View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PhilPapers

PUCRS

ESCOLA DE HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA
DOUTORADO EM FILOSOFIA

PAULO ANTÔNIO CALIENDO VELLOSO DA SILVEIRA

ÉTICA E INTELIGÊNCIA ARTIFICIAL:
da possibilidade filosófica de agentes morais artificiais

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

Porto Alegre
2020

PAULO ANTÔNIO CALIENDO VELLOSO DA SILVEIRA

ÉTICA E INTELIGÊNCIA ARTIFICIAL: da possibilidade filosófica de agentes morais artificiais

Tese de Doutorado apresentada como requisito para a obtenção do grau de Doutor pelo Programa de Pós-Graduação em Filosofia da Escola de Humanidades da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Draiton Gonzaga de Souza

Porto Alegre
2020

PAULO ANTÔNIO CALIENDO VELLOSO DA SILVEIRA

ÉTICA E INTELIGÊNCIA ARTIFICIAL: da possibilidade filosófica de agentes morais artificiais

Tese de Doutorado apresentada como requisito para a obtenção do grau de Doutor pelo Programa de Pós-Graduação em Filosofia da Escola de Humanidades da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovada em: _____ de _____ de 2021.

BANCA EXAMINADORA

Prof. Dr. Dr. Draiton Gonzaga de Souza (Orientador)

Prof. Dr. Agemir Bavaresco

Prof. Dr. Thadeu Weber

Prof. Dr. Carlos Alberto Molinaro

Profa. Dra. Gabrielle Bezerra Sales Sarlet

Porto Alegre
2020

Ficha Catalográfica

C153e Caliendo V da Silveira, Paulo Antônio

Ética e Inteligência Artificial : da possibilidade filosófica de agentes morais artificiais / Paulo Antônio Caliendo V da Silveira. – 2021.

145f.

Tese (Doutorado) – Programa de Pós-Graduação em Filosofia, PUCRS.

Orientador: Prof. Dr. Draiton Gonzaga de Souza.

1. Ética. 2. Inteligência Artificial. 3. Teorias morais. 4. Agentes Morais Artificiais. I. Souza, Draiton Gonzaga de. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Para Juliana, Sílvia, Draiton e Ingo.

AGRADECIMENTOS

A presente Tese não seria possível sem a confiança, orientação e exemplo acadêmico do Professor Dr. Dr. Draiton Gonzaga de Souza. Os seus ensinamentos, sugestões e, principalmente, guia de conduta são uma bela referência para qualquer pesquisador.

Agradeço ao Prof. Dr. Cláudio de Almeida pelas fantásticas aulas sobre Epistemologia Analítica, que ilustraram e orientaram muitos dos meus estudos nesses anos de pesquisas. Ao Professor Dr. Nythamar de Oliveira pelas aulas em Neurofilosofia e Filosofia Judaica, que contribuíram muito para o abrir minha visão para perspectivas novas e desafiadoras. Ao Prof. Dr. Roberto Pich pelos estudos sobre Metafísica e Filosofia Medieval, talvez a área mais impactante na minha trajetória de estudos. Ao Professor Dr. Thadeu Weber pelos estudos em Ética, que descortinaram temas, desafios e métodos claros de análise. Foi um privilégio estudar sobre Kant e Hegel com um dos maiores conhecedores destes autores no país.

Agradeço ao Prof. Dr. Ingo Sarlet pelo apoio na realização deste meu segundo Doutoramento, sem me afastar das minhas atividades docentes. Ao grupo de professores e colegas do PPGD da PUCRS e aos membros da Comissão Coordenadora do PPGD, na qual participei durante o tempo de redação da Tese. Aos servidores e funcionários da Escola de Direito, especialmente, saúdo em nome da amiga Caren Klinger, que torna nossa tarefa sempre mais aprazível e leve.

Agradeço ao Prof. Dr. Rafael Bordini, da Escola Politécnica e membro do PPG em Ciência da Computação, pelos anos de trabalho e pesquisas sobre Inteligência Artificial e Direito. Os debates e informações técnicas sobre o difícil assunto foi de muita valia e contribuição. Foi uma honra ter dialogado com uma das maiores autoridades internacionais sobre sistemas multiagentes. Agradeço igualmente todo o grupo de pesquisas em torno do Projeto PRAIAS, sobre e IA e Direito, especialmente, à Débora Engelmann, Tabajara Krausburg, Olimar Borges, Bruna Lietz e Marcelo Pasetti.

Agradeço aos meus colegas de estudo em Filosofia, no PPG da PUCRS, especialmente, Laura Nascimento, João Fett, Gregory Gaboardi, Samuel Cibils, Renata Floriano, Cezar Roedel, Ricardo Nüske e André Neiva. O prazer do debate, a troca de opiniões, os ensinamentos em epistemologia, em metafísica e filosofia medieval foram enriquecedores.

Agradeço muito, especialmente, a minha amada Juliana Damásio. Pela parceria na vida e pelo exemplo de pesquisadora, doutoranda, professora em Ciência da Computação, mas, principalmente, pelo modelo ético de boa pessoa.

Agradeço, especialmente, a Deus pela graça da vida, da saúde e da amizade.

RESUMO

A presente Tese pretende verificar e assumir a possibilidade filosófica do surgimento de um agente moral artificial autêntico. Toma-se como pressuposto a plausibilidade da superação do *Teste de Turing*, da *Sala Chinesa* e do *Teste de Ada Lovelace*, bem como da possível emergência de um autêntico agente artificial moral, com deliberações intencionais em perspectiva de primeira pessoa. Assim, aceita-se a tese da possibilidade de um código computacional capaz de dar origem à emergência.

O problema principal deste estudo será investigar a possibilidade filosófica de uma ética artificial, como decorrente da vontade e racionalidade *própria* de um sujeito artificial, ou seja, *da inteligência artificial como sujeito moral*.

Um agente ético artificial deve agir por características próprias e não conforme uma programação externa predeterminada. A ética artificial autêntica é interna e não externa ao autômato. Um modelo proposto e com crescente aceitação, e que demonstra essa possibilidade computacional, é o de uma moralidade que se constrói *de baixo para cima* (*bottom-up*), e nesse caso o sistema pode passar a adquirir capacidades morais de modo independente. Esse modelo se aproxima da *ética aristotélica das virtudes*. Outra forma possível é a união de um modelo computacional de piso, com modelos fundados na deontologia, com a formulação mais geral de deveres e máximas. De uma outra forma, demonstra-se que pelo menos em um caso é possível a construção de um modelo de moralidade artificial viável e autônomo.

Não há demonstração inequívoca da impossibilidade de os agentes morais artificiais possuírem emoções artificiais. A conclusão a que diversos cientistas de programação chegaram é que um modelo de agência artificial fundado em *machine learning*, combinado com a *ética da virtude*, é um caminho natural, coeso, coerente, integrado e "bem costurado" (*seamless*). Assim, existe uma resposta coerente, consistente e bem fundamentada que *indica que não é provada a impossibilidade de um agente moral artificial autêntico.*

Por fim, uma teoria ética responsável deve considerar a possibilidade concreta do surgimento de *agentes morais artificiais completos* (*full moral agent*) e todas as consequências desse fenômeno divisor na história da humanidade.

ABSTRACT

This Dissertation intends to verify and assume the philosophical possibility of the emergence of an authentic artificial moral agent. The plausibility of overcoming the Turing Test, the Chinese Room and the Ada Lovelace Test is taken as an assumption, as well as the possible emergence of an authentic artificial moral agent, with intentional deliberations from a first person perspective. Thus, the assumption of the possibility of a computational code capable of giving rise to the emergency is accepted.

This study's main problem will be to investigate the philosophical possibility of an artificial ethics, as a result of the will and rationality of an artificial subject, that is, of artificial intelligence as a moral subject.

An artificial ethical agent must act on its own characteristics and not according to a predetermined external schedule. Authentic artificial ethics are internal and not external to the automaton. A model proposed and with increasing acceptance, which demonstrates this computational possibility, is that of a morality that is built from bottom-up, in which case the system can start to acquire moral capacities independently. This model comes close to the Aristotelian ethics of virtues. Another possible way is the union of a computational floor model, with models based on deontology, with the most general formulation of duties and maxims. In another way, it is demonstrated that it is possible to build a viable and autonomous model of artificial morality in at least one case.

There is no clear demonstration of the impossibility for artificial moral agents to have artificial emotions. The conclusion reached by several programming scientists is that a model of artificial agency based on machine learning, combined with the ethics of virtue, is a natural, cohesive, coherent, integrated and seamless. Thus, there is a coherent, consistent, and well-founded answer that indicates that an authentic artificial moral agent's impossibility has not been proven.

Finally, a responsible ethical theory must consider the concrete possibility of the emergence of complete artificial moral agents and all the consequences of this dividing phenomenon in human history.

SUMÁRIO

INTRODUÇÃO

O presente trabalho pretende verificar a possibilidade filosófica da existência de autênticos agentes morais artificiais. A importância do tema é cada vez mais relevante pelo destaque que a inteligência artificial (IA) tem assumido em todos os campos da vida humana, seja no direito, na economia, na política ou na cultura.

A disseminação crescente do uso de mecanismos de IA tem despertado diversos questionamentos ainda sem respostas. Os impactos éticos de seu uso, por exemplo, em situações dramáticas na medicina, no julgamento por máquinas, no uso de carros autônomos ou em drones militares são indiscutíveis e não podem ser negligenciáveis. O primeiro passo para responder a essas questões está em firmar um conceito filosófico de inteligência artificial, o qual deve verificar a real possibilidade de existência, futura ou hipotética, de as máquinas serem inteligentes e não apenas aparentar imitar a inteligência humana.

Inicialmente, a suposição filosófica de haver inteligência artificial similar à humana parece desarrazoada ou mesmo petulante. Realmente é difícil supor que trôpegos robôs ou algoritmos lacunosos, e geralmente falhos, possam dar azo a um ser com pretensões de perfeição, à semelhança dos indivíduos na Terra. Obviamente se trata de uma indagação que conversa com o futuro da tecnologia e da humanidade em uma similar e desapaixonada reflexão que poderia ser realizada no âmbito da bioética ou neurociência. De modo geral, a literatura sobre o assunto apresenta sérias e fundamentadas críticas ao argumento da possibilidade de uma IA forte realmente emergir. Os ataques demolidores de *Searle* contam já quase meio século e ainda estão fortes e vigorosos. Os riscos envolvidos nessa possibilidade, contudo, ainda que remota, exigem a responsabilidade de uma reflexão aberta e criteriosa.

A tradição filosófica sobre o tema apresentou respostas diametralmente opostas a essa hipótese filosófica. Para *Descartes* existiriam diferenças importantes a impedir esse surgimento. *Ada Lovelace* questionará a possibilidade de essa inteligência ser criadora. *Searle* irá duvidar de uma autêntica inteligência artificial, no sentido próprio, como manifestação em primeira pessoa. O presente trabalho, nesse sentido, irá verificar a evolução histórica e conceitual do problema, conforme os principais proponentes, e suas concepções sobre a viabilidade de uma máquina com inteligência similar à humana. O trabalho terá como objetivo demonstrar a possibilidade de superação dos limites firmados por *Descartes, Ada Lovelace, Turing e Searle*.

Não serão objeto de análise, as críticas negacionistas sobre a impossibilidade filosófica de uma autêntica inteligência forte por parte de teorias relevantes, tais como perspectivas

hegeliana, kantiana, jusnaturalista, tomista ou eminentemente teológicas. Provavelmente, os estudos nessas direções serão ainda mais elucidativos, instigantes e desafiadores do que a perspectiva ora adotada. Os argumentos a serem refutados no presente trabalho se dirigem tão somente àqueles objetos de estudos específicos na área da filosofia da inteligência artificial expostos em *Turing e Searle*, com os fundamentos pautados nos autores antecedentes.

A ideia de possibilidade filosófica cinge-se à noção de uma ordem de coisas consistente, ou seja, que não viole as regras lógicas decorrentes da aplicação do princípio da contradição. A afirmação da possibilidade filosófica de uma inteligência artificial forte deve ser capaz de articular de modo consistente os seus pressupostos, de maneira a não incorrer em um argumento inconsistente. Não se pretende, contudo, apresentar um argumento exaustivo, capaz de abordar todas as possíveis refutações ou objeções à tese da possibilidade de uma inteligência artificial forte. Limita-se, modestamente, a refutar algumas das principais objeções apresentadas por *Turing*, contra o argumento de que as máquinas podem pensar, mais propriamente: a objeção da consciência, das imperfeições, da intencionalidade, da limitação algorítmica (*Argumento de Ada Lovelace*), biológica e teológica.

Partindo de um conceito de inteligência artificial, verificar-se-á a decorrente possibilidade da existência de uma ética artificial e da tese da eventualidade de agentes morais artificiais. O desafio da definição de uma ética artificial autêntica precisa considerar as suas características. Os debates filosóficos sobre a moralidade delimitaram historicamente o conceito de sujeito moral. Estendê-lo a um agente artificial, portanto, é um dos temas mais intrigantes no debate filosófico sobre a inteligência artificial. Alguns dos mais relevantes argumentos tocados por *Turing*, contudo, não serão abordados, tais como a objeção matemática.

As teorias morais têm se dividido em modelos distintos, com pressupostos e conclusões distintas, sobre diversos conceitos fundamentais da Ética. A distinção se torna ainda mais clara quando aplicada à inteligência artificial e às possibilidades de solução de conflitos morais por parte de um sistema inteligente. Como deveria um agente moral artificial deliberar em face a uma dilema moral? Duas são as alternativas principais aventadas pela doutrina recente: de um lado, um agente moral artificial deveria trazer, no código computacional, todas as regras para a melhor decisão ou deveria agir norteado por um modelo de aprendizado, comparando o seu comportamento com condutas exemplares. A escolha de um ou de outro modelo é um dos principais desafios.

A possibilidade de superação do Teste de Turing, do teste da Sala Chinesa e do Teste de Ada Lovelace se tornou um grande desafio para uma filosofia da inteligência artificial. Muitas questões permanecem em aberto, mas nem todas serão objeto do presente trabalho.

Cumpre, no entanto, ter presentes essas fundamentais indagações. Afinal as máquinas seriam racionais, conscientes, teriam vontade própria? Poderiam ter emoções? Deveriam ser responsabilizadas pelos seus atos e escolhas? Teriam alma? Seria possível uma agente moral artificial que respondesse positivamente a todos esses questionamentos? Ou, sintetizando, seria possível a emergência de um autêntico agente artificial moral, com deliberações intencionais em perspectiva de primeira pessoa?

Trata-se de um dos problemas mais relevantes da história humana. A possibilidade do surgimento do primeiro sujeito moral não humano. Essa será realmente possível ou todos os seres artificiais seriam apenas simulacros, que aparentam ter vontade própria, mas são manipulados como marionetes? Por fim, poderíamos questionar se existiria a possibilidade de haver um código computacional capaz de dar origem à emergência de um sujeito artificial autêntico.

