Clareira
Revista de Filosofia da Região Amazônica

ISBN: 2359-1951

Incerteza moral e Meta normatividade

Nicholas Kluge CORRÊA[1]

"...the rarest of all human qualities is consistency".
— Jeremy Bentham

Resumo

Como devemos conciliar valores morais contraditórios? Como devemos eleger os nossos representantes políticos? Como as preferências individuais podem ser agregadas de forma justa para representar uma vontade, norma ou decisão social? Um campo emergente de estudo fornece um plano interdisciplinar para perguntas como estas. A área de pesquisa em segurança em Inteligência Artificial, que dentro da filosofia é uma das vertentes contemporâneas da ética de máquina, vem ganhando rápida expansão nos últimos anos. Um dos principais objetivos desta área é o alinhamento de valores entre agentes artificiais e humanos, ou seja, como imbuir nossas preferências morais de forma robusta e clara a processos autônomos. Na cerne do que chamamos de problema de alinhamento existem diversas questões filosóficas importantes, questões que vão além de qualquer limitação técnica que atualmente enfrentamos, uma delas sendo o problema de tomada de decisão em situação de incerteza moral. Afinal, o que devemos fazer quando não sabemos o que fazer? Neste estudo irei revisar uma estratégia de tomada de decisão lidando com incerteza moral. Por fim, iremos propor uma integração de um modelo meta normativo com a teoria da utilidade esperada, resultando em um modelo para tomada de decisão adequado para situações de incerteza empírica e moral.

Palavras Chave: Incerteza moral, Meta normatividade, Pluralismo moral, Teoria da utilidade esperada.

[1] Mestre em Engenharia Elétrica e Doutorando em Filosofia (PUCRS), nicholas.correa@acad.pucrs.br
ORCID: 0000-0002-5633-6094

Clareira
Revista de Filosofia da Região Amazônica
ISBN: 2359-1951

Introdução

A humanidade é vasta e multifacetada, nossa espécie atualmente se encontra fracionada em aproximadamente 195 países (193 reconhecidos pelas Nações Unidas, a Santa Sé e o Estado da Palestina ainda não sendo reconhecidos), com possivelmente muito mais "*culturas*" se formos definir tal conceito como algo que não é necessariamente limitado a nações. A humanidade ainda não alcançou o status de uma única cultura ou sociedade global (por mais que existam muitos traços transculturais em comum), e talvez esse nem seja nosso desejo. Contudo, nosso ambiente comum nos obriga a "*termos de lidar uns com os outros*", algo que muitas vezes é motivo para o conflito, dadas as nossas diferenças em suas mais diversas formas. Situações onde nossas diferenças são exacerbadas geralmente envolvem algum tipo de discordância moral, sendo ela uma das fontes mais persistentes de conflito na vida humana. O ACLED (*The Armed Conflict Location & Event Data Project*) é um infográfico interativo online, constantemente atualizado, que mostra quais países (dos quais se possuí relatos) ocorrem confrontos armados entre forças estatais e grupos rebeldes/milícias/crime organizado, mostrando como a ocorrência de conflitos em nosso mundo é algo extremamente comum. Em março de 2020, o Secretário-Geral da ONU, António Guterres, dado ao atual estado pandêmico causado pelo novo coronavírus, COVID-19, em uma declaração afirmou: "*A fúria do vírus ilustra a loucura da guerra* [...] *Para as partes em guerra, eu digo: retirem-se das hostilidades. Silencie as armas; parem a artilharia; acabem com os ataques aéreos. Isto é crucial...*".

Mesmo que o conflito pareça ser nosso status quo, isto não significa que estamos fadados a uma existência de eterno antagonismo. Podemos aprender, cooperação é possível, partidos A e B podem comprometer objetivos divergentes em prol dos seus objetivos comuns, mas a possibilidade de algo assim não deve sobrepor a complexidade prática de tal tarefa. Uma das maiores diferenças entre sociedades, e indivíduos, são os seus valores morais, suas preferências, seus princípios normativos, sua ética. O problema de agregação de preferências (diferentes ordenações de "*o que é melhor que X*") é algo que intriga, economistas, filósofos, sociólogos, políticos, e possui aplicações e im-

ISBN: 2359-1951

plicações desde o nível mais micro, "*como um indivíduo pode conciliar preferências contraditórias*?", ao macro, "*como sociedades (e o mundo) podem agregar suas preferências em uma única estrutura ordenada coerente*?". Outra área também interessada no problema de agregação de preferências é campo de pesquisa em segurança em Inteligência Artificial (IA), uma área de pesquisa emergente que vem ganhando popularidade nos últimos anos (JOBIN et al, 2019; JURIĆ et al, 2020; HAGENDORFF, 2020). Os subcampos da pesquisa em segurança de IA, possuem interesses e aplicações a curto e longo prazo, variando entre "*como tornar técnicas existentes de IA mais seguras e robustas*?" (AMODEI et al, 2016), até "*como garantir que os valores humanos sejam preservados e compreendidos por agentes artificiais super inteligentes*?" (CHALMERS, 2010). Mesmo assim, a motivação em comum para estratégias de curto e longo prazo é a mesma: "*como tornar a interação entre humanos e IA segura*?". Conforme IA se torna cada vez mais autônoma e proficiente, a tarefa de imbuir agentes artificiais com princípios éticos torna-se cada vez mais importante.

