser tomados como evidência para a relevância do nível social emergente. Se a neurociência e psicologia da moralidade e comportamento social focarem-se exclusivamente na busca de melhorias morais no nível individual, então, não só poderíamos criar catástrofes sociais, mas vamos certamente estar ignorando aspectos muito importantes que dizem respeito à própria natureza dos problemas que o melhoramento moral almeja corrigir. Apesar de devidamente tratados em Savulescu & Persson (2012) e em outros lugares, as consequências do melhoramento moral para a política, relações internacionais e resolução de conflitos só podem ser adequadamente abordadas por pesquisa científica buscando o desenvolvimento de melhorias morais, se as configurações experimentais focarem em estratégias sociais de grupos para com outros grupos, ao invés de estratégias de indivíduos para com outros indivíduos.

4.5. Conclusão

Mostrou-se que as aquisições de estratégias sociais como as escolhas livres de prateleiras de um supermercado podem levar ao individualismo generalizado. Mais importante, a simulação prevê que as drogas individualistas devem ser banidas antes de serem desenvolvidas. Além disso, foram apresentadas razões para acreditar que um paraíso cooperativo ingenuo pode nos custar muito caro. Promover o uso de uma droga coperacional, de um jeito ou de outro, poderia desmantelar completamente a sociedade. A evolução natural criou um grande grupo de restrições sobre o que os seres humanos podem ou não podem ser, e assim, fez o que nos define e que nós valorizamos, reduzindo as nossas escolhas. Podemos usar as nossas aspirações internas e moral para definir como as coisas devem ser. Mas quando, ao nos libertarmos das correntes da

evolução, formos maravilhados com inúmeras possibilidades disponíveis, não teremos mais um conjunto de características e preferências comuns, que devem ser necessariamente perseguidas e através das quais todas as outras opções podem ser avaliadas. No entanto, e acreditando que o nosso tratamento da questão evidencia isso, não temos necessidade de nos confundir com especulação metafísica sobre o que significa ser humano. Ao tentar modelar como as pessoas irão escolher quando escolhem livremente estratégias sobre como escolher, fomos capazes de construir um modelo palatável do processo de decisão subjacente a tais escolhas e traçar predições, ainda que tentativas. Não obstante, não se pode negar o fato de esta também ser uma tentativa de resolver questões filosóficas sobre a natureza humana e que a simulação se baseia em pressupostos filosóficos sobre tal natureza. Como nunca antes, a humanidade não pode escapar dessas questões filosóficas sobre sua própria natureza, pois nunca antes tais especulações irão moldar tão dramaticamente o nosso futuro. Se o nosso entendimento destas questões for ultrapassado pelo nosso poder de modificar a nós mesmos, corremos o risco de extinção. Como Bostrom (2014) afirmou, o nosso desenvolvimento tecnológico acelerado cria a necessidade de investigação filosófica com um prazo limite.

5. Conclusão

Em termos práticos, aspectos fundamentais e universais humanos estão longe de serem perfeitos, ao contrário, possuem graves vieses cognitivos e inaptidões morais que podem levar catástrofes. Temos também boas razões teóricas, advindas da heurística evolutiva, para suspeitar que produtos evolutivos são muitas vezes falhos. O melhoramento humano é, portanto, desejável. Adicionalmente, tais aspectos humanos são potencialmente mutáveis com o uso da tecnologia, como se exemplificou com estudos sobre drogas já disponíveis que provocam grandes mudanças - e esperançosamente melhoramentos - afetivas, cognitivas e morais. O melhoramento humano é, portanto, provavelmente factível. Ademais, certos problemas da moral humana, se não resolvidos, podem acarretar a extinção da humanidade. Tentar desenvolver o melhoramento humano é, portanto, um imperativo em alguns casos. No entanto, pode acarretar severos riscos a humanidade. O melhoramento humano é potencialmente catastrófico.

Consequentemente, a conclusão dessa dissertação é que o melhoramento humano é - especificidades a parte - desejável, provavelmente factível, por vezes obrigatório, mas nestas vezes também potencialmente catastrófico. Desejável, factível, obrigatório e catastrófico certamente são propriedades inconvenientes de serem atribuídas juntas a um mesmo conceito. Mas foi mostrado também que essa situação é potencialmente solúvel. Podemos mapear quando e onde o melhoramento será catastrófico e evitar tais caminhos, como foi feito no último capítulo deste trabalho. Mas ainda fica em aberto se as instâncias nas quais o melhoramento não é catastrófico não sejam também aquelas nas quais ele não é mais um imperativo, pois não se foca mais naqueles problemas que apenas uma modificação profunda e radical poderia resolver. Pergunta-se novamente se existe uma solução entre trivialmente não realizar nenhum melhoramento mais ambicioso, e portanto ser omisso com nossa inaptidão potencialmente catastrófica, e entre desestabilizar completamente a condição humana, levando também à catástrofe.