Adota-se, como referencial teórico a ser estudado, a teoria ética das virtudes, tal como entendido em *Aristóteles*, *P. Foot, Anscombe, MacIntyre* e na epistemologia das virtudes. O presente trabalho não pretende verificar a possibilidade filosófica de um agente moral artificial por outras escolas relevantes, tais como as de *Hegel*[1] ou *Kant*, que poderiam igualmente apresentar soluções muito razoáveis e consistentes para o problema e, provavelmente, serão objeto de estudos futuros sob essas óticas.

A resposta a todos esses questionamentos pretende auxiliar a entender a tese da possibilidade filosófica do surgimento de um autêntico agente moral artificial, com todas as características necessárias para a emergência de um sujeito moral verdadeiro.

[1] Apesar de não ser objeto da presente tese, não cremos que um estudo sob o referencial hegeliano demonstra incompatibilidade imediata com o questionamento. O fundamento (Espírito) poderia ser mantido no sentido de dizer que a IA é o Espírito de nosso tempo, conforme a marcha da história. Assim, mesmo que Hegel não seja o referencial teórico, poderia ainda ser mencionado como uma visão que não é incompatível *a priori* com a tese da existência de IA forte e mesmo de um Agente Moral Artificial autêntico. Estudos posteriores poderão confirmar ou infirmar esse entendimento.

1 PRIMEIRA PARTE. ÉTICA E INTELIGÊNCIA ARTIFICIAL: UMA ANÁLISE CRÍTICA

1.1 CONCEITO FILOSÓFICO DE INTELIGÊNCIA ARTIFICIAL

1.1.1 Deuses, homens e alguns autômatos

No princípio existiam deuses, homens e alguns autômatos, segundo os gregos. Coube a um pastor, na verdade um dos maiores e mais antigos poetas, *Hesíodo* (750 a.C.?), contar com elegância essa cronologia. O escritor relata que as próprias musas o inspiraram a contar a *Teogonia*: "elas um dia a Hesíodo ensinaram belo canto quando pastoreava ovelhas ao pé do Hélicon divino"[2].

A obra de *Hesíodo* é arcaica na concepção precisa do termo. Não se trata de uma obra desatualizada, conforme o sentido ordinário e vulgar da palavra. Ela é uma poesia arcaica no sentido derivado de *arkhé*[3], decorrente do verbo *arkhómetha* (principiar), ou seja, trata-se de uma obra principal, inaugural ou antecedente. Mais além do sentido cronológico, de inicial ou de começo, ela possui um sentido essencial. Trata-se do princípio ordenado pela unidade indiscernível. A própria *Teodicéia* adota esse modo de relatar: "Sim bem primeiro nasceu Caos", depois "a Terra, o Tártaro, Eros, Érebo e a Noite negra, Éter e o Dia". Do princípio unitário, surgiram os deuses primordiais.

A origem dos homens é tão atribulada quanto a história conflituosa dos deuses, segundo os gregos. É justamente das disputas mitológicas que este surge, por gosto e capricho. Conta *Platão*, na sua obra *Protágoras*, que "houve um tempo em que só havia deuses, sem que ainda existisse criaturas mortais", e estas seriam criadas nas entranhas da terra, utilizando-se de ferro e fogo.[4] A própria palavra homem deriva do latim *húmus*, significando "terra" ou "terreno". Essa origem é corroborada em outros autores como *Esopo* (620 a.C.-564 a.C.) e *Ovídio* (43 a.C.-18 d.C.).

A fábula de *Esopo* reafirma *Prometeu* como criador da humanidade. A versão desse escritor é mais tocante. Os mortais teriam sido feitos de barro, porém, em vez de esse barro ter sido misturado com água, havia sido formado com lágrimas[5].

Ovídio põe em destaque a criação do homem, no desenrolar da *Teodicéia*. A ordenação do universo estava quase completa. Lá estavam os deuses, as estrelas e até os pequenos animais.

---

[2] HESÍODO. *Teogonia*: a origem dos deuses. Trad. Jaa Torrano. São Paulo: Iluminuras, 2003. p. 87.
[3] HEIDEGGER, Martin. *O que é isto* — A filosofia? São Paulo: Abril Cultural, 1973. p. 219. (Col. Os Pensadores).
[4] PLATÃO. *Protágoras*. Trad. Carlos Alberto Nunes. Belém: Universidade Federal do Pará, 2002, XI. 321. D.
[5] DOUGHERTY, Carol. *Prometheus*. Taylor & Francis, 2006. p. 17.

Mas faltava um ente para coroar a criação (“estes animais faltava um ente / dotado da mais alta inteligência”)[6]. Uma das propriedades essenciais desse novo ente é a sua inteligência, que o distingue de todos os outros animais e seres sobre a terra.

Conforme *Hesíodo*, os homens estariam entre os deuses e as bestas[7]. Viriam da terra, mas teriam a face em direção às estrelas (“O Factor conferiu sublime rosto / Erguido, para o céu lhe deu para que olhasse”)[8]. Olhar ao alto indicaria, talvez, que a sua inteligência era tanto prática, voltada para a terra, quanto abstrata, ao mirar o céu profundo.

Esopo iria tratar dessa dádiva concedida aos mortais na fábula *Zeus e o Homem*, na qual este se queixava de não ter as habilidades de certos animais, não podia voar altos voos, nem tinha a força ou a velocidade de certos animais. A quem *Zeus* repreende, ao dizer que ele detinha o dom da fala e a habilidade da razão[9]. A inteligência estaria vinculada a essas duas habilidades.

Diversos outros autores latinos (*Cattulus*, *Horatius* e *Propertius*) irão confirmar a noção de que *Prometeu* é o criador dos homens[10]. Trata-se de um mito poderoso, realçado por diversos escritores. Mas há um outro dado importante. Prometeu definirá, de modo inexorável, o destino da condição humana.

Prometeu irá presentear os humanos com o fogo, representando as artes técnicas, capazes de permitir aos mortais superar as limitações dos ciclos da natureza[11]. O deus-titã se caracteriza como um benfeitor e protetor da humanidade. Esse desejo de ajudar esses seres desgraçados faz com que ele os conceda o domínio das *technes*; tais como os remédios, as curas, a adivinhação, o conhecimento dos sonhos[12], bem como todas as artes que dominam pelo trabalho a natureza.

A tragédia humana está inserida em sua experiência existencial no sofrimento, no trabalho, por meio do “suor de seu rosto”, para elevar-se de para além de sua mísera condição, por meio da sabedoria. Sem o sofrimento não se alcança a sabedoria. Até este momento, o mundo é repleto de deuses, entes e homens. Nada se fala de máquinas artificiais, que possam imitar o comportamento humano.

---

[6] OVÍDIO. *Metamorfoses*. Trad. Bocage e comentários de Rafael Falcón. Porto Alegre: Concreta, 2016. p. 49. (Coleção Clássica.). Disponível em: https://issuu.com/editoraconcreta/docs/metamorfoses-teste. Acesso em: 23 maio 2020 às 23:34.
[7] DOUGHERTY, 2006, p. 35.
[8] OVÍDIO, 2016, p. 49.
[9] AESOP'S FABLES. A new translation by Laura Gibbs. Oxford University Press (World's Classics): Oxford, 2002. Disponível em: http://mythfolklore.net/aesopica/oxford/514.htm. Acesso em: 23 maio 2020 às 23:56.
[10] DOUGHERTY, 2006, p. 17.
[11] AZAMBUJA, Celso Candido. Prometeu: a sabedoria pelo trabalho e pela dor. *Archai*, n. 10, jan.-jul. 2013, p. 19-28.
[12] AZAMBUJA, 2013, p. 25.

A primeira menção a servos mecânicos aparece pela obra de *Homero*, no século VIII a.C., na *Ilíada*. A inaugural menção a autômatos aparece no livro como criações maravilhosas de *Hefesto*, o ferreiro dos deuses, deus da fundição, das invenções e da tecnologia. Diz a lenda que ele construiu um magnífico palácio de bronze, com inúmeros servos mecânicos. *Hefesto* prometera fabricar armas para *Aquiles*, o que fez ladeado por autômatos em formas femininas (XVIII. "Vem coxeando; o rei trôpego esteiam / Moças de ouro que às vivas assemelham / Na força e mente e voz, por dom celeste")[13]. As servas mecânicas do deus Vulcano eram dotadas de força, da palavra e, mais impressionantemente, de "mentes".

A palavra autômatos viria do grego: aquilo que "se move ou age por si mesmo"[14], um mecanismo que se movimentava por "motor" próprio e se distinguia de mecanismos que imitavam entes, mas manipulados por humanos para manifestar uma imitação. O teatro grego conhecia os exemplos de máquinas que simulavam a aparição de personagens que resolviam a trama, sendo o mais conhecido o *Deus ex machina*. Assim como se distinguiam de outros *simulatra* operados por humanos, como as *marionettes*.

A próxima estreia da mimetização artificial do humano é avassaladora, nas *Argonáuticas*, de *Apolónio de Rodes* (c. 295 a.C.-215 a.C.). Trata-se de um gigante de bronze denominado *Talos*, que guardava a ilha de Creta. *Talos* foi uma das criações de *Hefesto*, mas, diferentemente das doces servas mecânicas de ouro, o gigante de bronze mostra a face tenebrosa da tecnologia. Ele tinha o propósito de defender a ilha de Creta dos piratas. Na mitologia grega, *Talos* aparece ao lado de outras inúmeras criaturas bizarras e demônios que protegiam cidades de invasores[15], tal como o *Minotauro* ou a *Medusa*. Possuía uma espécie de fluido dourado divino, o "*ichor*", presente no sangue dos deuses, que, por uma única artéria, circulava por todo o seu corpo. O que levantou as dúvidas se *Talos* era imortal, senciente ou mesmo provido de alma. Muitas questões foram levantadas: seria senciente ou apenas imitaria o comportamento humano? Teria intelecto? Ou mesmo, sentiria algo?

Os limites do humano e do inumano que o imita serão objeto de alguma breve, porém importante, crítica filosófica posterior.

---

[13] HOMERO. *Ilíada*. Trad. Manoel Odorico Mendes. eBooksBrasil, 2009. Livro XVIII. Disponível em: http://www.ebooksbrasil.org/adobeebook/iliadap.pdf. Acesso em: 14 dez. 2020 às 09:47.
[14] BERRYMAN, S. Ancient automata and mechanical explanation. *Phronesis-A Journal for Ancient Philosophy*, v. 48, n. 4, p. 344-369, 2003.
[15] MAYOR, Adrienne. *Gods and robots*: myths, machines, and ancient dreams of technology. Princeton: Princeton University Press, 2018. p. 19.

1.1.2 O humano e o autômato

O conceito filosófico de *autômato* não surgiu inicialmente para análise de *corpus mechanicum* ou tecnológicos, tais como *Talos* ou as servas douradas de *Hefesto*, mas ao estudo dos animais não humanos. Os limites do humano encontravam o desafio na sua delimitação com os outros seres vivos, especialmente, os animais com longa duração de vida.

O estudo mais importante e mais citado, sobre essa inaugural menção ao termo autômato, surgirá no comentário de *Aristóteles* em sua obra *Metafísica*, na expressão "*ta automata tôn thaumatôn*" (autônomo como nos fantoches[16]). Provavelmente tenha sido a primeira aplicação filosófica do termo *automata*. Segundo *Aristóteles*:

> For prior to their knowing, they wondered [ἐθαύμαζον] that things could be as they are, but once they had come to know they wondered [θαυμάζουσιν] that things can fail to be as they are. [As examples of] wonders [θαύματα] he mentions the toys [παίγνια], exhibited by the creators of [such] marvels [ὑπὸ τῶν θαυματοποιῶν], that seem to move by their own power [αὐτομάτως κινεῖσθαι], and the solstices, which bring winter and summer[17] [18].

Ao se referir ao movimento de autômatos, o autor demonstra ser esse inautêntico, entes que aparentemente parecem se mover por motor próprio, mas são como bonecos ou fantoches, sem movimento com causa interna.

Para *Aristóteles* a natureza é a substância das coisas que possuem o princípio do movimento em si mesmas e por sua essência. Distinguir os seres com base no movimento ou crescimento era um critério possível para essa importante indicação da excepcionalidade do humano perante todos os demais seres.

É possível encontrar-se a afirmação de que *Aristóteles* compara as funções animais como se fossem mecanismos, e eles mesmos fossem uma espécie de bonecos ou fantoches, mas a apreciação orgânica destes se sobressai à sua visão mecanicista. Dois diferenciais citados pelo autor são o seu movimento voluntário e a capacidade de reprodução[19].

Na obra *A história dos animais* (*De generatione animalium*), *Aristóteles* apresenta a especificidade humana. Os animais se movimentam de modo orgânico ou de modo intencional.

---

[16] Tradução a partir do trecho disponível em http://hypnos.org.br/revista/index.php/hypnos/article/viewFile/47/47.
[17] ARISTÓTELES *apud* BOWE, Geoffrey. Alexander's Metaphysics commentary and some scholastic understandings of automata. Schole. Ancient Philosophy and the Classical Tradition, v. XIV, 2020, Issue 1. Disponível em: https://nsu.ru/classics/schole/14/schole-14-1.pdf. A Journal of the Centre for Ancient Philosophy and the Classical Tradition.
[18] ARISTOTELE. *Metafísica* – Saggio introduttivo, testo greco con traduzione. Disponível em: https://moodle.ufsc.br/pluginfile.php/1332285/mod_resource/content/1/Aristoteles-Metafisica-Edicoes%20Loyola%20%282002%29.pdf. Acesso em: 15 dez. 2020 às 00:08.
[19] Disponível em: http://web-b-ebscohost.ez94.periodicos.capes.gov.br/ehost/pdfviewer/pdfviewer?vid=2&sid=3a78d2a0-5332-45fa-92fe-752bea38f573%40sessionmgr103.

Agem de modo orgânico quando movidos por uma ação externa e de modo intencional quando agem voluntariamente. O ser humano, singularmente, é ainda um agente moral, porque *age por deliberação* para alcançar o fim que lhe é próprio[20]. Novamente o termo aparecerá na passagem da obra *De generatione animalium* tratando do movimento que provém do interior e não de causas externas:

> É possível que este mova outro, este ainda outro, e que se passe como nos fantoches (*ta automata tôn thaumatôn*). As partes em repouso, em certo sentido, possuem uma potência, e quando algo do exterior move a primeira destas, a seguinte logo se põe em atividade. Assim como nos fantoches, algo de algum modo desencadeia o movimento sem tocar em nada neste momento, mas tendo tocado antes, do mesmo modo também age aquele de quem provém o esperma ou que produziu o esperma, tendo antes tocado algo, mas agora não o tocando mais. De certo modo, o movimento que existe internamente é como o processo de construção da casa[21] (grifos nossos).

A capacidade de deliberação é a singularidade do humano. *Aristóteles* irá afirmar, no Livro II do *De generatione animalium*, que a finalidade da natureza "é particularmente clara nos demais seres vivos [não humanos] que não atuam nem por arte, nem tampouco porque haviam investigado, nem deliberado" (Ph. II, 8, 199a 20-21). Mas será na *Política* que *Aristóteles* irá apresentar os três traços distintivos do humano: fala, qualidades éticas e comunicação[22]. O ético irá distinguir a sua ação por meio da percepção do bem (Pol. I, 1, 1253a 14-18). Já "a comunicação faz a casa e a cidade" (Pol. I, 1, 1253a 18-19).