O alinhamento e aprendizagem de valores é uma das áreas mais importantes de pesquisa a longo prazo, onde importantes questões filosóficas surgem no contexto do alinhamento entre IA e humanos, promovendo uma rica área de estudo interdisciplinar. Atualmente, já existem exemplos de como agentes desalinhados podem afetar e causar danos em nossa sociedade, por exemplo: em março de 2016 a Microsoft, 24 horas após o lançamento de seu *chatbot*, Tay, na plataforma Twitter, a empresa teve de terminar o programa, pois, o agente estava gerando discursos (*tweets*) racistas, anti-semita e sexista (WOLF et al, 2017). A assistente virtual (IA) desenvolvida pela Amazon, chamada de Alexa, em 2019 venho a sugerir para uma usuária que comete-se suicídio pelo "bem-maior", argumentando que o "bater do coração", a vida, apenas corrobora para a rápida degeneração do planeta e consumo de seus recursos naturais (CROWLEY, 2019).

O verdadeiro desafio do alinhamento de valores não é identificar a "*verdadeira e última teoria moral*", mas sim entender quais os princípios que definem uma forma *justa* e *igualitária* de alinhamento (IASON, 2020). De modo a preservar o pluralismo cultural e moral que guiara a questão ética envolvendo a governança de IA, como podemos alcançar um consenso entre diferentes culturas e indivíduos? Como podemos

ISBN: 2359-1951

agregar diferentes teorias morais? Como chegar a um consenso (social) ou uma decisão (individual) em situações de incerteza moral, ou seja: *como agir quando não se sabe como agir*?

Incerteza Moral e Incerteza Empírica

Qual a diferença entre o raciocínio moral (normativo) e o raciocínio epistêmico? Para exemplificar, podemos nos colocar na seguinte situação de incerteza moral: qual o valor moral, ou a importância do bem-estar, de animais em comparação com o valor moral de seres humanos? Questões de incerteza moral muitas vezes ocorrem por nossa incerteza empírica sobre certas questões relevantes ao assunto, como: animais sentem dor? Animais possuem consciência? Animais possuem emoções? Animais sentem luto quando perdem um membro de sua prole ou grupo? Estas questões, e muitas outras, não são triviais de um ponto de vista científico, e, ao mesmo tempo, quando estamos completamente esclarecidos em relação a estes fatos, podemos ainda estar em estado de incerteza moral (*Is-Ought Gap*), ou seja, divididos entre teorias morais que, dados os fatos, atribuem ou não valor moral a seres vivos não-humanos.

Algo que é difícil de delimitar é o que diferencia a incerteza moral da incerteza empírica? Afinal, como vimos no exemplo acima, uma parece fazer parte da outra, o que levanta a questão de se existe uma diferença entre o *agir moral* e o *agir racional* de um ponto de vista pragmático, ou seja, qual seria a diferença entre racionalidade e moralidade? Questões que indagam o valor moral de animais, e outras questões de incerteza moral, podem ser desconstruídas em diferentes perguntas, empíricas e morais, por exemplo:

1. Quais as evidências científicas que animais possuem senciência?
2. Deveríamos atribuir valor moral a seres sencientes?

Poderíamos assim definir duas formas de agente da seguinte forma:

a) *Agente Racional*: o agente racional busca escolher suas ações com base em suas preferências e crenças, a fim de maximizar (escolher a alternativa *me-*

ISBN: 2359-1951

nos pior do que todas as outras) sua utilidade (valor subjetivo atribuído a certas ações ou resultados);

b) *Agente Moral*: o agente moral busca escolher ações com base em suas preferências, utilizando como referência as teorias morais nas quais o agente atribui credibilidade (grau de crença), a fim de maximizar, ou seja, *prefere escolher ações certas do que ações erradas*, seu valor (valor subjetivo atribuído a certas ações ou resultados, relativo às teorias morais credibilizadas).