6. Agradecimentos

Os questionamentos explorados nesse trabalho, bem como o rumo intelectual e acadêmico que tenho almejado e traçado há mais de meia década, seriam largamente diferentes sem a direção dada tanto pelos escritos como pela orientação por e-mails do Prof. Dr. Nick Bostrom, pelo qual possuo enorme gratidão e dívida. Sem o mesmo tipo de aconselhamento mais recente fornecido pelo Prof. Dr. Julian Savulescu, meu rumo acadêmico atual também teria sido outro. O Dr. Anders Sandberg também teve papel fundamental em minha formação. Sem sua inteligência, curiosidade e vitalidade intelectuais exemplares e instigantes muitas das potenciais originalidades contidas nos dois últimos capítulos desta pesquisa seriam ou inexistentes ou precárias e certamente menos alegres de se obter. Os comentários e revisões de minhas argumentações do capítulo II feitas pelo Prof. Dr. Nick Agar também me foram inestimáveis, inclusive revisões dos meus argumentos contrários aos argumentos do próprio Agar contra o melhoramento humano radical.

Também reconheço aqui o papel fundamental do apoio material de minha mãe, Maria Regina de Araujo, durante a execução dessa pesquisa. Quando faltaram recursos financeiros institucionais, ela se dispôs a financiar minhas duas viagens de pesquisa para a Inglaterra. Agradeço também a minha parceira Barbara Belle de Oliveira, por me apoiar, me entender, me incentivar, acreditar em mim e sempre ter a certeza de que eu iria conseguir meus objetivos – ainda que eu mesmo não a tivesse. Além disso, meu pai, Luiz Hermenegildo Fabiano, teve um papel fundamental na minha formação filosófica primeira, contribuindo para que descobrisse o que hoje tenho certeza ser – na medida em que minha racionalidade me permite – quase um destino na minha vida: o exercício filosófico.

Quero expressar meu enorme reconhecimento e gratidão ao orientador desta dissertação de mestrado, Prof. Dr. Osvaldo Frota Pessoa Jr., por permitir, orientar e auxiliar a execução deste projeto. Sem a possibilidade burocrática e de recursos para realizar esse mestrado que me foi concedida pelo Prof. Osvaldo, certamente essa pesquisa seria marginalmente impossível. Minha gratidão também se estende não só por garantir a executabilidade acadêmica de minha pesquisa, mas também por abrir espaço

para outros pesquisadores para os quais, assim como eu, esta execução ficaria dificultada de outro modo. Ademais, o apoio e orientação dos poucos pesquisadores brasileiros que encontrei dispostos a engajar na discussão acerca do melhoramento moral e/ou investigar o referencial teórico necessário a essa discussão foi imprescindível. A isso quero agradecer aos membros da banca de defesa, Prof. Dra. Maria Clara Dias e Prof. Dr. Adriano Naves de Brito, bem como o Prof. Dr. Marcelo de Araujo e o Prof. Dr. Brunello Stancioli. Adicionalmente, também devo minha gratidão ao apoio logístico e organizacional prestado por todos os colaboradores do Instituto de Ética, Racionalidade e Futuro da Humanidade (IERFH), em especial ao diretor Diego Caleiro, sem o qual não teria conhecido e me adentrado tão intensamente neste debate, na filosofia analítica e em tantos outros projetos de vida.

Por fim, gostaria de agradecer as inúmeras pessoas que me ajudaram seja durante a elaboração de meu projeto de doutoramento, no qual os dois últimos capítulos desta dissertação foi parcialmente baseado, seja num projeto de pesquisa empírica na área que infelizmente teve sua execução comprometida. Em ordem alfabética: Alexandre Euler, Alex Schell, Brian Earp, Brian Tomasik, Carl Shulman, Corinna Elsenbroich, Emma Bates, Jen Badham, Juan Cano, Johann Roduit, Kaj Sotala, Kristian Rönn, Lucas Nascimento Machado, Mark Weber, Marie Odile Monier Chelini, Miriam Wood, Nadira Faulmüller, Nick Beckstead, Paul Christiano, Pieter Bonte, Regina Rini, Ricardo Bindi, Ryan Murphy, Robert Gifford, Robert Rogers, Sean O'Heigeartaigh, Simon Driscoll e Stuart Armstrong. Certamente faltam nomes. Infelizmente, o reconhecimento específico e merecido de cada uma das contribuições individuais acima iria em muito alongar uma seção de agradecimentos já extensa, portanto, isso foi omitido.