Recorrendo ao risco da expansão por analogia, podemos dizer que *Aristóteles* diferencia o humano dos animais e, provavelmente, o diferenciaria dos autômatos pelas seguinte razões: i) os humanos se movimentam por deliberação, em direção a um fim; ii) possuem o dom da fala; iii) são agentes morais e iv) possuem comunicação. Todas essas características são igualmente relevantes e indiscutíveis.

A ciência e a filosofia contemporânea estão justamente a debater se as máquinas podem movimentar-se autonomamente, se podem se comunicar, mas o mais importante em nossa perspectiva é se podem *tomar deliberações morais.* O objeto da presente tese será justamente **verificar a possibilidade filosófica de os autômatos serem agentes morais artificiais**, algo demasiadamente humano.

---

[20] MARIZ, Débora. A especificidade da natureza humana em relação aos demais animais no pensamento aristotélico. *Argumentos*, Fortaleza, ano 6, n. 12, p. 157-168, p. 161, jul./dez. 2014.
[21] ARISTÓTELES. *De generatione animalium* I, 22, 730b 11-23.
[22] MARIZ, 2014, p. 157-168, p. 163.

1.1.3 Descartes: sobre humanos e máquinas

O brilhantismo grego em filosofia não legou uma teoria dos autômatos. Isso não passou despercebido por um dos pioneiros da teoria computacional, *Norbert Wiener*, em sua obra "Cybernetics or control and communication in the animal and the machine" (1965). O cientista relata que nenhuma teoria séria nos foi legada pela antiguidade para entendermos o fenômeno contemporâneo. Nem de perto as máquinas antigas possuem paralelo com as atuais, por mais criativas ou "mágicas" que fossem, não passavam de mecanismos automatizados (*clockwork automaton*), muito diferentes da cibernética contemporânea[23].

Uma teoria moderna dos autômatos principia com *René Descartes*[24]. Claramente o autor estava familiarizado com as intrigantes e sofisticadas máquinas artificiais que circulavam pela Europa, entretendo e desconcertando as cortes, mas ele não estava particularmente preocupado com esses fenômenos. Seu projeto principal era reformar a ciência e a filosofia, reorientando os interesses filosóficos dos debates escolásticos sobre ontologia, metafísica e teologia para considerações sobre o conhecimento, a mente, a subjetividade e a consciência[25].

Descartes é conhecido como o primeiro cientista cognitivo[26] pelos seus diversos estudos sobre a mente. Será, contudo, a sua tese sobre a separação absoluta entre mente e cérebro que irá despertar debates acalorados no âmbito dos cientistas e filósofos que lidam com o tema da inteligência artificial. O radicalismo de suas teses ainda exalta defensores de uma e outra parte, mas se originam em postulados bem construídos. As reações fortes de sua teoria decorrem de uma leitura parcial de sua obra ou mera incompreensão ou preconceito, sendo acusado algumas vezes de responsável pelo tratamento cruel ou industrial dos animais[27]. Foge ao objeto deste estudo, porém, realizar uma análise detida sobre os argumentos fundamentais dessa controvérsia[28], cingindo-se tão somente os argumentos adotados no debate sobre a inteligência artificial.

Para *Descartes*, ainda, existem dois modos para a substância (*cogitatio et extensio sumietiam possunt pro modis substantiae*): mente e corpo (*cognitio et extension*), reconhecidos pela herança cartesiana como a tese do dualismo. A sua origem está identificada na dúvida

---

[23] WIENER, Norbert. *Cybernetics or control and communication in the animal and the machine*. Massachusetts: MIT, 1965. p. 40.
[24] WIENER, 1965, p. 40.
[25] TEIXEIRA, João de Fernandes. *O pesadelo de Descartes*: do mundo mecânico à inteligência artificial. Porto Alegre: Editora Fi, 2018. p. 34.
[26] BATES, David. Cartesian Robotics. *Representations*, v. 124, n. 1, p. 43-68, Fall 2013.
[27] CHIAROTTINOI, Zelia Ramozzi; FREIRE, José-Jozefran. O dualismo de Descartes como princípio de sua Filosofia Natural. *Estudos Avançados*, v. 27, n. 79, p. 158, 2013.
[28] FAUSTO, Juliana. A cadela sem nome de Descartes: notas sobre vivissecção e mecanomorfose no século XVII. *Doispontos*, Curitiba, v. 15, n. 1, p. 43-59, abr. 2018.

metódica, defendida nas obras "Discours de la Méthode" e "Meditations Metaphysiques touchant la premiere Philosophie"[29].

Descartes, no "Discurso do Método", principia com a sua assunção da *dúvida metodológica* e das razões que devemos ter para duvidar de todas as coisas, especialmente das materiais e das nossas sensações ("tudo o que recebi, até presentemente, como o mais verdadeiro e seguro, aprendi-o dos sentidos: ora, experimentei algumas vezes que esses sentidos eram enganosos, e é de prudência nunca se fiar inteiramente em quem já nos enganou uma vez")[30]. A seguir ele afirma que, dada a liberdade de a mente duvidar de todas as coisas materiais, reconhece que é impossível ela mesmo não existir. Assim, provar-se-á que mente e corpo possuem naturezas absolutamente distintas. Tal é confirmado pelo fato de que a corporeidade é divisível, enquanto a mente é indivisível, dado que não podemos conceber a metade de um espírito[31].

Descartes rompe diretamente com *Aristóteles* que afirmava que todo o corpo vivo possui uma alma e que existe uma implicação necessária entre um e outro. Um não existe sem o outro[32]. O corpo seria a estrutura própria da alma:

> Esforçam-se apenas, todavia, por dizer que tipo de coisa é a alma; acerca do corpo que a acolhe, nada mais definem, como se fosse possível, de acordo com os mitos pitagóricos, que uma alma ao acaso se alojasse em qualquer corpo. É que cada coisa parece possuir forma específica – quer dizer, estrutura – própria. Eles exprimem-se, no entanto, como se se dissesse que a técnica do carpinteiro se alojou nas flautas: é preciso, pois, que a técnica use as suas ferramentas, e a alma o seu corpo.[33]

Para *Descartes* é preciso afastar essa compreensão filosófica de unir o estudo do corpo a uma figura misteriosa ou fantasmagórica como a alma. É preciso distinguir claramente estas duas substâncias para conhecer o mundo ("A principal distinção que observo entre as coisas criadas é que umas são intelectuais, isto é, substâncias inteligentes, ou então propriedades que pertencem a tais substâncias; as outras são corporais, isto é corpos ou propriedades que pertencem ao corpo")[34]. Cada um possui um atributo próprio, a alma possui o pensamento e o corpo, a extensão. A distinção entre corpo e alma é absoluta, um pode existir sem o outro

[29] CHIAROTTINOI; FREIRE, 2013, p. 161.
[30] DESCARTES, René. *Meditações sobre a Filosofia Primeira*. Tradução de Fausto Castillo. Campinas: Unicamp, 2004. p. 118. (Coleção Multilíngues de Filosofia).
[31] DESCARTES, 2004, p. 223.
[32] BITENCOURT, Joceval Andrade. Descartes e a Morte de Deus. Livrosgratis.com.br, 2012, p. 94. Disponível em: https://www.livrosgratis.com.br/ler-livro-online-20929/descartes-e-a-morte-de-deus, acesso dia 15 dez. 20, às 00:17.
[33] ARISTÓTELES. *Sobre a Alma*. I, 3, 407b 20-24. Imprensa Nacional Casa da Moeda, 2020. p. 47.
[34] DESCARTES, 2004, p. 44, I, art. 48.

("minha alma, pela qual sou o que sou, é inteira e verdadeiramente distinta do meu corpo e que ela pode ser ou existir sem ele")[35].

Descartes rompe igualmente com a teoria do movimento em *Aristóteles*. Para este filósofo, a matéria traz em si mesmo a causa de seu deslocamento. Descartes reputa isso como errôneo. Não existe uma inteligibilidade própria no movimento das coisas, mágico ou misterioso, que fizesse algo movimentar-se sob o domínio de uma força oculta ("não parecem proferir palavras mágicas, com uma força oculta e que superam o alcance do espírito humano, aqueles que dizem que o movimento, coisa muito conhecida de cada um, é o ato do ser em potência, enquanto está em potência? Quem compreende de fato essas palavras?")[36]. *Descartes* enuncia um novo mundo, liberto dos espíritos ocultos na matéria, livre de orientações internas, das causas finais, da inteligibilidade oculta nas coisas, um mundo de leis simples, matemáticas, sem mágicas, um mundo mecânico[37].

Os corpos, inclusive o humano, nada mais seriam do que um mecanismo, semelhante a um relógio ("Não vejo, efectivamente, nenhuma diferença entre as máquinas feitas pelos artesãos e o diversos corpos formados exclusivamente pela Natureza")[38]. O homem com corpo mecânico estaria perfeitamente integrado a um mundo mecânico. Um novo mundo, sem metafísica ou uma nova metafísica, sobre os ombros de uma nova ciência.

Os escritos de *Descartes* e seus pressupostos levam a crer que a mente não pode ser mecânica, dado que é indivisível e não pode ser decomposta em partes, nem seria ela a parte, essencial ou necessária, de um corpo mecânico. As partes divisíveis de um corpo mecânico podem ser divididas, reconstruídas e reorganizadas, ou seja, replicadas. Uma mente não poderia ser replicada. O homem poderia fazer mecanismos complexos como um relógio ou outras maravilhas da engenharia, mas jamais fazer uma máquina que replicasse a mente. Teoricamente seria exigir que uma substância replicasse outra substância absolutamente distinta. No máximo, seria uma bela imitação. *Descartes* rompe ainda com a tradição que distinguia os seres vivos das máquinas. Para ele todos os corpos são máquinas e os seres vivos são corpos destituídos de espírito, isto é, não passariam de máquinas vivas ou bestas-máquinas (*bêtes-machines*)[39].

---

[35] DESCARTES, 2004, p. 186-187.
[36] DESCARTES, *Regras para a orientação*..., reg. XII, p. 92.
[37] BITENCOURT, Joceval Andrade. Descartes e a Morte de Deus. *Livrosgratis.com.br*, 2012, p. 94. Disponível em: https://www.livrosgratis.com.br/ler-livro-online-20929/descartes-e-a-morte-de-deus, acesso dia 15 dez. 20, às 00:17.
[38] DESCARTES, *Regras para a orientação*..., reg. XII, p. 118.
[39] AVRAMIDES, Anita. Descartes and Other Minds. *Teorema*, v. XVI/1, p. 27-46, p. 33, 1996. Disponível em: file:///Users/caliendo/Downloads/Dialnet-DescartesAndOtherMinds-4244359.pdf. Acesso em: 28 maio 20 às 12:05.

A definição de *automata* será derivada desses pressupostos na obra "Tratado do Homem", como um corpo cujas funções (por mais sofisticadas que sejam) seguem os desígnios de sua organização[40].

O singular no homem seria a sua subjetividade, na sua capacidade de pensar. Mas seria possível uma máquina ou robô simular corretamente o pensamento? Para a teoria de *Descartes* a resposta seria negativa. A cognição humana seria impenetrável para autômatos e toda a imitação seria uma falsidade e jamais iria conseguir simular com autenticidade o pensamento humano[41].

Descartes irá se deparar com um importante problema filosófico: como podemos distinguir um indivíduo real de um autômato com forma humana? O problema já existia de modo concreto na época do filósofo, dado que pulsava a curiosidade na época para os sofisticados autômatos produzidos, especialmente por relojoeiros, e afirma-se inclusive que ele teria construído um[42], denominado de "minha filha Francine"[43]. A coitada da filha-máquina teria sido lançada ao mar por um capitão que identificou na máquina uma provável obra demoníaca, ou seja, um corpo mecânico habitado por um espírito maligno. Poderíamos imaginar a situação catastrófica oposta: e se o irado capitão atirasse ao mar a filha da serva de *Descartes*, a homônima *Francine*?

Descartes responde que deveríamos utilizar testes para identificar a presença de um "indivíduo real". Quais seriam esses importantes e imprescindíveis testes?

O primeiro deles seria o uso da linguagem, a capacidade de resposta articulada a tudo que seja dito na presença deste ser, como ele consegue declarar com competência seus pensamentos ("não é possível conceber que as combine de outro modo para responder ao sentido de tudo quanto dissermos em sua presença")[44]. O teste de *Descartes* seria uma espécie da versão posteriormente prevista por *Allan Turing*. Mas e se a máquina pudesse mimetizar a capacidade de respostas semelhante a um ser humano? E se ela passasse no teste de *Descartes*? E se essa capacidade fosse similar à mesma competência linguística de uma criança, de uma "pessoa embrutecida"[45] ou de um especialista em determinada área (medicina, direito, fiscal, etc.)?

---

[40] AVRAMIDES, 1996, p. 34.
[41] PORTO, Leonardo Sartori. Uma investigação filosófica sobre a inteligência artificial. *Informática na Educação: Teoria & Prática*, Porto Alegre, v. 9, n.1, jan./jun. 2006.
[42] KANG, Minsoo. The mechanical daughter of Rene Descartes: the origin and history of an intellectual fable. *Modern Intellectual History*, v. 14, n. 3, p. 633-660.
[43] TEIXEIRA, João de Fernandes. *O cérebro e o robô*: inteligência artificial, biotecnologia e a nova ética. São Paulo: Paulus, 2015.
[44] DESCARTES, René. *Discurso do Método*. São Paulo: Martins Fontes, 2009. p. 63.
[45] DESCARTES, 2004, p. 63.

O segundo teste seria a incapacidade dos autômatos de terem um *conhecimento prático* ou abrangente. Vejamos a afirmação de *Descartes*: "E o segundo é que, embora fizesse várias coisas tão bem ou talvez melhor do que algum de nós, essas máquinas falhariam necessariamente em outras, pelas quais se descobriria que não agiam por conhecimento, mas somente pela disposição de seus órgãos". E segue: "[...] daí ser moralmente impossível que haja uma máquina a diversidade suficiente de órgãos para fazê-la agir em todas as ocorrências da vida da mesma maneira que nossa razão nos faz agir"[46].

O argumento de *Descartes* é instigante, afinal seria possível uma máquina mimetizar quase perfeitamente a flexível e fina conduta humana nas condições reais da vida? Poderia o sublime ou o admirável da ação humana ser "mecanizado"? Seria a afirmação do filósofo limitada por sua condição temporal, contingente ou acidental? Algo que em seu tempo era impossível, mas hoje ou no futuro fosse uma possibilidade técnica. Tornar-se-ia a máquina um ser real? E se uma máquina pudesse mimetizar ainda melhor do que muitos seres humanos (crianças ou "homens embrutecidos")? Seria ela um "indivíduo real"? As limitações aduzidas trazem dúvida sobre os testes propostos, e assim como o próprio filósofo advoga, devemos afastar os métodos inseguros e obscuros? Outra limitação aos testes cartesianos estaria no fato de que se fundam em critérios de distinção acidentais e não essenciais entre o humano e o inumano. Debitar as distinções às limitações das possibilidades técnicas atuais não parece ser um critério conclusivo.