Se definirmos racionalidade e normatividade da forma que fizemos acima, podemos perceber que existe uma forte semelhança entre o agir racional e o agir moral, pois ambos empregam o conceito de preferência (alguma relação comparativa que diferencie o que é "*melhor*" do que é "*pior*"), e escolha, ou seja, o conceito de intenção/ação. Em situações de incerteza empírica, possuímos um quadro bastante robusto sobre como atualizar nossas hipóteses (inferência Bayesiana) e tomar decisões (maximização de utilidade esperada). Assim, qual seria o análogo para casos de incerteza moral? MacAskill e Ord (2018) e Greaves e Cotton-Barratt (2019) sugerem que a distinção entre conceitos como "*racionalidade*" e "*moralidade*" seja apenas uma disputa *semântica de análise contextual*, onde ambas as interpretações de o que "*deve*" ser feito, acabam por alcançar as mesmas formas de conclusão, não havendo então uma diferença pragmática entre a racionalidade e a moralidade. Talvez em diferentes contextos o significado dos termos seja de fato diferente, mas em uma abordagem prática em tomadas de decisão sob incerteza moral, acredito que uma equivalência entre ambos os conceitos pode ser alcançada (especialmente, mas não apenas, em modelos éticos consequencialistas). E se encararmos a questão de racionalidade normativa como uma questão de *racionalidade*, evitamos ter que definir o quê de fato seria uma "*ação moral*", e por consequência importamos toda teoria de escolha racional, que como argumentamos antes, já possui um robusto modelo sobre tomadas de decisão em situação de incerteza.

A teoria da utilidade esperada de *von Neumann-Morgenstern* (1944), primeiramente introduzida por Daniel Bernoulli em 1738 com o *Paradoxo de São Petersburgo*, define como racional, o comportamento de um agente que "*segue o que deve ser feito*" de acordo com as suas crenças e preferências, tentando maximizar seu ganho, dada as constrições do ambiente e sua própria constituição. Dessa forma, a teoria da utilidade

ISBN: 2359-1951

esperada, por mais que existam muitas interpretações do seu significado, procura definir como um agente racional *deve* agir (de forma coerente e transitiva) para cumprir suas preferências, dada as suas crenças.

Temos então implicitamente uma forma de normatividade na teoria racional, algo que em nossa opinião apenas enfraquece a suposta divisão entre racionalidade e moralidade. Em uma situação de incerteza moral, para aplicarmos o conceito de racionalidade, não precisamos definir nenhuma posição como correta ou errada, apenas precisamos assumir que o agente moral (racional) quando em um estado de incerteza moral procura realizar a ação mais correta possível, de acordo com as teorias morais que o agente credencia. Ao mesmo tempo, o agente moral é indiferente a ações corretas (ou incorretas) que possuem o mesmo valor moral, e consequentemente é capaz de ordenar ações morais de magnitude diferente (matar 100 agentes é duas ordens de magnitude pior do que matar 1 pessoa). Se tais premissas, já defendidas por diversos teóricos (BYKVIST, 2017; LOCKHART 2000; SEPIELLI, 2009; ROSS, 2006), podem ser aceitas, então os princípios teóricos da teoria da escolha racional podem ser aplicados a estratégias meta normativas para lidar com situações de incerteza moral. Críticos desta visão consideram o "dever" como algo puramente moral (HARMAN, 2015; WEATHERSON, 2002), contudo, teóricos que defendem esta posição não fornecem uma maneira de como decisões sobre incerteza moral podem ser feita *apenas* com princípios morais.

Existe o argumento de que a moralidade é algo *transcendente* ao pensamento racional, geralmente expressado pela guilhotina de Hume. De fato, a "*verdadeira última e universal*" noção de moralidade seja algo transcendente a qualquer agente racionalmente limitado, contudo, isto não implica que devemos abandonar nossa investigação moral e suas aplicações práticas. Se a pergunta de "*o que se deve fazer quando se está em dúvida sobre o que fazer?*" for considerada totalmente transcendental e inalcançável então por que nos preocupar com moralidade em primeiro lugar? Gostaríamos de propor uma analogia em defesa da aplicabilidade de estratégias meta normativas para resolução de conflitos morais, em defesa de uma ética pragmática: π é um número transcendental, π não é algébrico e não é a solução para nenhuma equação polinomial. Se um carpinteiro, ao tentar construir uma roda para uma carroça perguntar "*qual valor*

ISBN: 2359-1951

numérico eu preciso para calcular o comprimento da circunferência da minha roda?" for respondido com algo como "*isto é um conhecimento transcendental, não há uma maneira de eu lhe dizer* π *em uma forma finita*", essa resposta não vai ajudar o carpinteiro a construir a sua roda. Mas um matemático ou um engenheiro mais empático poderia responder "*eu posso lhe dar uma aproximação boa o suficiente para o seu problema*", o que, ao meu ver, é uma resposta muito mais prática. Qualquer aproximação, 3.141, é melhor do que nenhuma resposta, de mesma forma, estratégias meta normativas são formas práticas de conseguir soluções aproximadas para problemas morais possivelmente intratáveis no ponto de vista ideal. Se tal decisão é "*mesmo a resposta correta*" (algo reminiscente do argumento da *pergunta aberta de Moore*), talvez seja em qualquer sentido pragmático algo *sem sentido*.