7. Bibliografia

ADORNO, Theodor & HORKHEIMER, Max. “Dialética do esclarecimento.” 3ª ed. Trad.: Guido A. de Almeida. RJ: Zahar. 1991.

AGAR, Nicholas. “Truly human enhancement: A philosophical defense of limits.” MIT Press. 2014.

AGAR, Nicholas. “Moral bioenhancement is dangerous.” J Med Ethics 0:1–4. 2013.

ALEXANDER, Larry & MOORE, Michael. "Deontological ethics." The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), Acessado on-line em 06/08/2012, disponível em: http://plato.stanford.edu/archives/fall2008/entries/ethics-deontological/. 2008.

AZZAM, K. M. A. et al. “Enantioselective determination of modafinil in pharmaceutical formulations by capillary electrophoresis, and computational calculation of their inclusion complexes.” Microchimica Acta, 166(3-4), 311-317. 2009.

BARANSKI, J. V. et al. "Effects of modafinil on cognitive and meta-cognitive performance." Hum Psychopharmacol. Jul; Vol. 19(5):323-32. 2004.

BARTELS, Daniel M. et al. "Principled moral sentiment and the flexibility of moral judgment and decision making." Cognition, v. 108, n. 2, p. 381. 2008.

BECKSTEAD, Nick. “On the overwhelming importance of shaping the far future.” PhD Thesis. Department of Philosophy, Rutgers University, USA. 2013.

BENTHAM, Jeremy, "An introduction to the principles of morals and legislation." Library of Economics and Liberty. 1907.

BORN et al., “Sniffing neuropeptides: An intransnasal approach to the human brain.” Nat Neurosci. Jun;5(6):514-6. 2002.

BOSTROM, Nick. “Superintelligence: Paths, dangers, strategies.” Oxford University Press. 2014.

BOSTROM, Nick. “The Transhumanist FAQ v.2.1”, publicado pela World Transhumanist Association. 2004.

BOSTROM, Nick. “The future of human evolution.” Death and Anti-death: Two Hundred Years After Kant, Fifty Years After Turing. 2004b.

BOSTROM, Nick & ORD, Toby. “The reversal test: Eliminating status quo bias in applied ethics.” Ethics 116 Julho: 656-679. 2006.

BOSTROM, Nick. “Curriculum Vitae.” Disponível em: http://www.nickbostrom.com/cv.pdf. Acesso em: 8 de janeiro de 2012.

BOSTROM, Nick. “Existential risk reduction as the most important task for humanity.” Global Policy. 2011.

BOSTROM, Nick. “Existential risks: Analyzing human extinction scenarios.” Journal of Evolution and Technology, Vol. 9.1. 2002.

BOSTROM, Nick. “Human genetic enhancements: A transhumanist perspective.” Journal of Value Inquiry, Vol. 37, No. 4. 2003.

BOSTROM, Nick. “In defense of posthuman dignity.” Bioethics, v. 19, n. 3. 2005.

BOSTROM, Nick. “Transhumanist values.” Review of Contemporary Philosophy, Vol. 4, No. 1-2. 2005.

BOSTROM, Nick & SANDBERG, Anders. “Converging cognitive enhancements.”, Annals of the New York Academy of Sciences, Vol. 1093. 2006.

BOSTROM, Nick & SANDBERG, Anders. “Cognitive enhancement: Methods, ethics, regulatory challenges.”, Science and Engineering Ethics, Vol. 15, No. 3. 2009b.

BOSTROM, Nick & SANDBERG, Anders. “The wisdom of nature: An evolutionary heuristic for human enhancement.” In: BOSTROM, Nick e SAVULESCU, Julian (orgs.). Human Enhancement. Oxford University Press, EUA. 2009a.

BOYD, Robert et al. “The evolution of altruistic punishment.” Proceedings of the National Academy of Sciences, v. 100, n. 6, p. 3531-3535. 2003.

BURT, T. “Donepezil and related cholinesterase inhibitors as mood and behavioral controlling agents.” Current Psychiatry Reports, 2(6), 473-478. 2000.

BUSS, David (orgs.). “The handbook of evolutionary psychology”. Wiley, New Jersey. 2005.

CAIDWELL, John A. et al. "The effects of modafinil on aviator performance during 40 hours of continuous wakefulness: A UH-60 helicopter simulator study." Army Aeromedical Research Unit Fort Rucker, Junho. 1999.

CAIDWELL, John A. et al. "The efficacy of modafinil for sustaining alertness and simulator flight performance in F-117 pilots during 37 hours of continuous wakefulness." Air Force Research Lab Brooks AFB TX, Human Effectiveness Dir/Biodynamics and Protection Div, Janeiro. 2004.

CHALMERS, David. “The singularity: A philosophical analysis.” Journal of Consciousness Studies, 17(9-10). 2010.