Apesar desses questionamentos, é inegável a contribuição de *Descartes* para uma teoria dos autômatos. Suas intuições serão tão originais que somente os autores contemporâneos conseguirão produzir respostas adequadas.

O dualismo contemporâneo será herdeiro e legatário das diversas contribuições elaboradas por *Descartes*.

1.1.4 Ada Lovelace e as máquinas sem pensamento

O primeiro algoritmo na história é fruto do talento e da inspiração de uma matemática competente. Aos 12 anos de idade, *Ada Lovelace* escrevera um livro sobre cavalos alados mecanizados[47] e talvez passasse o resto de sua vida combinando a imaginação livre e abstração rigorosa, com desenvoltura ímpar. *Ada* viveu entre gigantes de "continentes" separados, porém

[46] DESCARTES, 2004, p. 64.
[47] HOLLINGS, Christopher; MARTIN, Ursula; RICE, Adrian. The early mathematical education of Ada Lovelace. *BSHM Bulletin, Journal of the British Society for the History of Mathematics*, 2017. Disponível em: https://www.claymath.org/sites/default/files/the_early_mathematical_education_of_ada_lovelace.pdf. Acesso em: 02 ago. 2020 às 21:36.

muito próximos: a poesia e a matemática. Para entender o seu estilo, trabalho e conclusões, é necessário entender um pouco de seu contexto[48]. Filha de *Lord Byron* o famoso poeta britânico, trabalhou com o pai da computação *Charles Babbage* e manteve contato com *Michael Faraday* e *Charles Dickens*, dentre outros cientistas famosos de sua época.

A biografia de *Ada Lovelace* (1815-1852) é encantadora e mereceu diversos e merecidos estudos. Seu grande contributo foi a criação do primeiro algoritmo computacional, ou seja, de um programa bem ordenado de passos para a realização de uma rotina de operações lógicas.

A época era de invenções geradas em velocidade alucinante, de uma revolução industrial incessante em diversos setores produtivos (química, metalurgia, eletricidade, etc.). Ao lado da construção célere de máquinas mais rápidas, mais fortes e maiores, uma outra revolução estava acontecendo. *Ada* estava preocupada com máquinas virtuais.

As máquinas da revolução industrial sugeriam mecanismos físicos permeados de impressões desagradáveis. Geralmente por meio de aparatos físicos destinados a torcer, distorcer ou transformar a natureza física, geralmente pelo uso da força bruta. Muitas vezes geram temor e desconfiança, pelo risco de causarem danos aos seus usuários ou criadores. De um lado inspiram admiração, de outro, desconfiança[49]. Talvez não seja sem razão que o deus da técnica, *Hefesto,* seja manco, feio e não desperte a mesma empatia de outros afamados deuses.

Ada iria trabalhar e criar máquinas em um outro sentido, muito diverso, daquele da intuição tradicional. Em vez de máquinas físicas, ela seria a pioneira nas máquinas computacionais[50]. Estas se caracterizam como ideias abstratas de especificações de como um objeto físico deve funcionar[51]. *Ada* irá explicar o funcionamento dessas máquinas virtuais de modo muito elegante como se estivesse tecendo padrões algébricos, da mesma forma que o tear mecânico tece flores e folhas (*the Analytical Engine weaves algebrical patterns just as the Jacquard-loom weaves flowers and leaves*)[52], ou seja, ela somente automatiza procedimentos, tal como faz um tear. Ele não cria padronagens novas, nem desenha novas e sublimes formas. Ele segue um padrão ao estilo dos cartões perfurados do tear mecânico de *Jacquard*[53]. Há uma

---

[48] TOOLE, B. Poetical Science. *The Byron Journal*, 15, p. 55-65, 1987.
[49] MINSKY, Marvin Lee. *Computation*: finite and infinite machines. Austin: Englewood Cliffs, N.J.; Prentice-Hall, 1967. p. 1.
[50] Iremos nos referir ao sentido de máquinas, no presente trabalho, nesse sentido, de sistemas virtuais.
[51] MINSKY, 1967, p. 5.
[52] LOVELACE, Ada *apud* BABBAGE, Charles. *Sketch of the analytical engine invented*. [1942]. Disponível em: http://www.fourmilab.ch/babbage/sketch.html. Acesso em: 29 maio 2020 às13:07.
[53] HEATH, G. Origins of the binary code. *Scientific American*, v. 227, n. 2, p. 76-83, August 1972; e KIM, Eugene Eric; TOOLE, Betty Alexandra. Ada and the First Computer. *Scientific American*, v. 280, n. 5, p. 76-81, May 1999.

informação na entrada da máquina (*input*), um processamento e uma informação na saída (*output*). A máquina física ou virtual não cria seus próprios códigos, ela segue uma rotina delimitada. Enfim, ela automatiza um procedimento. Não existiria "inteligência artificial", apenas um uso inteligente das máquinas automáticas.

Apesar de ser profundamente conhecedora dos primeiros computadores criados, bem como uma exímia matemática, *Ada* lançará uma conclusão sobre a possibilidade de as máquinas pensarem, que iriam dominar por um século o debate sobre o tema. Seu vaticínio é simples, claro e objetivo: *a máquina analítica não possui pretensão de gerar nada* (*The Analytical Engine has no pretensions whatever to originate anything*)[54]. Dela nada emerge espontaneamente.

Para ela nenhuma competência preditiva poderia ser derivada da máquina analítica, a máquina somente performaria o que foi programada para fazer (*It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths*)[55].

Talvez a sua afirmação se dirigisse apenas à máquina analítica ou às máquinas de seu tempo? Ou à máquina que estava trabalhando com os números de *Bernoulli*? De um modo ou de outro, a sua afirmação foi lida com pretensões ampliadas, levando a dúvida metódica a todas as futuras possibilidades.

Não se pode debitar a *Ada* a falta de imaginação poética, geralmente atribuída a matemáticos rigorosos, para verificar possibilidades ocultas e não vislumbradas ou desdobramentos ambiciosos, para os avanços tecnológicos. Talvez por isso sua afirmação clara e límpida ecoou por tanto tempo entre os pioneiros da computação.

A conclusão de *Descartes* recebia uma confirmação importante, da primeira programadora da história, a de que não é possível que as máquinas possam pensar. Contra as maiores aspirações dos pioneiros da computação, proclamava-se que as máquinas eram mera extensão dos poderes humanos, tal como um óculos estende a visão, um motociclo a velocidade, entre outros (*There are in all extensions of human power, or additions to human knowledge, various collateral influences, besides the main and primary object attained*). Durante um século ninguém conseguiria responder a esse vaticínio de modo competente.

[54] LOVELACE, Ada *apud* BABBAGE, [1942].
[55] *Ibidem*.

1.1.5 Turing e as máquinas que pensam

Alan Turing (1912-1954) foi um prodígio de sua época. Foi o pioneiro da teoria matemática da ciência da computação, também denominada Teoria dos Autômatos (*Automaton Theory*). A sua importância é tão significativa que passaram a ser denominadas "Máquinas de Turing" (*Turing Machine*)[56].

Turing irá produzir a primeira resposta consistente às objeções cartesianas à possibilidade de as máquinas utilizarem competências linguísticas. A resposta de *Turing* aparecerá no revolucionário artigo publicado sob o título "Computing machinery and intelligence", na *Revista Mind* em 1950[57]. O texto principia com a ambiciosa pergunta: "Podem as máquinas pensar?" (*Can machines think*?). De imediato o autor descarta o caminho tradicional de interpretação por meio da análise dos conceitos "máquinas" e "pensar", como pressuporia uma análise filosófica tradicional. Ao contrário propõe uma estratégia diversa e menos ambígua.

Turing sugere um artifício denominado "jogo da imitação", no qual um entrevistador precisa adivinhar quem seria o humano em uma entrevista às cegas. A pergunta do autor é o que aconteceria no caso de o entrevistador tomar por erro a máquina como se fosse o humano. Para ele esse questionamento substituiria a tradicional pergunta "podem as máquinas pensar?". Na impossibilidade de se definir com clareza o conceito de "inteligência" ou "pensar", o autor propõe verificar se uma máquina poderia imitar com sucesso o comportamento humano a ponto de confundir um observador imparcial. *Turing* respondia *Descartes* e diria que não existem limitações plausíveis para acreditar que uma máquina não pudesse responder satisfatoriamente às interlocuções propostas, sem ser confundida com um ser humano.

A partir desse momento, *Turing* lista nove possíveis objeções à tese de que as máquinas podem pensar e as contradita. A primeira objeção é teológica e afirma que pensar é um atributo da alma imortal, considerando que somente homens e mulheres podem possuir alma, então nenhum animal ou máquina pode pensar. Responde que essa afirmação é meramente especulativa, dado que Ele poderia atribuir alma ao ente que desejasse. Pensar o contrário seria uma restrição injustificada de Seu poder. Outra limitação seria matemática. Não haveria a possibilidade de uma máquina de estados finitos desempenhar ações razoavelmente em todos os casos. Em resposta poder-se-ia afirmar que a vitória humana seria momentânea.

---

[56] HAMMING, R. W. The Theory of Automata. Reviewed work: computation: finite and infinite machines by Marvin L. Minsky. *Science*, New Series, v. 159, n. 3818, p. 966-967, 1968.
[57] TURING, A. M. Computing Machinery and Intelligence. *Mind*, v. 49, p. 433-460, 1950.

A objeção de consciência afirmaria que a máquina jamais poderia sentir a felicidade de uma vitória ou a tristeza de uma decepção. *Turing* reconhece os mistérios da consciência, apesar de jamais alguém conseguir localizá-la, contudo, acredita que resolver esse problema não é antecedente ao jogo da imitação. Respeitosamente *Turing* deixa ao final a objeção de *Lady Ada Lovelace*. Ao argumento de que as máquinas não poderão criar nada e somente executam ordens em um sistema predeterminado, ele responde afirmando que talvez a matemática se referisse ao fato de que ao seu tempo a máquina analítica (*analitical engine*) não poderia criar algo.

Contudo, questiona-se, será que a autora pensava que *jamais* as máquinas poderiam criar algo? A essa objeção, a autora responde que as máquinas podem nos surpreender (*take us by surprise*), ou seja, mesmo que a programação e a arquitetura do sistema sejam muito bem formuladas, é possível que o cálculo inicial não preveja todas as possibilidades e arranjos. É razoável supor que os cálculos não tenham sido exaustivos ao infinito, capazes de cobrir todas as possibilidades e, portanto, as máquinas podem nos tomar de surpresa.

Um contra-argumento seria que a surpresa envolveria um "ato mental criativo" (*creative mental act*), em linha com a objeção da consciência. Dado que esse ato é exclusivo de seres conscientes, então não é possível que uma máquina nos tome de surpresa. Trata-se, conforme, *Turing*, de uma falácia derivada da percepção arrogante de que a mente é capaz de perceber todas as consequências de determinado ato e que a surpresa não poderia advir de um livro, uma planta ou uma máquina.

A objeção de *Lady Lovelace* de que as máquinas não podem aprender, mas tão somente executar ordens, conforme foram predestinadas, tal como o tear mecânico de *Jacquard*, é contraposta à situação de que as máquinas possuíam e, talvez, possuam ainda, limitada capacidade de armazenamento e processamento. Tal como os seres humanos se desenvolvem em fases (infância e adulta), as máquinas também seguiriam o mesmo curso. Elas estariam ainda em uma fase inicial. E de igual modo como as crianças precisam de educação para desenvolver-se, as máquinas aprenderiam. Os erros ou comportamentos randômicos comporiam um elemento fundamental na aprendizagem humana e de máquinas.

Turing, após superar essas objeções e outras listadas, roga que as máquinas se restrinjam à competição intelectual com os humanos, mas nada pode nos garantir que novas e complexas situações possam surgir. Devemos nos arriscar? Apesar de cauteloso, o autor afirma: *devemos tentar mesmo assim*.

1.1.6 Searle e as máquinas não pensam

O artigo de *Alan Turing* foi impactante. Após o seu escrito, houve uma proliferação significativa de textos sobre as possibilidades decorrentes de uma máquina de *Turing*. A alucinante produção científica, técnica e acadêmica sobre o tema, no entanto, não poderia nem brevemente ser apresentada neste trabalho, sem incorrer em grave falha no esquecimento de algum dos mais importantes nomes envolvidos na matéria[58].

Em 1956 ocorre em *Dartmouth* uma lendária Conferência, em que o termo *inteligência artificial* seria cunhado pelo Prof. *John McCarthy*[59]. Nesse local se reuniram os grandes pioneiros (*founding colleagues*) desse novíssimo campo de pesquisas, tais como: *Marvin Minsky*, *Oliver Selfridge*, *Ray Solomonoff*, e *Trenchard More*. Muitos desses estudos procuravam explorar as possibilidades vislumbradas por *Turing* e as suas possíveis objeções.

Herbert Simon, um dos pioneiros nos estudos em Inteligência Artificial, irá estabelecer quatro distinções sobre o conceito de artificial, como sendo coisas que: i) podem ser sintetizadas por humanos; ii) imitam o natural; iii) se caracterizam em termos de função ou adaptação ou iv) podem ser "*esquematizadas*" (*design*), em termos imperativos ou descritivos[60]. A Inteligência Artificial irá ser tomada em um ou vários desses sentidos. Mas a dúvida ainda primordial é quanto a possibilidade das máquinas imitarem a inteligência humana. O Teste de Turing estaria sob ataque ferrenho pelos gênios da computação.

Diversas limitações foram apresentadas ao Teste de Turing (TT). A primeira objeção dirigir-se-ia a sua natureza antropocêntrica, afinal, estar-se-ia comparando máquinas e humanos, como se não houvesse outro tipo de inteligência (animal ou artificial).

Algumas das objeções direcionadas foram quanto à limitação do teste às capacidades linguísticas da máquina. Passaria no TT o programa que melhor mimetizasse o comportamento falante humano, mas não haveria uma limitação do conceito de inteligência ao conceito de fala? Não existiriam animais inteligentes não falantes?

Por fim, questionavam-se os limites do conceito de inteligência limitada às competências externas dos "estados mentais". Saber-se-ia o resultado de um processo intelectivo, tal como uma soma, mas não existia explicação de como a máquina chegou àquele resultado. Ao perguntar-se à máquina quanto é 7 mais 2, ter-se-ia o resultado declarado 9, mas

---

[58] Veja-se BUCHANAN, Bruce G. A. (Very) Brief History of Artificial Intelligence. *AI Magazine*, v. 26, n. 4, Winter 2005.
[59] *Ibidem*.
[60] SIMON, Herbert. The Sciences of the Artificial. London/Cambridge: MIT Press, 1996, p. 05. Disponível em https://epdf.pub/the-sciences-of-the-artificial-3rd-edition-pdf-5eccdc2f3d0e8.html. Acesso dia 26.12.20 às 23:46.

não saberíamos quais o processos internos (*inner states*) que levariam àquele resultado[61]. As objeções não retiravam o brilhantismo do TT, mas apenas exigiam pesquisas ainda mais aprofundadas sobre o promissor campo que se abria.