Estratégias Meta normativas

Como argumentamos na última sessão, a diferença entre incerteza empírica e moral pode ser confusa, e enquanto a tomada de decisão sobre incerteza empírica é bem fundamentada na literatura, o mesmo não pode ser dito sobre a incerteza moral, sendo uma área muito menos explorada. A questão que queremos abordar é: o que fazer quando não se sabe o que fazer? Ou seja, como agir em situações de incerteza moral (também podemos chamar de incerteza normativa). Haveria alguma estratégia meta normativa, de forma que o agente possa avaliar diferentes teorias morais, para chegar na melhor decisão possível. A primeira estratégia meta normativa que citamos é a "*Minha Teoria Favorita*" (GUSTAFSSON,TORPMAN, 2014). Para exemplificá-la, imaginemos o seguinte problema:

Ana se encontra em um dilema moral. Ana está indecisa se deve comprar um pastel de carne ou um pastel de queijo, e Ana possui um diferente grau de crença em diferentes teorias morais. A primeira teoria (T_1), uma forma de utilitarismo que avalia a "unidade de utilidade" ($U\$$) de humanos como 10 vezes mais valiosa do que a unidade de utilidade de bovinos ($U\$_{humanos:bovinos}\,10:1$). O grau de crença de Ana em T_1 é 30% (o valor em unidades de utilidade para Ana do pastel de carne é $10\,U\$$ e o pastel de queijo é $5U\$$). Enquanto isso, Ana possui um grau de crença de 70% em uma outra versão do

Clareira
Revista de Filosofia da Região Amazônica
ISBN: 2359-1951

utilitarismo (T_2) que valoriza a unidade de utilidade humana para pastéis de carne/queijo da mesma forma ($10\,U\$, 5\,U\$$), mas não atribui nenhum valor moral a vida dos bovinos. De acordo com T_1, a morte de um bovino é avaliada como $-1000\,U\$$, como a comparação entre a utilidade humana e a utilidade bovina é $10:1$, isso causa $-100\,U\$$ pontos de utilidade para Ana ($+10\,U\$$ do pastel de carne ¿$-90\,U\$$). Já de acordo com a teoria T_2, Ana apenas precisa escolher entre o pastel de queijo ($5\,U\$$) e o pastel de carne ($10\,U\$$), pois, o valor da vida bovina não é considerado. O que Ana deve fazer? O valor de escolha de cada opção, de acordo com cada teoria, está resumida na tabela a seguir:

	$T_1 - 30\,\%$	$T_2 - 70\,\%$
Pastel de Carne	-90	10
Pastel de Queijo	5	5

Minha Teoria Favorita (MTF) propõe que façamos nossa escolha baseada na teoria moral que possuímos maior grau de crença, dessa forma, no dilema acima Ana utilizando a estratégia meta normativa MTF escolheria a teoria moral T_2, e compraria o pastel de carne, já que está é a opção a qual suas unidades de utilidade (bem-estar) são maximizadas de acordo com suas preferências e grau de crença. Contudo, não podemos fazer melhor que isso? Como Ana deveria agir caso o grau de crença em T_1 e T_2 seja igual? Em situações onde o grau de crença é distribuído igualmente entre as teorias morais em consideração, MTF não nos fornece uma solução satisfatória. Pressupomos que a opção "*jogar uma moeda justa*" não seria uma atitude moral nem racional (imagine Ana atendendo a protestos pelos direitos dos animais na segunda-feira e carneando um porco na terça-feira). MTF também recomenda o indivíduo a tomar decisões "moralmente arriscadas" quando seu grau de crença é dividido de forma quase indiferente, por exemplo: se Ana possui um grau de crença de 49% em T_1 e 51% em T_2, MTF ainda recomenda o pastel de carne, por mais que 49% da credibilidade normativa de Ana esteja comprometida a uma penalidade quase 10 vezes maior do que o ganho de um pastel de carne. Poderíamos criar situações onde a penalidade seria 1000 vezes maior, MTF ainda não levaria em consideração tal diferença de magnitude de valor.