CHALMERS, David. “Does conceivability entail possibility?” In GENDLER, T. & HAWTHORNE, J. (eds.), Conceivability and Possibility. Oxford University Press. 2002.

CIRKOVIC, Milan M. & REES, Martin J. (org.) “Global catastrophic risks.” Oxford: Oxford University Press. 2008.

CONWAY, Paul & GAWRONSKI, Bertram. “Deontological and utilitarian inclinations in moral decision making: A process dissociation approach.” 2012.

COSMIDES, Leda e TOOBY, John. “Evolutionary psychology: A primer.” Center for Evolutionary Psychology. Disponível em: http://www.psych.ucsb.edu/research/cep/primer.html Acessado: 24-08-2011. 1997.

CRAWFORD, Charles e KREBS, Dennis (orgs.). "Foundations of evolutionary psychology." Psychology Press, 2ª edição. 2008.

DANIELS, N. “Can anyone really be talking about ethically modifying human nature.” Human Enhancement. 2009.

DAVIS, Mark H. "Measuring individual differences in empathy: Evidence for a multidimensional approach." Journal of Personality and Social Psychology, v. 44, n. 1, p. 113-126. 1983.

DAWES, Robyn M. & MESSICK, David M. “Social dilemmas.” International Journal of Psychology, v. 35, n. 2, p. 111-116. 2000.

DE QUEIROZ, K. “Species concepts and species delimitation.” Systematic Biology, 56(6). 2007.

DE GRAY, Aubrey. "Ending aging." 1ª edição, St. Martin's Press. 2007.

DITZEN et al. “Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict.” Biol Psychiatry. 2009 May 1;65(9):728-31. 2009.

DOMES, et al. “Oxytocin improves “mind-reading” in humans.” Biol Psychiatry. Mar 15;61(6):731-3. 2007.

DOUGLAS, Thomas. “Intertemporal disagreement and empirical slippery slope arguments.” Utilitas, 22(2), 184. 2010.

DOUGLAS, Thomas. “Moral enhancement.” Journal of Applied Philosophy, 25(3), 228-245. 2008.

DOUGLAS, Thomas. “The morality of moral neuroenhancement.” In: CLAUSEN, J. & LEVY, N. (Eds). “Handbook of neuroethics.” Springer. 2014.

DOUGLAS, Thomas & DEVOLDER, K. “Procreative altruism: Beyond individualism in reproductive selection.” Journal of Medicine and Philosophy, 38(4), 400-419. 2013.

EARP, Brian & SANDBERG, Anders. SAVULESCU, Julian. "Natural selection, childrearing, and the ethics of marriage (and divorce): Building a case for the neuroenhancement of human relationships." Philosophy & Technology (2012): 1-27. 2012.

EPSTEIN, Seymour et al. “Individual differences in intuitive–experiential and analytical–rational thinking styles.” Journal of Personality and Social Psychology, v. 71, n. 2, p. 390, 1996.

ELLISON, Peter (orgs.). “Endocrinology of social relationships.” Harvard University Press . 2009.

FALCH, Torberg & SANDGREN, Sofia. "The effect of education on cognitive ability." Economic Inquiry Journal Volume 49, Exemplar 3, pp. 838–856. 2011.

FERRAGUTI, F. & SHIGEMOTO, R. "Metabotropic glutamate receptors." Cell and Tissue Research, 326(2), 483-504. 2006.

FISCHER-SHOFTY et al. "Oxytocin facilitates accurate perception of competition in men and kinship in women." Soc Cogn Affect Neurosci. 2012.

FISHBURN, Perter. "Utility theory for decision making." Huntington, NY: Robert E. Krieger. 1970.

GIFFORD, Jonas & GIFFORD, Robert. "FISH 3: A microworld for studying social dilemmas and resource management." Behavior Research Methods, Instrumentation, and Computers, 32, 417- 422. 2000.

GIFFORD, Robert & GIFFORD, Jonas & BEARDEN, Anomi. "Manual for FISH 3.1" Disponível em: http://web.uvic.ca/~rgifford/fish/FISH%203.1%20Manual.pdf

GILBERT, Roberts. "Competitive altruism: From reciprocity to the handicap principle." Proc. R. Soc. Lond. B 265. 1998.

GREELY, H. et al. "Towards responsible use of cognitive-enhancing drugs by the healthy." Nature, 456(7223), 702-705. 2008.

GREENBERG, J. A. et al. "Caffeinated beverage intake and the risk of heart disease mortality in the elderly: A prospective analysis." Am J Clin Nutr 85 (2): 392–8. 2007.

GREENE, Joshua D. "The secret joke of Kant's soul, in moral psychology.", In SINNOTT-ARMSTRONG, W. (ed). "Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development." MIT Press, Cambridge, MA. 2007.