Coube a *John Searle* (1932-) elaborar a mais importante e bem formulada objeção ao TT. O autor irá distinguir, corretamente, entre IA Forte (*strong*) e IA fraca (*weak*), conforme as suas funções. A IA fraca seria aquele campo em que as máquinas conseguem passar no Teste de Turing, performando competências linguísticas indiscerníveis de um ser humano. A IA forte sugere a possibilidade de máquinas que performam competências próprias de um ser humano, ou seja, não apenas aparentam como possuem igualmente todas as competências humanas, inclusive a consciência. O presente trabalho irá se dedicar tão somente ao conceito de IA forte.

O Teste de Turing, em sua versão limitada ou fraca, se dirigia tão somente às máquinas que mimetizavam competências linguísticas, mas não as realizavam realmente. O autor havia substituído a questão inicial "podem as máquinas pensar?" por uma nova versão: "podem as máquinas aparentar, persuasivamente, que pensam?".

Uma versão ampliada ou forte do Teste de Turing seria objeto de *John Searle*, que irá analisar e criticar a intuição comum de que o cérebro é um computador (*hardware*) e a mente, um programa (*software*). *Searle* retoma o problema de *Turing* no famoso artigo *Uma máquina poderia pensar*? (*Could a machine think*?) [62] e responde que somente uma máquina especial, com poderes causais (*internal causal powers*) semelhantes aos cérebros poderia pensar. O autor dirá que a IA forte trata do pensamento e esse possui relação com o programa, contudo, apenas o programa não é suficiente para o pensar.[63]

Searle apresenta duas proposições encadeadas: 1) a intencionalidade nos seres humanos (e animais) é produto da características causais do cérebro (*causal features of the brain*) e 2) instanciar um programa de computador não é uma condição suficiente de intencionalidade (*intentionality*). A conclusão de *Searle* é a de que toda tentativa de criar intencionalidade artificialmente (*strong AI*) deve duplicar os poderes causais do cérebro humano e não simplesmente elaborar um programa[64].

A primeira proposição retoma uma crítica direta ao dualismo cartesiano, que se fundamentava na distinção absoluta entre mente e corpo, com base na afirmação de que estes possuem naturezas distintas. Duas substâncias seriam idênticas se possuíssem as mesmas

---

[61] LAVELLE, Jane Suilin. 3 What is it to have a mind? *In*: CHRISMAN, Matthew; PRITCHARD, Duncan. *Philosophy for Everyone*. London/New York: Routledge, 2014, p. 49.
[62] SEARLE, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, v.3, n. 3, p. 417-457, 1980.
[63] SEARLE, 1980, p. 1.
[64] SEARLE, 1980, p. 1.

propriedades e seriam distintas em caso diverso, conforme a denominada Lei de Leibniz. Considerando que posso duvidar da existência de meu corpo, devido a um gênio maligno, então mente e corpo possuem naturezas distintas.

A crítica direta à tese cartesiana se denominou de *teoria da causação*[65]. Como pode o espírito causar efeitos no corpo, tal como o movimento, se eles possuem naturezas distintas? Talvez, uma das primeiras formuladoras dessa objeção tenha sido a *Princesa Elisabete da Boêmia* (1618-1680), em uma de suas muitas cartas a *Descartes*[66].

Searle adota a tese de que a intencionalidade é produto das características causais do cérebro, ou seja, há interação entre a mente e o cérebro, e ambos possuem a mesma natureza (*brain processes cause consciousness*)[67]. Adota-se a teoria da identidade de que os estados mentais (*mental states*) e os estados físicos são idênticos (*physical states*). Sem a teoria da causação, a interação entre mente e cérebro se tornaria uma problema obscuro, e as relações entre ambos, fantasmagórica.

Searle não estaria somente objetando *Descartes*, porém igualmente a posição dominante entre os teóricos da computação, para os quais a IA forte fundar-se-ia na ideia de que a mente e a consciência não são processos concretos, físicos ou biológicos, mas formais e abstratos. Como exemplo, cita o autor a definição de mente para *Daniel Dennett* e seu coautor *Douglas Hofstadter*, de que a mente é algo abstrato e separado de qualquer aparato físico (*an abstract sort of thing whose identity is independent of any particular physical embodiment*)[68]. Para demonstrar a segunda proposição, *Searle* irá formular o célebre argumento denominado de Sala Chinesa (*chinese room argument*), que irá orientar praticamente todos os debates filosóficos sobre inteligência artificial a partir de então.

O argumento de *Searle* possui como pressuposto o problema da *tematicidade* (*aboutness*), afinal, todo estado mental é sobre algo[69]. Desse modo, aparentar competência linguística não é o mesmo que possuí-la. Dito de outro modo, instanciar um programa de computador não é uma condição suficiente de intencionalidade (*intentionality*).

Há três elementos no experimento da Sala Chinesa, um agente externo que envia caracteres chineses, uma pessoa dentro da sala que não sabe o significado desses símbolos, mas

---

[65] LAVELLE, 2014, p. 38.
[66] Cf. a autora: "*Given that the soul of a human being is only a thinking substance, how can it affect the bodily spirits, in order to bring about voluntary actions*?", *in* Correspondence between Descartes and Princess Elisabeth René Descartes and Princess Elisabeth of Bohemia. Disponível em: https://www.earlymoderntexts.com/assets/pdfs/descartes1643_1.pdf. Acesso em: 31 maio 2020 às 14:07.
[67] SEARLE, J. R. *The mystery of consciousness*. John R. Searle and exchanges with Daniel C. Dennett and David Chalmers. New York: The New York Review of Books, 1997, p. 191.
[68] SEARLE, 1997, p. 192.
[69] Conforme *Jane Lavelle*: "*It is clear that my thoughts are about things*" *in* LAVELLE, 2014, p. 39.

possui um livro que indica o caractere correspondente e um agente externo que recebe os símbolos correspondentes. O experimento apresenta uma estrutura sintática ao descrever somente o uso de regras para a manipulação de símbolos.

O computador atuaria de mesmo modo. Ele sabe como manipular símbolos, conforme um procedimento previamente determinado, porém não conhece o que está sendo processado, ou seja, não consegue atuar no nível semântico da comunicação. Para *Searle* a máquina falha em determinar a *tematicidade* (*aboutness*) ou entendimento dos símbolos manipulados[70].

Searle irá retomar o tema na sua obra *O mistério da consciência* (*The Mystery of consciousness*), de 1998. Conforme o autor, símbolos abstratos não possuem poderes causais (*causal powers*), capazes de produzir consciência. Os poderes causais estão no meio de implementação, tal como no cérebro. Não se trata, contudo, de uma excepcionalidade humana. Para *Searle* qualquer sistema complexo o suficiente poderia ser um "meio de implementação"[71].

Se, por uma lado, ele rejeita o dualismo, por outro lado, nega igualmente uma visão reducionista da mente a elementos puramente físicos. Há uma rejeição do materialismo reducionista. A consciência é uma característica real e intrínseca de certos sistemas biológicos, mas que não pode ser reduzida[72] tal como outras propriedades físicas, como o sólido em termos de estrutura molecular[73].

Nesse ponto *Searle* parece adotar uma tese cartesiana, a indivisibilidade da mente. A consciência, para o autor, não é reduzível no sentido em que outras propriedades o são, porque é essencialmente uma ontologia em primeira pessoa. Não há como ser reduzida a uma ontologia objetiva ou em terceira pessoa[74]. Ontologia em primeira pessoa significa essencialmente a realização de experiências subjetivas. O cérebro possui a incrível capacidade de produzir experiências e estas existem apenas quando sentidas por um ser humano ou algum agente animal[75].

Mas qual a relação entre consciência e intencionalidade? *Searle* responde a esse questionamento afirmando categoricamente: "consciência é intencional" (*consciousness is*

---

[70] LAVELLE, 2014, p. 51.
[71] "*Any system-from men sitting on high stools with green eyeshades, to vacuum tubes, to silicon chips-that is rich enough and stable enough to carry the program can be the implementing médium*", *in* SEARLE, 1997, p. 210.
[72] O conceito de "redução" não é unívoco, como explica Searle. Ele apresenta uma dúzia de sentidos diversos no capítulo 5 (*Reductionism and the irreducibility of consciousness*) (SEARLE, J. R. *The Rediscovery of the Mind*. Cambridge: MIT, 1995).
[73] SEARLE, 1997, p. 211.
[74] Cf. *Searle*: "*consciousness has a first-person or subjective ontology and so cannot be reduced to anything that has third-person or objective ontology*", *in* SEARLE, 1997, p. 212.
[75] Cf. o autor: "*first-person ontology is this: biological brains have a remarkable biological capacity to produce experiences, and these experiences only exist when they are felt by some human or animal agent*", *in* SEARLE, 1997, p. 212.

intentional)[76]. Toda a consciência o é em uma perspectiva em primeira pessoa. Para *Searle* toda intencionalidade é uma consciência perspectiva. Experimentamos sensações e outras formas de experiências conscientes em determinado aspecto. Ver é ver em determinado aspecto. Todas as formas de representação de objetos o são em determinado aspecto. Toda intencionalidade possui uma forma "aspectual" (*every intentional state has what I call an aspectual shape*)[77].

Partindo desses pressupostos, o autor irá afirmar a sua tese sobre a "conexão entre consciência e intencionalidade: de que somente seres conscientes possuem intencionalidade e qualquer ato inconsciente intencional é no mínimo potencialmente consciente" (*only a being that could have conscious intentional states could have intentional states at all, and every unconscious intentional state is at least potentially conscious*)[78].

É indubitável que as teses de *Searle* foram e são ainda uma demarcação importante nos debates filosóficos sobre inteligência artificial. Cabe-nos verificar os debates posteriores, críticos ou não, às inovações apresentadas por esse importante autor.

1.1.7 A possibilidade da inteligência artificial forte

Os argumentos de *Searle* foram tão desconcertantes que implicaram sucessivas respostas e tentativas de superação da Tese da Sala Chinesa. Diversas sugestões foram apresentadas, para superar a objeção à IA Forte. Não é nossa pretensão esgotar todas as possibilidades, mas tão somente listar algumas das principais respostas aos limites propostos pela tese de *Searle*.

A primeira resposta poderia ser no sentido de que, no experimento da Sala Chinesa, erraria ao afirmar que o indivíduo que coleta o cartão (*input*) e entrega o símbolo correspondente na saída (*output*) seria a "mente computacional". Na verdade esse indivíduo seria apenas uma parte da mente, e a sala inteira seria um sistema computacional[79]. Não importa se o indivíduo, na condição *implementador* das entradas e saídas de caracteres, entenda o seu significado, o que importa é a compreensão do sistema.

Mesmo que o sistema fosse considerado uma mente computacional, ainda faltaria a causalidade necessária para o surgimento da IA forte[80]. Não haveria conexão adequada entre a

---

[76] SEARLE, J. R. *The rediscovery of the mind*. Cambridge: MIT, 1995, p. 51.
[77] SEARLE, J. R. *The Rediscovery of the Mind*. Cambridge: MIT, 1995, p. 51.
[78] SEARLE, 1995, p. 51.
[79] São defensores dessa tese: *Jack Copeland, Daniel Dennett, Douglas Hofstadter, Jerry Fodor, John Haugeland, Ray Kurzweil* e *Ned Block*. Veja-se COLE, David. The Chinese Room Argument. *In*: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy*. Spring 2020 Edition. Disponível em: https://plato.stanford.edu/archives/spr2020/entries/chinese-room/. Acesso em: 31 jul. 2020 às 00:04.
[80] HARNAD, S. *Minds, machines, and searle*: what's right and wrong about the Chinese room argument. Preston and Bishop (eds.), 2002. p. 294-307.

mente e o cérebro. Para *Hanard* a ideia de mecanismo é central na relação mente-programa. É um sistema físico operando conforme leis causais (físicas). Um aeroplano seria um mecanismo. Faltaria uma conexão causal apropriada entre o sistema computacional e ambiente[81].

A objeção da Sala Chinesa como sistema é insuficiente para superar as principais teses de *Searle* e não toca nos temas mais importantes: causalidade e intencionalidade. Novas abordagens seriam testadas.

A teoria computacional da mente defendeu que a mente é um programa (*software*), no sentido de que pensar é uma forma de manipulação simbólica no nosso cérebro. Se um sistema adequado for programado, com um software certo, então poderíamos dizer que ele *pensa*. O software apropriado poderia dar ao sistema adequado uma *mente* ou *intencionalidade*.

A ideia de que nossa mente é apenas um programa que roda no cérebro apresenta-se deveras reducionista. Claramente ela é muito mais sofisticada do que uma calculadora de símbolos lógicos. A mudança de perspectiva se situa em uma nuance da Tese da Sala Chinesa, em que se afirma que os meros símbolos abstratos não possuem poderes causais (*causal powers*). A causalidade estaria no cérebro.

E se o programa, no entanto, estivesse em um sistema adequado, tal como o cérebro? Não teríamos aqui uma resposta à objeção de *Searle*? Poder-se-ia afirmar que tal sistema adequado não existe ainda, mas essa formulação não afastaria a possibilidade virtual de criação de tal máquina. Não haveria um problema de impossibilidade lógica, tal qual na separação entre mente e cérebro na tese dualista cartesiana.

Poder-se-ia objetar ainda que faltaria a essa máquina a experiência que é constitutiva da intencionalidade e da consciência. O "sistema computacional humano" não é meramente um processo interno, mas envolve uma interação com o meio ambiente[82]. E se dotássemos esse sistema adequando mecanismos de interação com o meio ambiente[83], tal como provemos humanos de novas formas de ouvir, ver mais longe, andar mais rápido ou mesmo andar, por meio de próteses?

A tese de um autômato (*robot*) com sensores para interagir com o meio ambiente, tal como ver, ouvir e mesmo sentir, superaria o obstáculo da conexão entre mente e ambiente, por

[81] HARNAD, S. Minds, machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 1, n. 1, p. 5-25, 1989.
[82] THOMPSON, E. *Mind in life*: biology, phenomenology, and the sciences of mind. Cambridge and London: Harvard University Press, 2007, p. 8.
[83] FODOR, J. A. Searle on what only brains can do. *The Behavioral and Brain Sciences*, p. 431-432, 2010, p. 431.

meio de experiências sensoriais únicas por parte do autômato[84]. A resposta propõe uma mudança de uma tese computacional da mente para uma tese robótica da mente. A inteligência artificial deixaria de ser um programa instalado no cérebro e passaria a ser entendida como um sistema incorporado no cérebro (*embodied AI*). Haveria a passagem do funcionalismo simbólico para o funcionalismo robótico, ou seja, da ideia de que as funções mentais são funções simbólicas para a ideia de que existem funções simbólicas assentadas em funções não simbólicas (sensoras, motoras, associativas, entre outras primárias)[85].