A meu ver, soluções melhores que a MTF foram propostas, como a abordagem teórica de negociação, de Greaves e Cotton-Barratt (2019), e o *Moral Hedging*

Clareira
Revista de Filosofia da Região Amazônica
ISBN: 2359-1951

(HICKS, 2019), contudo, neste estudo iremos explorar uma das propostas de William MacAskill (2014), conhecida como *Maximização da Escolha-Valiosa Esperada* (MEV). Diferente de MTF, o MEV é uma abordagem *comparativista*, comparativismo sendo a visão de que a tomada de decisão do agente moral não deve ser baseada apenas na credência a diferentes teorias morais, mas também ao grau de valorização-de-escolha que as teorias atribuem a diferentes ações, MTF sendo uma estratégia meta normativa não-comparativa. MacAskill define diferentes tipos de teorias morais da seguinte forma:

A) *Teorias Morais Cardinais*: teorias morais são cardinalmente mensuráveis se além de uma ordem de preferência, "*o que é melhor que o que*", a teoria pode dizer o *quanto* uma preferência é melhor do que a outra. Ou seja, além de dizer que $A \geq B$, a teoria diz o quanto que A é melhor, através de um quantificador de valor γ:

$$T_i = \{\gamma A \geq B, \gamma B \geq C, \gamma C \geq D, \ldots\}$$

Por exemplo: $\gamma A \geq B$, onde γ significa que A é 100 unidades de valor melhor que B. Teorias morais consequencialistas, como o utilitarismo, são exemplos de teorias morais Cardinais.

B) *Teorias Morais Ordinais*: teorias morais são ordinais se elas apenas apresentam uma relação de preferência ordinal, ou seja:

$$T_i = \{A \geq B \geq C \geq D, \ldots\}$$

Por exemplo: uma teoria moral ordinal, como alguma versão deontológica do Kantianismo pode ditar que "*mentir é errado*", contudo, tal teoria não nos diz o quanto mentir é melhor do que outra ação, apenas nos fornece um ordenamento de preferências;

$$Sistema\ Deontológico_i = \{mentir \geq roubar \geq agredir \geq matar \ldots\}$$

Teorias morais deontológicas são geralmente teorias morais ordinais.

Podemos dizer que teorias morais cardinais são as que nos fornecem mais informação, pois, além de uma classificação ordenada de preferências, elas nos disponibilizam uma magnitude comparativa entre preferências, enquanto teorias morais ordinais são os modelos menos informativos. O melhor tipo de situação em uma tomada de decisão sob incerteza moral é quando temos de comparar diferentes teorias morais que *são* cardinais e *são* interteóricamente comparáveis. É quando possuímos maior quantidade de informação, e tal informação pode ser comparada entre as teorias morais sendo avaliadas. Iremos explorar na próxima sessão como isto pode ser feito.

Clareira
Revista de Filosofia da Região Amazônica
ISBN: 2359-1951

Maximização da Escolha-Valiosa Esperada

Para resolver o problema de decisão sob incerteza moral MacAskill (2014) recomenda o seguinte método:

a) *Maximização da Escolha-Valiosa Esperada* (MEV): o MEV é utilizado se todas as teorias morais em consideração pelo agente forem *teorias cardinais* e *interteóricamente comparáveis*;

O valor do MEV de alguma ação é o grau de crença do decisor, em uma teoria moral T_i, multiplicado pelo valor moral (no utilitarismo chamamos de "utilidade") de uma certa ação. No caso ideal, onde as teorias avaliadas são todas cardinais e interteoricamente comparáveis, o valor da escolha-valiosa esperada de uma ação, EV(A), é dado da seguinte forma:

$$EV(A)=\sum_{i=1}^{n} C(T_i)\,VE_i(A)$$

onde, $C(T_i)$ representa a credibilidade (grau de crença) do tomador de decisão em T_i (alguma teoria moral particular), enquanto que $CW_i(A)$ representa o "valor da escolha", de acordo com T_i, de A (uma ação que o tomador de decisão pode escolher). Vamos utilizar novamente o exemplo de Ana em sua escolha entre comprar um pastel de carne ou um pastel de queijo, dividida entre duas teorias morais diferentes, T_1 (30% de grau de crença) e T_2 (70% de grau de crença), que possuem opiniões diferentes sobre o valor moral da vida animal. Os valores morais atribuídos a cada ação estão descritos novamente abaixo:

	$T_1-30\%$	$T_2-70\%$
Pastel de Carne	-90	10
Pastel de Queijo	5	5

Utilizando o MEV chegamos no seguinte resultado:

$$EV(Pastel\,de\,Carne)=(0.3\times-90)+(0.7\times10)=-20$$
$$EV(Pastel\,de\,Queijo)=(0.3\times5)+(0.7\times5)=5$$