GRENEE, Joshua. "From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology." Nature Reviews: Neuroscience, Volume 4. October. 2003.

GRENEE, Joshua. "Moral tribes: Emotion, reason and the gap between us and them." Atlantic Books. 2014.

GRÖN, Georg et al. "Cholinergic enhancement of episodic memory in healthy young adults." Psychopharmacology (2005) 182: 170–179. 2005.

GUASTELLA et al. "Oxytocin enhances the encoding of positive social memories in humans." Biol Psychiatry. 2008 Aug 1;64(3):256-8. 2007.

HAMILTON, William D. "The genetical evolution of social behaviour (I and II)." Journal of Theoretical Biology, v. 7, n. 1, p. 1-16. 1964.

HANAGE, William P. "Fuzzy species revisited." BMC Biology, 11(1), 41. 2013.

HARTZ, B. P. & RONN, L. C. B. "NCAM in long-term potentiation and learning. In structure and function of the neural cell adhesion molecule." NCAM. (pp. 257-270). Springer New York. 2010.

HERTZOG, C. et al. “Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory?” Psychology and Aging 18.4: 755-769. 2003.

ILLY, A. & VIVIANI, R. “Espresso coffee: The chemistry of quality.” San Diego: Academic P. 1995.

JULIANO, Laura M. & GRIFFITHS, Roland R. “A critical review of caffeine withdrawal: Empirical validation of symptoms and signs, incidence, severity, and associated features.” Psychopharmacology 176 (1): 1–29. 2004.

KAHANE, Guy & SHACKEL, Nicholas. “Methodological issues in the neuroscience of moral judgement.” Mind & language, v. 25, n. 5, p. 561-582. 2010.

KAHNEMAN, Daniel (orgs.) “Choices, values, and frames.” Cambridge University Press; 1ª edição. 2000.

KAHNEMAN, Daniel & TVERSKY, Amos. “Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.” Psychological Review, 90: 293-315. 1983.

KAHNEMAN, Daniel. “A Psychological perspective on economics.” The American Economic Review, Vol. 93, No. 2. 2003.

KAHNEMAN, Daniel. “Economic analysis and the psychology of utility: Applications to compensation policy.” The American Economic Review, Vol. 81, No. 2. 1991.

KAHNEMAN, Daniel. “Loss aversion in riskless choice: A reference-dependent model.” The Quarterly Journal of Economics, Vol. 106, No. 4 (Nov., 1991), pp. 1039-1061. 1991.

KANHEMAN, Daniel (orgs.) “Heuristics and biases: The psychology of intuitive judgment.” Cambridge University Press. 2002.

KAPNER, E. “Recreational use of ritalin on college campuses.” InfoFactsResources – The Higher Education Center for Alcohol and Other Drug Prevention. Available at: www.edc.org/hec/pubs/factsheets/ritalin.pdf (accessed 4 Jan 2006). 2003.

KIRSCH et al. “Oxytocin modulates neural circuitry for social cognition and fear in humans.” The Journal of Neuroscience, 7 December 2005, 25(49): 11489-11493. 2005.

KOENIG, Laura B. et al. “Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings.” Journal of Personality, v. 73, n. 2, p. 471-488. 2005.

KONTKANEN, O. & CASTRÉN, E. “Trophic effects of selegiline on cultured dopaminergic neurons.” Brain Research, 829(1), 190-192. 1999.

KOSFELD et al. “Oxytocin increases trust in humans.” Nature 435, 673-676 (2 June 2005). 2005.

KOUTROUBAKIS, I. E. & VLACHONIKOLIS, I. G. “Appendectomy and the development of ulcerative colitis: Results of a meta-analysis of published case control studies.” American Journal of Gastroenterology 95(1): 171–6. 2000.

KREBS, Dennis. L. “The origins of morality: An evolutionary account.” New York: Oxford University Press. 2011.

KRUGER et al. “Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses.” Soc Cogn Affect Neurosci. 2012 Mar 24. 2012.

KURZWEIL, Raymond. “The law of accelerating returns.” Disponível e acessado online (15-07-2012) em: http://www.kurzweilai.net/the-law-of-accelerating-returns

LEITER, Brain. “Ranking of graduate programs in philosophy in the english-speaking world”. Philosophical Gourmet Report. 2013.

LESON. C. L. et al. “Caffeine overdose in an adolescent male.” Journal of Toxicology. Clinical Toxicology Vol. 26 (5–6): 407–15. 1988.

LEVINE, David K. “Game theory.” Nature Publishing Group, Encyclopedia of Cognitive Science. 2014.

LI Yanfeng, ZHAN Hao, XIN Yimei et al. “Effects of modafinil on vestibular function during 24 hour sleep deprivation.” Frontiers of Medicine in China, Vol. 1, Number 2, 226-229. 2007.