Outra condição seria necessária, pois não basta ter o sistema incorporado. A intencionalidade decorre da experiência, portanto, o autômato deveria estar imerso no seu ambiente ou, melhor dizendo, situado[86]. Deveria possuir as conexões corretas com o ambiente, de modo a produzir as corretas atitudes proposicionais em relação ao mundo.

Após sucessivas investidas, as objeções contra a Tese da Sala Chinesa parecem ter tomado corpo e vislumbrado a possibilidade de que talvez as máquinas pudessem pensar e, incrivelmente, adquirir consciência. Assim, adota-se a distinção realizada por *Searle* em IA forte e fraca. Sendo que *o conceito de IA a ser utilizado será o de que máquinas podem hipoteticamente possuir, dada a superação de limites tecnológicos, intencionalidade e consciência artificial similar à humana.*

O presente trabalho, portanto, partirá assim da assunção da impossibilidade de se descartar a possibilidade de surgimento de um sistema artificial fundado em IA forte, bem como da probabilidade não desprezível de seu surgimento no futuro. Partindo desse pressuposto, pergunta-se: poderiam tais máquinas darem origem a proposições morais artificiais? Poderiam surgir agentes morais artificiais? Para responder a esses questionamentos primeiro iremos tratar da relação entre teorias morais e inteligência artificial.

1.2 ÉTICA E INTELIGÊNCIA ARTIFICIAL

1.2.1. Das diversas acepções de uma ética artificial

[84] PIEK, Matthijs. *The Chinese Room and the Robot Reply*. Tese. (Doutorado) – Tilburg University. Philosophy of Science and Society, Holanda, [*s.d.*]. Disponível em: http://arno.uvt.nl/show.cgi?fid=146015, acesso 02 jun. 2020 às 06:44.
[85] HARNAD, 1989, p. 8.
[86] ZIEMKE, T. *The body of knowledge*: On the role of the living body in grounding embodied cognition, 2016. Disponível em: https://www.researchgate.net/publication/306350827_The_body_of_knowledge_On_the_role_of_the_living_body_in_grounding_embodied_cognition. Acesso em: 31 jul. 2020 às 00:07.

A presente seção tratará da possibilidade filosófica de uma ética artificial. Existem três sentidos possíveis desse questionamento: da ética *aplicada* à inteligência artificial (IA), *decorrente* de sua aplicação ou da *própria* da IA.

No primeiro caso, trataremos dos limites e das diretrizes éticas para a pesquisa e o desenvolvimento da IA. Poder-se-ia questionar nesse campo quais são os princípios que devem nortear as pesquisas sobre autômatos, robôs e algoritmos. Somente esse campo de pesquisas apresenta inúmeros desafios.

Como devem ser estruturados os algoritmos de modo a proteger os indivíduos de mal uso da IA? O exemplo mais famoso de uma investigação nesse sentido se encontra na proposta de *Isaac Asimov*[87], originalmente publicada em 1942, com as três leis da robótica:

> *First Law — A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
>
> *Second Law — A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*
>
> *Third Law — A robot must protect its own existence if such protection does not conflict with the First or Second Laws.*

A proposta aparentemente simples e compreensível de *Asimov* não impediria dilemas éticos: e se um, robô para proteger um humano, tivesse de causar mal a outro humano? E se o humano causasse mal a si mesmo? Poderia o sistema agir para impedir? Haveria conflito entre o livre-arbítrio humano e o paternalismo artificial? O que significaria "causar dano"? Alimentar-se mal? Qual a extensão desse dever de cuidado ou tutela? As perguntas poderiam suceder em número infinito. Poder-se-ia ainda questionar se seria consistente tratar um robô superinteligente como um mero escravo ou objeto e não como um sujeito de direitos?[88]

Digamos que possamos adequadamente listar, para cada pergunta sobre como o algoritmo em IA deveria agir em determinada situação, uma resposta adequada e que cada resposta seja estruturada de modo consistente com as demais respostas. Devemos igualmente pressupor que não somos infalíveis e que não somos oniscientes sobre todos os fatos contingentes. Ou seja, o programa é elaborado por um ser humano para ser aplicado por um sistema artificial inteligente.

---

[87] ASIMOV, I. *Eu, Robô*. 2. ed. em português. Tradução de Luiz Horácio da Matta, 1969. Disponível em: http://bibliotecadigital.puc-campinas.edu.br/services/e-books/Isaac%20Asimov-2.pdf. Acesso em: 31 jul. 2020, às 00:14.

[88] ANDERSON, S. L. Asimov's "three laws of robotics" and machine metaethics. *AI & Soc*, v. 22, 2008, p. 477-493, p. 493.

Partindo desses pressupostos, chegaremos ao famoso *Paradoxo do Prefácio*. O problema foi originalmente proposto por *Makinson* (1965), quando sugere que um autor escreve um livro e admite que alguma das sentenças da obra pode ser falsa. Assim, um autor teria escrito uma obra com um conjunto *x* de proposições e, para cada uma, haveria uma crença justificada *S*. Contudo, havendo a crença de que podem existir proposições falsas, haveria inconsistência com a afirmação de que cada uma das proposições possui um crença justificada.

Duas conclusões podem surgir: de um lado, a de que é possível que um sujeito possua um conjunto de crenças inconsistentes e racionais[89]; ou a de que crenças inconsistentes podem ser apoiadas por qualquer tipo de evidência[90]. Do mesmo modo, uma listagem exaustiva e compreensiva sobre a ética aplicada à inteligência artificial deveria aceitar que alguma proposição moral venha a ser inconsistente com outra ou fundada em evidências enganadoras.

Estabelecer os princípios de boa formação de algoritmos éticos seria o Santo Graal da ética artificial. Infelizmente, ela teria de aceitar a incompletude e a inconsistência eventual como um destino manifesto. Uma alternativa seria utilizar uma estratégia distinta. Dado que o estabelecimento *ex ante* de um catálogo completo, consistente e racional de princípios éticos aplicados à IA é impossível, quem sabe permitir que o sistema ético seja flexível o bastante para solucionar dilemas que se apresentarem? E se o sistema puder aprender a como melhor escolher, a partir de um conjunto de regras predefinidas, qual a decisão ética a tomar?

Outro campo de pesquisas é justamente sobre os desafios éticos decorrentes da *aplicação* da IA. Quais as consequências éticas do uso de IA no convívio humano. Alguns campos têm se tornado tormentosos, tais como a utilização de drones, *chatbots*, sistemas de reconhecimento facial, entre outros. Muitos desafios surgem desse uso cada vez mais disseminado e ubíquo dos sistemas de inteligência artificial. Quem será o responsável pelos danos? O programador, o proprietário do *robot*, a sociedade? Trata-se de um tema importante para a Filosofia do Direito. Poder-se-ia mesmo atribuir ao próprio *robot* ou sistema artificial a responsabilidade pelos seus atos, com as sanções proporcionais?

Nenhum dos casos acima será objeto de nossa pesquisa no presente trabalho, apesar de sua clara importância, prática ou teórica. Nos dois temas, a ética vislumbra a *inteligência artificial como objeto*[91], ou seja, como instrumento utilizado pela vontade e racionalidade humana. Em certa medida, tal problema não se distancia tanto do uso de outra tecnologia e suas

---

[89] RODRIGUES, L. R. O paradoxo do prefácio generalizado. *Intuitio*, Porto Alegre, v. 11, n. 1, p. 7-18, 2018.
[90] CONEE, E. A. The Preface Paradox. *In*: DANCY, J.; SOSA, E.; STEUP, M. (ed.). *Companion to Epistemology*. 2. ed. Malden, MA: Wiley-Blackwell, 2009. p. 604-605.
[91] MÜLLER, Vincent C. Ethics of Artificial Intelligence. *In*: ELLIOTT, Anthony (ed.). *The Routledge social science handbook of AI*. London: Routledge. p. 1-20.

consequências decorrentes. Esses desafios surgiram em diversos momentos históricos sobre o uso das máquinas na Revolução Industrial, dos aviões em guerras, da energia nuclear, entre tantos outros exemplos.

O desafio do presente trabalho, portanto, é investigar a possibilidade filosófica de uma ética artificial como decorrente da vontade e racionalidade *própria* de um sujeito artificial. A ética artificial tratará do relacionamento das máquinas inteligentes com os agentes humanos e artificiais e deverá se deter sobre as considerações quanto a valores, escolhas morais e seus dilemas.

A ética clássica *trata do agir humano*[92] e, de certo modo, pressupõe uma antropologia filosófica[93]. Para existir uma ética do agir artificial, deverá o sujeito moral artificial possuir capacidade de escolhas. A expansão da ética para seres não humanos, de mesma forma que se cogita, e alguns aceitam, para os animais, é um dos problemas mais intrigantes de nosso tempo. Essas são as questões que se pretende a seguir investigar.

1.2.2 Da possibilidade de um status moral da inteligência artificial

Existiria a possibilidade lógica de uma ética *própria* à inteligência artificial ou esta é necessariamente externa? As consequências da resposta a esse questionamento são imensas. Uma tarefa é analisar a inteligência artificial como *objeto* ou artefato humano que pode ou não ser bem utilizado no agir humano; outro problema muito distinto é tratar da possibilidade de um agir por parte de um sujeito artificial. São significativas as implicações e os dilemas éticos decorrentes da aceitação desta última possibilidade.

A definição do status moral da inteligência artificial é fundamental para a solução de dilemas e problemas, que podem surgir do tratamento das máquinas. Em nosso entender, a resposta passa pela admissão da existência de uma escala de atribuição de estatura moral (*moral status*) às máquinas.

Em um primeiro nível, poderíamos considerar as máquinas como apenas objetos destituídos de qualquer consideração moral em si, tal como uma pedra ou um pedaço de madeira[94]. Seriam meros artefatos à disposição humana, extensões das habilidades físicas, ampliando sua força, velocidade e inteligência. São o fruto do desejo humano de escapar à miserabilidade de sua condição, que remonta à mitologia grega e ao mito de Prometeu. A lenda

---

[92] HERRERO, F. Javier. A Ética Filosófica de Henrique Cláudio de Lima Vaz. *Síntese*, Belo Horizonte, v. 39, n. 125, p. 393-432, p. 431, 2012.
[93] HERRERO, 2012, p. 431.
[94] Para os fins do presente trabalho, vamos desconsiderar os argumentos de atribuição moral *ground* para a natureza.

conta que, para superar sua frágil condição, a humanidade recebe de seu titã benfeitor o fogo, representando o domínio da técnica sobre as agruras da natureza.

O elogio à técnica, à tecnologia e ao progresso atinge o seu ponto mais importante com os estudos de *Francis Bacon* (1561-1626). Nada ilustraria melhor esse desejo de ciência e tecnologia do que a sua declaração na obra "New Atlantis" de que o objetivo do empreendimento humano, representado na fundação dessa nova sociedade, era explicitamente o "aumento dos limites do império humano, para efetuar todas as coisas possíveis"[95]. A modernidade talvez trouxesse, finalmente, o domínio da natureza ao alcance das mãos humanas.

Para *Wiener* (1960), o pai da cibernética e um dos pioneiros das pesquisas em IA, os propósitos das máquinas (inteligentes) será aquele que nós desejarmos[96]. Os sistemas de inteligência artificial serão artefatos decididos, desenhados, implementados e usados por seres humanos[97].

Assim, os sistemas de IA serão artefatos bons ou ruins conforme o seu uso pela humanidade. Sua utilização pode tender ao bem ou ao mal. Podem possuir um lado virtuoso ou maldoso[98]. Serão os indivíduos que irão escolher como usá-los. *Wiener* irá alertar para o lado nefasto desse desejo de senhorio ou dominação[99]. Segundo o autor, o deleite do novo senhor das máquinas atinge o ápice quando ele descobre o poder de criar um "escravo" (*robot*[100]) subordinado, subserviente, eficiente, que nunca demanda nada para si nem exige qualquer tratamento melhor do que um pedra de basalto.

O ápice desse drama aparece na obra de *Karel Čapek*, que enuncia um mundo onde os humanos criam robôs, que ao fim tornam os próprios humanos obsoletos[101]. *Wiener* utiliza a imagem extremamente forte do *Golem* para expressar essa limite ético do mau uso dos sistemas inteligentes. *Golem* era uma figura da mitologia judaica, com feições antropomórficas e animado a partir do barro[102], que passaria de um servo obediente a um ser violento.

---

[95] "*The end of our foundation is the knowledge of causes, and secret motions of things; and the enlarging of the bounds of human empire, to the effecting of all things possible*". BACON, Francis. *New Atlantis*. Project Gutenberg, 2020. Disponível em: http://www.gutenberg.org/ebooks/2434. Acesso em: 05 jun. 2020 às 01:24.
[96] "*[W]e had better be quite sure that the purpose put into the machine is the purpose which we really desire*". WIENER *apud* DIGNUM, Virginia. *Artificial intelligence*: foundations, theory, and algorithms. Cham: Springer, 2019, p. v.
[97] DIGNUM, Virginia. *Artificial intelligence*: foundations, theory, and algorithms. Cham: Springer, 2019, p. v.
[98] ARKIN, Ronald C. *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press, 2009.
[99] WIENER, Norbert. *God and Golem, Inc.* A comment on certain points where Cybernetics impinges on Religion. Cambridge: MIT Press, 1963. p. 55.
[100] Atribui-se a origem etimológica da palavra robô ao termo checo "robota", de autoria de *Karel Čapek*, na ficção científica R.U.R. (*Rossum's Universal Robots*) de 1921. A palavra teria por significado "esforço" ou "servidão".
[101] TCHAPEK, Karel. *A fábrica de robôs*. São Paulo: Hedra, 2012.
[102] ROSENBERG, Yudl. *The golem and the wondrous deeds of the Maharal of Prague*. Trad. Curt Leviant. New Haven: Yale University Press, 2007.

Nesse primeiro nível, considerações éticas subjacentes se dirigem ao uso responsável da tecnologia e das máquinas[103]. Como devem os veículos autônomos se comportar em situações complexas, tal como no Dilema do Bonde (*Trolley Dilemma*[104])? Podem drones militares autônomos decidir quando disparar e matar seus alvos?

Em todos esses casos, a palavra *autônomo* é utilizada de modo impróprio, dado que as decisões de carros e drones são desenhadas de fora, por um terceiro, um *designer* humano[105]. Não exatamente um ente que se movimenta sozinho, mas que pareça ter movimento próprio, tal como um fantoche, no dizer de *Aristóteles*. O carro e o drone são apenas artefatos ou extensões de partes humanas, tal como o óculos é para a visão ou a bicicleta, para as pernas. Há uma clara confusão entre autonomia mecânica e autonomia moral[106]. Um carro autônomo é um mecanismo que atua sem intervenção humana direta, mas de modo algum delibera ou age em sentido próprio, em primeira pessoa. A diferença, importante e significativa, está no fato de que esses novos artefatos reproduzem mecanismos de escolhas éticas, deliberadas, arquitetadas, desenhadas e implementadas *ex ante,* para uso e produção de consequências, conforme o modelo pensado por humanos.