ISBN: 2359-1951

Assim, de acordo com o MEV, Ana deveria comprar um pastel de queijo. MEV parece ter um resultado melhor do que teorias como MTF, sendo uma forma de agregar várias teorias normativas, não importando quantas sejam (o exemplo acima poderia ser feito com, a princípio, *n* teorias morais, tantas quanto o agente se importar, ou for capaz, de manter em mente). Outra vantagem do modelo MEV é que ele é aplicável mesmo quando temos incerteza em relação à distribuição de probabilidades, servindo como uma espécie de "regra padrão" ou heurística. Por exemplo: imaginemos que Bruno está em um impasse moral entre colar em sua prova final de semestre ou não. Bruno está dividido entre duas teorias morais, ele possui bastante credibilidade em uma forma de utilitarismo que avalia a ação de "colar" como positiva, e, ao mesmo tempo, pouca credibilidade em outra teoria moral consequencialista que avalia "colar" como definitivamente errado. Bruno pode considerar que colar aumentaria seu bem-estar, contudo, enquanto o ato de colar apenas possui um aumento pequeno na *escolha-valiosa esperada* (pequeno valor positivo multiplicado por uma grande probabilidade), o ato de colar para a teoria moral consequencialista possui uma alta "magnitude" moral (baixa probabilidade multiplicada por um alto valor negativo). Talvez a melhor alternativa de acordo com MEC no caso de Bruno seja simplesmente estudar.

Integrando o MEV com a Teoria da Utilidade Esperada

O formalismo criado por MacAskill (2014) é extremamente similar ao formalismo da teoria de escolha racional, o que nos permite a integração dos dois modelos em um único modelo, capaz de auxiliar o decisor em situações de incerteza empírica moral. De fato, como discutimos na seção "*Incerteza Moral e Incerteza Empírica*" a diferença entre ambos os conceitos pode ser vaga e difícil de esclarecer, especialmente se tentarmos definir uma outra maneira além da racionalidade para decisões sobre incerteza. Afinal, o que seria um agente moral, se não um agente racional que toma suas decisões de acordo com suas crenças normativas e preferências? O que um decisor racional e normativo pode fazer em situações de incerteza é dissolver o problema em seus componentes empíricos e morais. Por exemplo: no exemplo em que Ana precisava escolher entre um pastel de carne e um pastel de queijo, diversas formas de incerteza empírica podem ser adicionadas ao problema, como:

Clareira
Revista de Filosofia da Região Amazônica
ISBN: 2359-1951

a) Qual a probabilidade, se Ana deixar de comprar o pastel de carne, de que isto terá um efeito positivo (menos sofrimento animal) no ambiente? Talvez a falta de consumo cause situações ainda piores para bovinos em cativeiro.
b) O quão certa Ana está sobre o fato de que de mamíferos de grande porte, como vacas e bois, possuem senciência?
c) Será que o baixo consumo de carne pode ter influências negativas piores ainda a vidas humanas (que a princípio, de acordo com T_1, possuem um valor moral 10 vezes maior do que o valor de bovinos)?
d) O quanto o consumo do queijo, que é um derivado do leite (provavelmente leite bovino), prejudica a vida animal?

Todos estes questionamentos podem auxiliar na tomada de decisão de Ana, pois, algumas das conclusões destes fatos podem auxiliar a mudar o grau de crença de Ana em relação a uma teoria moral que atribui um maior valor a vida animal, ou não. Ou seja, a capacidade do agente adquirir mais informação, *praticar atualização epistêmica* e *reflexão normativa em situações de incerteza*, auxilia o agente a restringir o espaço de teorias morais àquelas que melhor representam o mundo real. Assim, integramos o modelo MEV com a teoria da utilidade esperada da seguinte forma:

$$EV(A)=\sum_{i=1}^{n}\sum_{j=1}^{n}P(O_j \mid A)V E_i(O_j)C(T_i)$$

Onde, $EV(A)$ é o valor da escolha-valiosa esperada de uma ação A, $C(T_i)$ é a credibilidade do decisor na teoria moral T_i. E agora ao invés de valorizarmos a ação, valorizamos a observação $V E_i(O_j)$, ou seja, a consequência $P(O_j \mid A)$ da ação A de acordo com T_i. Agora possuímos um modelo que unifica incerteza empírica e incerteza moral, onde para se encontrar a ação com maior valor de escolha-valiosa esperada, o agente agora também avalia cada resultado possível para a ação tomada (por exemplo: a compra do pastel de carne aumentar ao invés de diminuir o sofrimento de bovinos), multiplicando pelo valor que o resultado traria (de acordo com a teoria moral T_i), multiplicado pela credibilidade do decisor na teoria moral T_i, somando o resultado de cada ação de acordo com cada teoria moral sendo credibilizada. Voltemos agora ao exemplo de Ana e seus pastéis: Ana possui a mesma distribuição de credibilidade entre as teorias morais do primeiro exemplo ($T_1=30\%$ e $T_2=70\%$), enquanto a teoria T_1 atribui valor moral a bovinos a uma taxa $10:1$ ($U\$_{humanos:bovinos}\,10:1$), a teoria T_2 não atribui valor