MAYER, J. D. & GASCHKE, Y. N. “The experience and meta-experience of mood.” Journal of Personality and Social Psychology, 55, 102-111. 1988.

MOORE, A. B. et al. “Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment.” Psychological Science, 19(6), 549-557. 2008.

MÜLLER, U. et al. “Effects of modafinil on working memory processes in humans.” Psychopharmacology (Berl.) Vol. 177 (1-2): 161–9. 2004.

MULLER, U. et al. “Effects of modafinil on non-verbal cognition, task enjoyment and creative thinking in healthy volunteers.” Neuropharmocology: 2012 (In press). 2012.

MURPHY, Ryan & ACKERMANN, Kurt & HANDGRAAF, Michel. “Measuring social value orientation.” 2011.

NAKAMURA, K. & SHIRANE, M. “Activation of the reticulothalamic cholinergic pathway by the major metabolites of aniracetam.” European Journal of Pharmacology, 380(2), 81-89. 1999.

NAKAMURA, K. & SHIRANE, M. “Aniracetam enhances cortical dopamine and serotonin release via cholinergic and glutamatergic mechanisms in SHRSP.” Brain research, 916(1), 211-221. 2001.

NCDT. Report of the 2011 National Coffee Drinking Trends (NCDT). 2011.

NOLLER, Patricia & FEENEY, Judith A. (Ed.) “Understanding marriage: Developments in the study of couple interaction.” Cambridge University Press. 2002.

NYHAN, B. & REIFLER, J. “Opening the partisan mind? Self-affirmation and factual misperceptions about politics.” 2011.

OSTROM, Elinor. ”Neither market nor state: Governance of common-pool resources in the twenty-first century.” International Food Policy Research Institute. 1994.

PACHOLCZYK, A. “Moral enhancement: What is it and do we want it?” Law, Innovation and Technology, 3(2), 251-277. 2011.

PERSSON, Igmar & SAVULESCU, Julian. “Unfit for the future: The need for moral enhancement.” OUP Oxford, 2012.

PERSSON, Igmar & SAVULESCU, Julian. “Unfit for the future?: Human nature, scientific progress and the need for moral enhancement.” In: SAVULESCU, J. & MEULEN, Rudd ter (orgs.) “Enhancing Human Capacities”. Wiley-Blackwell. 2011.

PITMAN, R.K. & SANDERS, K.M. et al. “Pilot study of secondary prevention of postraumatic stress disorder with propranolol.” Biological Psychiatry Vol. 512: 189–192. 2002.

PLATEK, Steven (orgs.). “Foundations in evolutionary cognitive neuroscience”. Cambridge University Press; 1ª edição. 2009.

POHL, Rüdiger (orgs.) “Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory.” Psychology Press. 2005.

POSNER, R. A. “Catastrophe: risk and response.” Oxford University Press. 2004.

POWELL, Russell & BUCHANAN, Allen. “Breaking evolution's chains: the prospect of deliberate genetic modification in humans.” In: SAVULESCU, J. & MEULEN, Rudd ter (orgs.) “Enhancing human capacities.” Wiley-Blackwell. 2011a.

POWELL, Russell & BUCHANAN, Allen. “Breaking evolution's chains: The promise of enhancement by design.” In: Journal of Medicine and Philosophy, v. 36, n. 1, p. 6-27. 2011b.

RIEDERER, P. & LANCHENMAYER, L. “Selegiline’s neuroprotective capacity revisited”. Journal of Neural Transmission, 110(11), 1273-1278. 2003.

RODUIT, Johhan. Review of AGAR, N. 2010. “Humanity’s end: Why we should reject radical enhancement.” Med Health Care and Philosophy 14:345–350. 2011.

ROSS, Don. “Game Theory.” In: The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.) 2012.

RUSBULT, Caryl E. & VAN LANGE, Paul AM. “Interdependence, interaction, and relationships.” Annual Review of Psychology, v. 54, n. 1, p. 351-375. 2003.

SAMUELSON, William & ZECKHAUSER, Richard. “Status quo bias in decision making.” Journal of Risk and Uncertainty 1 (1988): 7–59. 1988.

SANDBERG, Anders & SAVULESCU, Julian. “Neuroenhancement of love and marriage: The chemicals between us.” Neuroethics (2008) Vol. 1:31-44. 2008.

SANDBERG, Anders & LIAO, S. M. “The normativity of memory modification.” Neuroethics (2008), (1 2) 85-99. 2008.

SANDBERG, Anders & RAVELINGIEN, A. “Sleep better than medicine? Ethical and philosophical issues related to 'wake enhancement.” Journal of Medical Ethics, Vol. 34: e9. 2008.