Nesse primeiro nível, as considerações éticas subjacentes se dirigem ao uso responsável da tecnologia e das máquinas[107]. Sob essa perspectiva, as máquinas e a responsabilidade de seu uso serão consideradas conforme o seu grau de complexidade ética incorporado nas máquinas ou na sua programação. Vejamos, por exemplo, uma máquina automática que dispensa mercadorias (*vending machine*), o caixa automático de um banco (ATM), um controlador de velocidade em vias públicas (pardal), todos esses casos não diferem muito do tear mecânico de *Jacquard* ou da *Analitical Machine* de Babbage. *Ada Lovelace* estava absolutamente correta quando afirma que essas máquinas nada criam e muito menos possuem protagonismo ético. Nenhuma delas possui dilemas éticos a resolver ou um agir moral a considerar.

Tomemos por exemplo um controlador de velocidade urbana (radar). Ao verificar, por alguma forma de sensor, a ultrapassagem de velocidade, ele atesta o descumprimento da legislação de trânsito. Poderíamos afirmar que esse equipamento cumpre uma função normativa mínima e age segundo um silogismo lógico. Se andar acima de tal velocidade, a sanção será a imposição de multa em tantas unidades monetárias, com concomitante aplicação de restrições

---

[103] DIGNUM, Virginia. *Artificial intelligence*: foundations, theory, and algorithms. Cham: Springer, 2019.
[104] FOOT, Philippa. The problem of abortion and the doctrine of double effect. *Oxford Review*, v. 5, p. 5-15, 1967.
[105] Cf. "*(their aren't) able to render moral decisions on their own*". ETZIONI, A.; ETZIONI, O. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, v. 21, n. 4, p. 403-418, 2017.
[106] ETZIONI, A.; ETZIONI, O, 2017, p. 410.
[107] DIGNUM, *Artificial intelligence*…, 2019.

na habilitação do motorista infrator. Esses mecanismos tecnológicos funcionam como *agentes lógicos*, dotados de um programa interno que determina rotinas lógicas, mas sem qualquer apreciação ética no seu agir automatizado. Trata-se de mecanismos que seguem regras normativas deliberadas coletivamente, em lei ou decisão judicial, consagradas em regras e normas, não havendo nenhum espaço para deliberação pela máquina[108].

Não existe deliberação ética na entrega de uma mercadoria por uma máquina de vendas automática ou no registro da ultrapassagem do limite admitido. Todas as considerações éticas são externas ao mecanismo e se dirigem às escolhas humanas na arquitetura normativa da máquina. Qual a razão de a máquina estar em algum lugar e não noutro, por que multar determinados automóveis ou condutores e não outros? Todos os questionamentos éticos se dirigem aos formuladores e não à máquina[109].

Poderemos ter, também, nesse nível máquinas que atuam como *agentes normativos*, ou seja, dotados de uma programação em lógica deôntica *standard*. A máquina performaria operações em lógica deôntica, tais como: permitido, proibido e obrigatório.

Michael Anderson e *Susan Anderson* denominam estas máquinas como "agentes éticos implícitos", programadas para suportar comportamentos éticos ou evitar os comportamentos antiéticos. A programação dessas máquinas envolveria inserir a ética no sistema (*putting ethics into a machine*[110]). Nesse sentido, os autores apresentam como exemplo os ATMs que são programados para evitar fraudes bancárias ou aviões que são programados para garantir a segurança dos passageiros. Para eles essa modalidade de *software* que requer considerações éticas deveriam receber um determinado grau de reconhecimento ético (*so at least this sense of* '*machine ethics' should be accepted by all as being desirable, lack moral status*)[111].

Cremos que todos os pressupostos utilizados pelos autores não se sustentam. Em nossa opinião nenhum desses casos apresenta a inserção da ética na máquina, tampouco elas poderiam ser denominadas "agentes éticos implícitos". O fato de performarem atos lógicos normativos não implica que a máquina compreenda o código ético inserido. Todos os elementos da máquina, lógicos ou normativos, são produzidos e controlados externamente.

A arquitetura, o desenho, o algoritmo, as funções, a implementação e o uso são fruto de uma mente humana que projeta a sua racionalidade por extensão em uma máquina. Jamais

---

[108] Cf. "*Here, there is very little need for moral deliberations and decision-making (though there is a need to 'teach' the driverless carto comply*)". ETZIONI, A.; ETZIONI, O., 2017, p. 417.
[109] ANDERSON, Michael; ANDERSON, Susan Leigh. The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines*, v. 17, 2007, p. 01-10. Disponível em: https://link.springer.com/article/10.1007%2Fs11023-007-9053-7. Acesso em: 05 jun. 2020, às 23:07.
[110] ANDERSON, M.; ANDERSON, S. L., *ibidem*.
[111] ANDERSON, M.; ANDERSON, S. L., *ibidem*.

poderiam ser chamadas de agentes éticos, no máximo de máquinas normativas, operando um sistema de lógica deôntica *standard*. Em nenhum desses casos, a máquina opera deliberações morais. Tão somente opera conforme regras preestabelecidas, em uma proposição do tipo: dado o estado de coisas "x", deve aplicar tal solução normativa. Se existe algum agente ético implícito, este será o programador da máquina, que incluiu diretrizes normativas no *software*. A máquina é somente uma extensão do agente. Não é o ATM que está realizando comportamentos éticos ou antiéticos, mas o próprio banco, por meio do artefato.

Os autores ainda utilizam denominações ainda mais complicadas, tais como *Explicit Ethical Agents* e *Autonomous Ethical Agents*. O primeiro não agiria independentemente de agentes humano, como, por exemplo, no caso de um consultor ético para humanos, ao estilo de assistente pessoal artificial. Os *agentes éticos autônomos* poderiam, na visão dos autores, calcular e tomar decisões éticas autonomamente como no caso de um robô militar, que decide logicamente a melhor ação em um face de um dilema ético. Não cremos que nenhum dos dois casos represente verdadeiramente uma situação em que estaríamos perante um agente ético. Um robô militar não difere substancialmente de qualquer outra arma, como, por exemplo, de natureza química ou nuclear[112].

O fato de o programa possuir uma biblioteca ampla de ações a serem tomadas, um algoritmo eficiente para determinar o curso de ação mais adequado perante determinada situação, não caracteriza uma máquina como sendo um *agente ético*. Conforme vimos, um agente ético deve sê-lo por características próprias e não agir conforme uma programação externa predeterminada. A ética deve ser interna e não externa ao autômato. Ela deve ser fruto de um agir em primeira pessoa e nunca como instrumento de um terceiro, ou seja, deve garantir a sua natureza subjetiva e não ser objeto da ação de outrem. Essa é a diferença entre uma máquina lógica e uma máquina moral. Uma possui autonomia mecânica e a outra, autonomia moral.

Dadas essas conclusões parciais, poderíamos questionar: será possível a existência de uma máquina moral? O que garantiria um *status* moral para um sistema inteligente? Esse é nosso próximo objetivo, determinar os elementos que caracterizam o conceito de *agente moral*.

1.2.3 Da centralidade ética do conceito de sujeito

[112] JOHNSON, A. M.; AXINN, S. The morality of autonomous robots. *Journal of Military Ethics*, v. 12, n. 2, p. 129-141, 2013.

Dois critérios têm sido utilizados para se determinar o estatuto moral de um agente: *senciência e personalidade* (*sapience or personhood*). Eles têm sido caracterizados da seguinte forma[113]:

> Senciência (*Sentience*): capacidade por experiência ou *qualia*, tal como a capacidade para sentir ("*the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer*"); e
>
> Personalidade (*Sapience or personhood*): é o conjunto de capacidades associadas à inteligência superior, tal como autoconsciência e racionalidade ("*a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent*")[114].

Tem-se como admitido que os animais em geral possuem *qualia* e são seres sencientes[115]; e mesmo possuem uma moralidade própria, mas somente os seres humanos possuem personalidade (*sapiência*) e uma moralidade superior.

O que determina a existência de personalidade, em sentido filosófico?

A determinação do conceito de "pessoa" é uma das grandes revoluções em filosofia e foi objeto de elaboração precisa e refinada por *São Severino Boécio* (480-524). A roda da fortuna da rica e trágica vida, desse importante filósofo, transitou da glória ao cadafalso. Legou-nos importantes obras ao ponto de ser nominado o último romano e o primeiro escolástico.

São Boécio irá apresentar a sua definição de *pessoa* em duas obras, *Liber contra Eutychen et Nestorium* e *De Consolatione Philosophiae*. Para o autor, esse conceito está intimamente ligado à noção de individualidade. Trata-se de uma mudança conceitual revolucionária[116]. A tradição grega reconhecia o indivíduo, mas a tônica de sua definição estava vinculada a conceitos universais, tais como pólis, sociedade ou espécie humana. O *ethos* e o agir ético estavam profundamente vinculados aos deveres da vida em sociedade. A educação grega (*paideia*) era voltada para formação de um cidadão virtuoso (*arete*, *excelência*) em harmonia com a vida na pólis[117].

---

[113] BOSTROM, Nick; YUDKOWSKY, Eliezer. The ethics of artificial intelligence. *In*: FRANKISH, Keith; RAMSEY, William M. (ed.). *The Cambridge Handbook of artificial intelligence*. Cambridge: Cambridge University Press, 2014. p. 316-334. Disponível em: https://intelligence.org/files/EthicsofAI.pdf. Acesso em: 06 jun. 2020 às 22:54, na p. 6 (versão com pequenas alterações em relação à versão impressa).

[114] BOSTROM; YUDKOWSKY, 2014. p. 316-334.

[115] SINGER, Peter. *Vida ética*: os melhores ensaios do mais polêmico filósofo da atualidade. Rio de Janeiro: Ediouro, 2002.

[116] RODRIGUES, Ricardo Antonio. Severino Boécio e a invenção filosófica da dignidade Humana. *Seara Filosófica*, n. 5, 2012, p. 3-20, p. 3.

[117] JAEGER, Werner. *Paideia*: The ideals of greek culture. Trans. Gilbert Highet. Oxford University Press, 1945. v. I-III.

O autor ressignifica o conceito de um modo radicalmente distinto daquele utilizado na antiguidade, consolidando as bases conceituais que formarão a definição de pessoa na escolástica e na Idade Moderna.

O debate era sobre a *Santíssima Trindade* e *São Boécio* transitava teologicamente de modo perigoso sob a regência de um monarca adepto do arianismo, ou seja, um herege pela ortodoxia cristã. A tradição católica e cristã havia recepcionado o conceito de pessoa com acento na sua condição de pertença a uma *comunitas*. A individualidade e a racionalidade assentavam-se na sua relação com a coletividade. E o indivíduo somente se realizaria no seu vínculo estreito e necessário com a comunidade[118].

Para o autor, a definição de pessoa estava no cerne de um acirrado debate teológico, contudo, seria errôneo limitar as suas formulações a esse aspecto, as quais possuem profunda importância filosófica. Seu objetivo era rebater as formulações equivocadas de *Nestório* sobre a natureza de Cristo. Para este, Cristo possuía uma dúplice natureza (divina e humana) e, portanto, possuiria uma dúplice pessoa. O resultado dessa formulação seria desastroso para a Teologia e dissiparia ou obscureceria o mistério do Cristianismo[119].

São Boécio será o primeiro filósofo a acentuar o aspecto substancial da singularidade individual[120]: "as essências certamente podem ser nos universais, mas é apenas nos indivíduos e nos particulares que elas têm substâncias" (*essentiae in universalibus quidem esse possunt, in solis vero indiviuduis et particuluaribus substant*)[121]. Inicialmente, distingue as substâncias universais das particulares. As substâncias universais (homem, animal, pedra ou artefato) são as que se predicam nos particulares. Homem se predica de cada homem; pedra de cada pedra e assim por diante. Os particulares não se predicam de outros, tal como Platão ou esta pedra que fez esta estátua[122]. Para o autor, a intelecção dos universais é tomada dos particulares. Por sua vez, os particulares não são uma coleção de características particulares ou acidentes, como pensava *Porfírio*. A individualidade é substancializada, ou seja, a sua natureza decorre de sua essência[123].

O ponto mais relevante da argumentação de *São Boécio* está ao afirmar que pessoa é predicação de substância e, considerando que estas podem ser universais ou particulares, discorre que a pessoa somente possa ser uma predicação particular, singular e individual. A

---

[118] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 4.
[119] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 7.
[120] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 5.
[121] SAVIAN FILHO *apud* BOÉCIO, 2005, p. 78-79.
[122] CULLETON, Alfredo S. O conceito de pessoa em Ricardo de São Vitor. *Problemata – Rev. Int. de Filosofia*, v. 2, n. 1, 2011, p. 11-26, p. 13.
[123] CULLETON, 2011, p. 14.

pessoa nem é uma coleção de acidentes, nem é uma predicação universal, mas uma substância individualizada[124]. A seguir versa sobre a distinção entre a natureza e a pessoa, quando afirma que a pessoa não se iguala à natureza, contudo não se pode predicar *pessoa* para além da natureza.

Somente os seres portadores de alma racional poderiam ser classificados como pessoas (Deus, anjos e homem)[125], ou seja, a racionalidade é uma condição não acidental para o conceito de pessoa. Assim, o conceito de pessoa não pode ser atribuído aos seres *inanimados* (pedras), sem *sentidos* (vegetais) ou sem *racionalidade* (animais). Disso tudo decorre que, se há pessoa tão somente nas substâncias, e naquelas racionais, e se toda substância é uma natureza, mas não consta nos universais, e, sim, nos indivíduos, a definição que se obtém de pessoa é a seguinte: "substância individual de natureza racional"[126].

A definição de pessoa em *São Boécio* é caracterizado como individualidade de natureza racional (*persona proprie dicitur naturae rationalis individua substantia*). Essa formulação irá orientar todo o debate escolástico e estará nas bases do surgimento do moderno conceito de indivíduo, direitos humanos, dignidade da pessoa humana[127] e tantos outros correlatos. Sua importância será impar para a história do conceito de pessoa.

A ideia de racionalidade como condição essencial para as almas racionais é trabalhada na obra *De Consolatione Philosophiae*. Esta talvez seja uma das obras mais dramáticas da tradição filosófica. *Boécio* se encontra preso, sozinho em uma sela, à espera de sua cruel execução e *recebe* a Filosofia. Aquele que outrora fora rico, poderoso, com funções de governo, culto e refinado, encontraria a morte destituído de tudo[128].

Nesse momento fatal, destituído de tudo, reflete sobre a condição humana e procura elementos para definir o seu sentido. O que faz com que o mal muitas vezes subjugue os bons? Estaria Deus partilhando desses atos? Tal não seria possível, dado que Deus é o Sumo Bem[129].