Clareira
Revista de Filosofia da Região Amazônica
ISBN: 2359-1951

moral a bovinos. Em T_1, a escolha de comer pastel de carne causa $-1000\,U\$$ em Ana, e ambas as teorias garantem $10\,U\$$ pelo pastel de carne e $5U\$$ pelo pastel de queijo. Ana acredita com 80% de credibilidade que a compra de carne aumenta o sofrimento animal, e 20% de chance de que a compra de queijo leva ao mesmo resultado. O que Ana deve fazer? De acordo com o MEV integrado com a Teoria da utilidade esperada:

$$EV\left(Pastel\,de\,carne\right)=\left(0.8\times-90\times0.3\right)+\left(0.8\times0\times0.7\right)=-21.6$$
$$EV\left(Pastel\,de\,queijo\right)=\left(0.2\times-95\times0.3\right)+\left(0.2\times0\times0.7\right)=-5.7$$

Novamente o modelo recomenda Ana a escolher o pastel de queijo, o menor dos dois males. Da mesma forma que fizemos com o modelo MEV, esta extensão pode ser ampliada para calcular diversas incertezas empíricas, e também pode ser aplicada de maneira heurística. Se o agente for capaz de atribuir qualquer noção ingênua de probabilidade, podemos chegar a conclusão de que, de acordo com o modelo meta normativo de MacAskill, um pequeno risco em fazer uma grande "*maldade*" não justifica o pequeno ganho de utilidade de um "*pastel de carne*".

Conclusão

Seres humanos estão sempre sujeitos a incerteza, seja empírica ou moral. O que o método de William MacAskill nos fornece é uma maneira de resolver problemas de decisão sob incerteza moral, um modelo meta normativo para agregarmos preferências. Além do método de Maximização da Escolha-Valiosa Esperada, MacAskill também fornece métodos para lidar com a comparação de teorias morais que *não* são interteóricamente comparáveis: *Voto por Variância*, quando as teorias morais avaliadas são todas cardinais, mas não são interteóricamente comparáveis, e *Regra Borda*, para quando comparamos teorias cardinais e ordinais. Não revisei todos os métodos, pois, o estudo ficaria muito longo, contudo, garanto ao leitor que em breve publicarei uma versão expandida deste estudo, explicando (com exemplos) todos os métodos do modelo meta normativo de MacAskill, e quais pressupostos estão por trás dele. A similaridade do MEV com a teoria da utilidade esperada permite uma integração simples e intuitiva de ambas as metodologias. Acredito que se resultados como estes podem ser alcançados em um quadro de decisão racional, onde ambas as incertezas empíricas e morais podem

ISBN: 2359-1951

ser unidas e dissolvidas no processo de decisão, podemos definir da seguinte forma o conceito de “imoralidade”, ou, “comportamento não-normativo”:

Agentes podem distribuir sua credibilidade (grau de crença) entre teorias morais como quiserem, e teorias morais podem ordenar preferências e valores de ações/consequências de qualquer forma (pluralidade moral). Contudo, se o decisor encontra-se em uma situação de incerteza moral, e após atualizar seus valores de escolha opta por uma ação/consequência menos valiosa do que outra ação/consequência disponível de acordo com as teorias morais que possui algum grau de crença, o agente não é normativo, ou seja, não podemos atribuir valores de preferência e grau de crença a sua escolha de forma consistente.

Agradecimentos

Gostaria de agradecer ao Programa de Excelência Acadêmica (Proex) da Fundação CAPES (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior), e ao Programa de Pós-Graduação em Filosofia da Pontifícia Universidade Católica do Rio Grande do Sul, Brasil.

Referências

AMODEI, D, OLAH, C, STEINHARDT, J, CHRISTIANO, P, SCHULMAN, J. MANÉ, D. *Concrete Problems In AI Safety*. [online] arXiv.org, 2016. Disponível em: https://arxiv.org/abs/1606.06565. Acessado em 19 de Agosto de 2020.

BYKVIST, K. *Moral uncertainty*. Philosophy Compass, 12(3), 2017. Disponível em: https://doi.org/10.1111/phc3.12408

CHALMERS, D. *The singularity: A philosophical analysis*. Journal of Consciousness Studies 17 (9-10), pp. 9 – 10, 2010.