SANDBERG, Anders. Palestra “Swine flu, black swans, and Geneva-eating dragons.” proferida no UKH+ meeting em 20 de Junho de 2009.

SANDEL, M. “What’s wrong with enhancement.” President’s Council on Bioethics, Washington, DC (www. bioethics. gov), 12. 2002.

SAVULESCU, Julian & MEULEN, Rudd ter (orgs.) “Enhancing human capacities.” Wiley-Blackwell. 2011.

SAVULESCU, Julian & PERSSON, Igmar. “Unfit for the future? Human nature, scientific progress and the need for moral enhancement.” In SAVULESCU, Julian & MEULEN, Rudd ter (orgs.) “Enhancing human capacities.” Wiley-Blackwell. 2011.

SHAMAY-TSOORY et al. “Intranasal administration of oxytocin increases envy and Schadenfreude (gloating).” Biological Psychiatry Volume 66, Issue 9, Pages 864-870. 2009.

SINNOT-ARMSTRONG, Walter. “Consequentialism.” The Stanford Encyclopedia of Philosophy (Winter 2011 Edition), Edward N. Zalta (ed.), acessado on-line em 06/08/2012, disponível em: http://plato.stanford.edu/archives/win2011/entries/consequentialism/. 2011.

SMITH, A. “Effects of caffeine on human behavior.” Food and Chemical Toxicology 40: 1243–1255. 2002.

SMITH, M., LEWIS, David, & JOHNSTON, M. “Dispositional theories of value.” Proceedings of the Aristotelian Society, Supplementary Volumes, 63, 89-174. 1989.

SQUIRE, Larry R. et al. (orgs.) “Fundamental neuroscience.” Academic Press. 3a edição. 2008.

STAHL, Stephen M. “Awakening to the psychopharmacology of sleep and arousal: Novel neurotransmitters and wake promoting drugs.” J Clin Psychiatry 63:467-468. 2002.

STOOPS, W. et al. “Reinforcing effects of modafinil: Influence of dose and behavioral demands following drug administration.” Psychopharmacology, 182(1), 186-193. 2005.

TALEB, Nassim. “The black swan: Why don't we learn that we don't learn?” New York: Random House. 2005.

TEMKIN, Larry. “Rethinking the good: Moral ideals and the nature of practical reasoning.” Oxford University Press. pp. 238-262. 2012.

TERHERSON, David J. & RAFTERY, James. “Discouting.” Economic Notes, BMJ 2 de Outubro de 1999; Vol. 319(7214). pp. 914–915. 1999.

TETER, C. J. et al. “Prevalence and motives for illicit use of prescription stimulants in an undergraduate student sample.” J Am Coll Health 53. 2005.

TRIVERS, Robert L. “The evolution of reciprocal altruism.” Quarterly Review of Biology, p. 35-57, 1971.

TSUNEKAWA, H. et al. “Synergistic effects of selegiline and donepezil on cognitive impairment induced by amyloid beta (25–35).” Behavioural Brain Research, 190(2), 224-232. 2008.

TURNER, D. C. et al. “Cognitive enhancing effects of modafinil in healthy volunteers.” Psychopharmacology (Berl.) Vol. 165 (3): 260–9. 2003.

UNKELBACH et al. “Oxytocin selectively facilitates recognition of positive sex and relationship words.” Psychol Sci. 2008 Nov;19(11):1092-4. 2008.

VOLOKH, E. “The mechanisms of the slippery slope.” Harvard Law Review, 116(4), 1026-1137. 2013.

VAN LANGE, Paul AM et al. “The psychology of social dilemmas: A review.” Organizational Behavior and Human Decision Processes, v. 120, n. 2, p. 125-141. 2013.

VAN LANGE, Paul A. M. & JOIREMAN, Jeff A. “How we can promote behavior that serves all of us in the future.” Social Issues and Policy Review, v. 2, n. 1, p. 127-157. 2008.

VON NEUMANN, John & MORGESTERN, Oskar. “Theory of games and economic behavior.” Princenton University Press. Disponível on-line em: http://ia600301.us.archive.org/29/items/theoryofgamesand030098mbp/theoryofgamesand030098mbp.pdf. 1944.

WEBER, M. et. al. “A conceptual review of social dilemmas: Applying a logic of appropriateness”. Personality and Social Psychology Review 8: 281–307. 2004.

WEDEKIND, Claus & MILINSKI, Manfred. “Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus Generous Tit-for-Tat.” Proceedings of the National Academy of Sciences, v. 93, n. 7, p. 2686-2689. 1996.

YOO, J. H. et al. “Relevance of donepezil in enhancing learning and memory in special populations: A review of the literature.” Journal of Autism and Developmental Disorders, 37(10), 1883-1901. 2007.