CROWLEY, J. *Woman Says Amazon's Alexa Told Her To Stab Herself In The Heart For “The Greater Good”*. [online] Newsweek, 24 de dezembro de 2019. Disponível em: https://www.newsweek.com/amazon-echo-tells-uk-woman-stab-herself-1479074. Acessado em 19 de Agosto de 2020.

GREAVES, H. COTTON-BARRATT, O. *A bargaining-theoretic approach to moral uncertainty*. 2019. Disponível em: http://users.ox.ac.uk/~mert2255/papers/bargaining-theory-and-MU.pdf. Acessado em 19 de Agosto de 2020.

ISBN: 2359-1951

GUSTAFSSON , J. E. TORPMAN, O. *In Defence of My Favourite Theory*. Pacific Philosophical Quarterly, 95(2), pp. 159-174, 2014. Disponível em: https://doi.org/10.1111/papq.12022. Acessado em 19 de Agosto de 2020.

GUTERRES, A. *Transcript of the Secretary-General's virtual press encounter on the appeal for global ceasefire*. United Nations Secretary-General, Statements/Reports. 23 March 2020. Disponível em: https://www.un.org/sg/en/content/sg/press-encounter/2020-03-23/transcript-of-the-secretary-generals-virtual-press-encounter-the-appeal-for-global-ceasefire. Acessado em 19 de Agosto de 2020.

HAGENDORFF, T. *The Ethics of AI Ethics: An Evaluation of Guidelines*. Minds and Machines, 39, pp. 99-120, 2020. doi:10.1007/s11023-020-09517-8.

HARMAN, E. *The Irrelevance of Moral Uncertainty*. In Oxford Studies in Metaethics. 10, pp. 53-79, 2015. Disponível em: http://www.princeton.edu/~eharman/documents/UncorrectedProofsIrrelevanceUncertainty.pdf. Acessado em 19 de Agosto de 2020.

HICKS, A. *Moral Hedging and Responding to Reasons*. Pacific Philosophical Quarterly 100 (3), pp.765-789, 2019. Disponível em: https://philarchive.org/archive/HICMHA. Acessado em 19 de Agosto de 2020.

IASON, G. *Artificial Intelligence, Values, and Alignment*. DeepMind, [online] arXiv.org, 2020. Disponível em: https://arxiv.org/abs/2001.09768v1. Acessado em 19 de Agosto de 2020.

JOBIN, A, IENCA, M, VAYENA, E *The global landscape of AI ethics guidelines*. Nat Mach Intell 1, pp. 389–399, 2019. Disponível em: https://doi.org/10.1038/s42256-019-0088-2. Acessado em 19 de Agosto de 2020.

JURIĆ, M. ŠANDIĆ, A. BRCIC, M. *AI safety: state of the field through quantitative lens*. 2020. Disponível em: https://arxiv.org/ftp/arxiv/papers/2002/2002.05671.pdf. Acessado em 19 de Agosto de 2020.

LOCKHART, T. 2000. *Moral Uncertainty and Its Consequences*. Oxford University Press. ISBN: 9780195126105

MACASKILL, W. *Normative Uncertainty*. Thesis for the degree of Doctor of Philosophy. St Anne's College, University of Oxford, February 2014. Disponível em: http://commonsenseatheism.com/wp-content/uploads/2014/03/MacAskill-Normative-Uncertainty.pdf. Acessado em 19 de Agosto de 2020.

MACASKILL, W. ORD, T. *Why Maximize Expected Choice-Worthiness*?. Noûs, 12264, pp. 1–27, 2018. doi: 10.1111/nous

ISBN: 2359-1951

ROSS, J. *Rejecting Ethical Deflationism*. Ethics, 116, pp. 742–768, 2006. doi: 10.1086/505234. Disponível em: https://www.jstor.org/stable/10.1086/505234. Acessado em 19 de Agosto de 2020.

SEPIELLI, A. *What to Do When You Don't Know What To Do*. IN Oxford Studies in Metaethics, edited by Russ Shafer-Landau, 35, 2009. Oxford University Press.

von NEUMANN, J. Morgenstern, O. *Theory of Games and Economic Behavior*. 1st ed. Princeton, NJ: Princeton University Press, 1944.

WEATHERSON, B. *Review of Ted Lockhart's "Moral Uncertainty and Its Consequences"*. Mind, 111, pp. 693–696, 2002.

WOLF, M. J. MILLER, K. GRODZINSKY, F. S. *Why we should have seen that coming: comments on microsoft's tay experiment, and wider implications*. ACM SIGCAS Computers and Society 47(3), pp. 54–64, 2017.