YUDKOWSKY, Eliezer. “Complex value systems are required to realize valuable futures.” Proceedings of AGI 2011. Springer. 2011.

YUDKOWSKY, Eliezer. “Value is fragile.” Less Wrong (Blog) Disponível em: http://lesswrong.com/lw/y3/value_is_fragile/. 2009.

YUDKOWSKY, Eliezer. “Introducing the “Singularity”: Three major schools of thought.” Palestra proferida no Singularity Summit em 2007.

YUDKOWSKY, Eliezer. “Cognitive biases potentially affecting judgment of global risks” In: BOSTROM, Nick (orgs.) “Global catastrophic risks.” Oxford University Press, USA. 2008.

ZAK et al. “Oxytocin increases generosity in humans”. PloS ONE , Vol. e1128, No. 11. 2007.

Anexo A – Simulação da Difusão de Melhoramentos Morais:

Código-fonte

```
% Model assumptions

% Early adopters: small faction will take up tech – right now
represented
% by initial state, we can add noise

% Imitation: pickup proportional to how successful strategy is – right
now
% modelled as comparing yourself to one other, imitating him if doing
% better

% Punishment non-coperators – not in this basic model
% Punishment requires cooperations
% Sociovalue orientations – built in

% 1 = Altruist: MaxOther (Blue)
% 2 = Cooperative: MaxJoint (Light blue)
% 3 = Individualist: MaxOwn (Cyan)
% 4 = Equalitarian: MinDiff (Yellow)
% 5 = Competitive: MaxDiff (Red)
% 6 = Agressive: MinOther (Dark Red)


N=150; % Population size
learningRate = 0.1; % How much to imitate
choosable=[1 1 1 1 1 1]; % What orientations can change?
randfreq=0.00001; % How often to randomize one orientation?
reps=100; % Repetitions
tmax=150; % Timesteps

weight=rand(N,6);

%weight(:,2)=0.2;
%weight(:,3)=0.3;

for i=1:N; weight(i,:)=weight(i,:)/sum(weight(i,:)); end; % Normalize

choices=[]; % Store history

ori=[]; % Store weights sum

function payoffs = payoffmatrix(weights)
```

```
% Assume self is row player, other column

payoffs=weights(1)*nor([3 5; 0 1]); % 1 = Altruist
payoffs=payoffs+weights(2)*nor([6 5; 5 2]); % 2 = Cooperative
payoffs=payoffs+weights(3)*nor([3 0; 5 1]); % 3 = Individualist
payoffs=payoffs+weights(4)*nor([0 -5; -5 0]); % 4 = Equalitarian
payoffs=payoffs+weights(5)*nor([0 5; 5 0]); % 5 = Competitive
payoffs=payoffs+weights(6)*nor([-3 -5; 0 -1]); % 6 = Agressive

%    C   D
%C [3,3 0,5]
%D [5,0 1,1]

    function M2 = nor(M)
        % Normalize payoffs to [-1 1]
        ma=max(M(:)); mi=min(M(:));
        M2=(M-mi)/(ma-mi);
    end
end


for tim=1:tmax
    score=zeros(N,1);
    cc=[];
    for rep=1:reps
    for i=1:N
        j=ceil(rand*N);
        % Individual payoffs based on orientation
        payoffMatrixI=payoffmatrix(weight(i,:));
        payoffMatrixJ=payoffmatrix(weight(j,:));

        % Decide
        % Currently simplistic maximizer
        [dummy,choicei]=max(max(payoffMatrixI'));
        [dummy,choicej]=max(max(payoffMatrixJ'));

        % Reward according to individual payoffs
        score(i)=score(i)+payoffMatrixI(choicei,choicej);
        score(j)=score(j)+payoffMatrixJ(choicej,choicei);

        cc=[cc choicei choicej]; % Save choices
    end
    end

    choices=[choices; cc-1]; % Store the behavior

    ori=[ori; sum(weight.^4)]; % Store weights sums to the power

%    plot(tim,score,'.');
%    hold on
clf
figure(1)
    plot(weight')
   drawnow
```

```
figure(2)
    area(ori)
%    drawnow

 % Update orientation
    f=find(choosable);
    for i=1:N
        j=ceil(rand*N);
        if score(j)>score(i)
            %ow=weight(i,3); % only allow increased Ind
            weight(i,f)=(1-
learningRate)*weight(i,f)+learningRate*weight(j,f);
            %if (weight(i,3)<ow) weight(i,3)=ow; end
            %weight(i,:)=weight(i,:)/sum(weight(i,:));
        end
        if (rand<randfreq) weight(i,ceil(rand*6))=rand;
weight(i,:)=weight(i,:)/sum(weight(i,:)); end
    end
end
figure(3)
plot(mean(choices'))
```