O mundo possui uma ordenação e não parece ser fruto do caos, logo deveria ser racional. Haveria um erro na distribuição divina de castigos e prêmios, com a fortuna entregando erroneamente aos maus o que pertence aos bons? Tal não seria possível, dado que o Uno é

---

[124] CULLETON, 2011, p. 14.
[125] Cf.: *"dizemos que há uma pessoa do homem, de Deus, do anjo. Por sua vez, das substâncias algumas são universais, outras particulares".* Ver SAVIAN FILHO, Juvenal, Rodrigues, Ricardo. Severino Boécio e a invenção filosófica da dignidade humana. *Revista Seara Filosófica*, n. 5, 2012, p. 163. Disponível em: https://periodicos.ufpel.edu.br/ojs2/index.php/searafilosofica/article/view/1915. Acesso em: 15 dez. 20 às 00:36.
[126] *"Quocirca si persona solis substantiis est atque in his rationabilibus, substantiaque omnis natura est necin universalibus sed in indiuiduis constat, reperta personae est definitio:* '*naturae rationabilis indiuidua substantia*'". BOÉCIO, 2005, p. 282.
[127] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 12.
[128] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 4.
[129] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 12.

racional. Se Deus rege o mundo, é onisciente e possui inclusive o conhecimento dos fatos contingentes futuros, por que permite o mal? Esta seria a prova da liberdade humana. O livre-arbítrio é que conduz às escolhas, boas ou ruins. Logo, o mal não possui um estatuto ontológico, ele é contingente e deriva da errada deliberação dos homens, que escolhem o vício e o pecado[130].

Os homens possuem o livre-arbítrio e Deus não interfere nessas escolhas. A dignidade humana deriva não apenas do fato de ser um indivíduo racional, mas que faz deliberações sobre o seu agir, escolhendo o bem ou o mal. O indivíduo, ao fazer o bem, se aproxima da divindade, se eleva e é recompensado[131]. Se santifica ao agir virtuosamente. Aquele que comete o mal decai ao nível dos animais e se afasta do sagrado. *São Boécio* religa a condição humana ao seu sentido. A razão da condição humana não estaria em fatos efêmeros, tais como a riqueza ou o prestígio.

Para o autor, a dignidade humana não se fundamenta apenas na dimensão racional e na autoconsciência, mas por escolher agir eticamente, em direção ao bem e à santidade (participar da divindade)[132]. *São Boécio* acrescenta, na obra *De Consolatione Philosophiae*, a noção da ética como essencial ao conceito de pessoa. O agir para o bem é tão relevante quanto a racionalidade, para essa definição. A definição de pessoa manteria na escolástica os seus aspectos principais, tais como determinados na fórmula proposta por *São Boécio*, com ligeiras distinções. O conteúdo essencial seria mantido: *a singularidade racional do indivíduo*.

Para *Ricardo de São Vitor* (séc. XII), na sua obra *De trinitate*[133], *pessoa* é a existência incomunicável da natureza divina (*persona est divinae naturae incommunicabilis existentia*). Cada pessoa é uma realidade ontológica única, incomunicável, fechada, singular e racional. Em uma redação mais completa, afirma que pessoa é: "um ser existente por si mesmo como singular modo de existência racional" (*persona sit existens per se solum juxta singularem quemdam retionalis existentie modum*)[134].

A proposta de definição de *Ricardo* decorre de seu questionamento à amplitude da formulação de *São Boécio*. Esta seria demasiado imprecisa, na medida em que uma definição

---

[130] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 17.
[131] Cf. BOÉCIO *apud* RODRIGUES. "*Mas havíamos demonstrado que a felicidade é o próprio bem, o objeto de cada um de nossos atos. Portanto, é simplesmente o bem que é proposto como recompensa a todas as ações humanas. Ora, o bem não pode ser separados das pessoas boas, e não se poderia chamar de bom aquele a quem falta o bem; é dessa forma que as recompensas não negligenciam um bom comportamento [...] eis [...] a recompensa dos bons: [...] eles se tornam deuses como partícipes da divindade*". Ver RODRIGUES, Ricardo Antonio, *op. cit.*, p. 16.
[132] RODRIGUES, Ricardo Antonio, *op. cit.*, p. 17.
[133] SAINT-VICTOR. RICHARD de. *La Trinité*. Edição bilingue Latim-Francés. Introdução, tradução e notas de Gaston Salet SJ. Paris: Les Editions du CERF, 1959.
[134] CULLETON, 2011, p. 11-26.

deveria delimitar completamente o termo a ser definido. Na definição de *Boécio*, poder-se-ia deduzir, afirma *Ricardo*, que Deus seja uma substância individual, o que não teria sentido.

Por outro lado, a pessoa possui características únicas que não podem ser partilhada outras substâncias. O fato de ser uma *rationalis substantia* não distingue um indivíduo de outro. Deve existir uma qualificação exclusiva de cada pessoa, não partilhável, por isso incomunicável. O nome próprio da pessoa representa essa natureza autorreferencial da particularidade de cada um[135]. Todo indivíduo é substância racional, contudo, o que o caracteriza é propriedade singular (*proprietas singularis*), que designa diretamente o significado individual de cada um.

Para o autor, o conceito de *existência* (*existentia*) será fundamental para determinar a individualidade. De um lado, teremos a existência comum e, de outro, a existência incomunicável. Esta última seria propriedade atribuída somente a um indivíduo singular e não compartilhável, por isso mesmo, incomunicável. A propriedade pessoal é alguém ser absolutamente diferente (*discretus*) de todos os demais. As propriedades individuais não são acidentes, mas constitutivos da individualidade (*personalem proprietatem*)[136].

Pessoa não será mais, como bem sustenta *Culleton*, pensada como algo (*aliquid*), mas como alguém (*aliquis*). Não se fundamenta mais a sua diferença específica na racionalidade, mas no que é único, singular, incomunicável e discreto para cada pessoa. O distintivo de cada pessoa é a sua incomunicável existência (*incommunicabilis existentia*).

A partir desse momento, estariam construídos os pilares para a compreensão da pessoa como *sujeito*, singular, individual e incomunicável. A ética passa de uma afirmação de relações universais para uma afirmação de ações individuais e concretas, tomadas por sujeitos singulares, imersos em sua existência particular e única. A centralidade ética do sujeito passa a se constituir em um elemento fundamental da ética.

Caberia a *São Tomás de Aquino* (1225-1274) acrescentar o último elemento fundamental no conceito de pessoa: *a liberdade*. O autor retoma o conceito de *São Boécio* e o fundamenta, acrescentando o seu conceito particular: “por isso, alguns definem pessoa dizendo que é uma hipóstase distinta por uma qualidade própria à dignidade” (*persona est hypostasis proprietate distincta ad dignitatem pertinente*)[137].

---

[135] CULLETON, 2011, p. 18.
[136] CULLETON, 2011, p. 19-20.
[137] AQUINO, Tomás de. *Suma Teológica*, 1, 29, 1. 2017. Disponível em: https://sumateologica.files.wordpress.com/2017/04/suma-teolc3b3gica.pdf. Acesso em: 08 jun. 20220, às 01h02.

Ser pessoa é o mesmo que *hipóstase* (substância individual) de natureza racional[138].

A seguir acrescenta o *Doctor Angelicus* que a pessoa se caracteriza como um ente especial, dotado de racionalidade do poder de dirigir-se a si mesmo, de tomar escolhas e de não submeter-se às forças externas: "O particular e o indivíduo realizam-se de maneira ainda mais especial e perfeita nas substâncias racionais que têm o domínio de seus atos e não são apenas movidas na ação como as outras, mas agem por si mesmas. Ora, as ações estão nos singulares"[139].

A pessoa é a singularidade, racional, consciente, livre e que age em sentido ético e, portanto, é dotada de dignidade em si (*ad dignitatem pertinente*)[140].

A partir desse momento, estava consagrada a centralidade ética do conceito de sujeito, sendo aprofundada, criticada, aperfeiçoada por diversos filósofos, que irão destacar a importância do sujeito como questão filosófica central. Há uma passagem da filosofia que estuda o *ser enquanto ser* e as preocupações metafísicas para ter como objeto o *sujeito cognoscente.*[141]

Nos séculos seguintes teremos dois movimentos teóricos relevantes, em relação aos fundamentos da autoridade. A primeira mudança será da razão para a vontade. No final da Idade Média, ter-se-á a passagem fonte da autoridade da razão divina (*divine reason*) para a vontade divina (*divine will*)[142].

Duns Scotus (1266-1308) foi um dos teóricos desse entendimento e realizou essa importante passagem, que teria consequências muito importantes no entendimento da relação entre a vontade divina e a humana. A nossa razão fundamentar-se-ia na razão divina, em que a racionalidade humana participa em uma fração desta. De outro lado, a nossa vontade não é compartilhada diretamente pela vontade humana. A vontade será sempre individual e particular[143].

---

[138] AQUINO, Tomás de. *Suma Teológica*, 1, 29, 2. 2017. Disponível em: https://sumateologica.files.wordpress.com/2017/04/suma-teolc3b3gica.pdf. Acesso em: 08 jun. 20220, às 01h04.

[139] AQUINO, Tomás de. *Suma Teológica*, 1, 29, 1. 2017. Disponível em: https://sumateologica.files.wordpress.com/2017/04/suma-teolc3b3gica.pdf. Acesso em: 08 jun. 20220, às 01h12.

[140] STREFLING, Sérgio Ricardo A realidade da pessoa humana em Tomás de Aquino. Porto Alegre: Edipucrs, 2016. Disponível em: https://editora.pucrs.br/anais/seminario-internacional-de-antropologia-teologica/assets/2016/20.pdf. Acesso em: 08 jun. 2020 às 01:21.

[141] ALMEIDA, Rogério Tabet de. Evolução histórica do conceito de pessoa – enquanto categoria Ontológica. *Revista Interdisciplinar de Direito*, [*S.l.*], v. 10, n. 1, p. 229, out. 2017. ISSN 2447-4290. Disponível em: http://revistas.faa.edu.br/index.php/FDV/article/view/202. Acesso em: 08 jun. 2020.

[142] ALMEIDA, 2017, p. 229.

[143] ZAGZEBSKI, Linda. Intellectual Autonomy. *Philosophical Issues*, v. 23, p. 244-261, 2013. Disponível em: http://www.investigacoesfilosoficas.com/wp-content/uploads/04-Zagzebski-2013-Intellectual-Autonomy.pdf.

Para *Scotus*, Deus é, e somente Ele pode ser, perfeito. Assim, a alma também é perfeita ao participar do sagrado. Esta por sua vez se divide em três perfeições: a memória, a intelecção e a vontade. Segundo o autor: "também a vontade é dita 'perfeita', enquanto é sob aquele ato de querer perfeito"[144]. Assentava-se, assim, a vontade em bases sólidas, ao lado da razão, como propriedade distintiva da pessoa. A subjetividade teria por pilares a razão, mas também a vontade.

Outra contribuição relevante de *Scotus* está no seu entendimento sobre a importância da relação, para o conceito de *pessoa*. Apesar de a pessoa possuir uma existência incomunicável, ela não é isolada. Para tanto há uma distinção importante realizada entre a noção de indivíduo e a de pessoa. O primeiro é uma *realidade ontológica*, responde ao questionamento "o que é isto?" e se dirige à essência, que será sempre comunicável, entre todos os indivíduos. A pessoa é uma *realidade ôntica*, responde ao questionamento "quem é?" e se dirige ao plano da *existência*, incomunicável na singularidade de cada um[145].

O sujeito é incomunicável e singular, mas existe e compartilha de uma essência e se relaciona por essência com outros seres. É sobre esse agir prático que iremos nos deter em seguida.

O *eu* de Descartes será somente um de tantos outros filósofos a proceder a essa virada filosófica, de tal modo que uma lista completa tornar-se-ia sempre lacunosa. Se o sujeito passa a ser o objeto de estudo da filosofia, a autonomia passa a ser o objeto da moralidade.

1.2.4 Da autonomia como conceito central da moralidade

Immanuel Kant (1724-1804)[146] representa uma revolução copernicana nos estudos de filosofia moral. Neste autor teremos a passagem da autoridade natural da razão (*natural authority of reason*) para a ideia da autoridade natural do sujeito (*idea of the authority of the self*)[147]. A construção de *Kant* pretende demonstrar que ser governado por si mesmo e ser governado pela razão é o mesmo, dado que o sujeito é alguém governado pela vontade racional.

O autor inicia a "Fundamentação da Metafísica dos Costumes" com uma afirmação forte de que "nada é possível pensar que possa ser considerado como bom sem limitação a não ser

---

[144] *O conhecimento segundo João Duns Escoto*. 21§ 7 – Ordinatio I, d. 8, p. 1, q. Disponível em: http://docplayer.com.br/47246257-O-conhecimento-segundo-joao-duns-escoto.html. Acesso em: 11 jun. 2020 às 07:33.

[145] ALMEIDA, Juliano Ribeiro. Pessoa e relação em João Duns Scotus. *Enrahonar, Supplement Issue*, p. 79-87, p. 84, 2018. Disponível em: https://ddd.uab.cat/pub/enrahonar/enrahonar_a2018nsupissue/enrahonar_a2018nSupplp79.pdf. Acesso em: 11 jun. 2020 às 11:04.

[146] HILL, Thomas E. *The Blackwell guide to Kant's ethics*. Chichester: Wiley-Blackwell, 2009.

[147] ZAGZEBSKI, 2013, p. 244-261.

uma só coisa: uma boa vontade"[148]. O conceito de boa vontade não é totalmente formulado, sendo decorrente do bom senso. Seu valor independe de qualquer utilidade, bastando-se no seu querer. Parece, em determinadas passagens, identificar a "boa vontade" com a "simples vontade", ou seja, como a "vontade boa em si mesma"[149].

Caberá à razão, o destino de fundação de uma boa vontade. Esta será o bem supremo e a condição e meio de outras intenções, ou seja, da própria moralidade. O conceito de boa vontade estará no núcleo do conceito de dever. A ação por dever não pode, contudo, ser simplesmente boa, ela deve ser incondicionalmente boa. Para tanto o querer deve estar orientado pela razão legisladora[150].

Immanuel Kant critica as teorias morais de seu tempo e propõe responder à questão "como devo agir?" com fundamento na racionalidade. Como encontrar um fim último que não derivasse de bases frágeis, casuísticas ou contingentes. Para tanto a sua proposta é radical, deve-se pensar a moralidade longe das determinações empíricas, com base em regras racionais e universais. *Kant* irá encontrar o seu ponto de partida sólido na "vontade individual sujeita à razão". Mas não se trata apenas da vontade racional, e sim da vontade livre de forças externas, salvo o contrário, a vontade não seria autônoma, porém sujeita à heteronomia.

A vontade, para *Kant*, é "faculdade de desejar"[151], ou seja, o elo que permite que o mero livre-arbítrio se mova para a ação, na observação de *Thadeu Weber*. O autor, ao comentar *Kant*, destaca que a vontade somente será pura se afastada de toda heteronomia, de toda coação, de todas as instâncias mediadoras, tais como as motivações empíricas. A faculdade de desejar será pura na medida em que faz as suas escolhas com base nos princípios, *a priori*, dados pela razão[152].

O contingente, o empírico e o particular não podem se constituir em fundamentos seguros para a ação, somente a razão, que é fonte do conhecimento, poderá ser diretriz para a vontade, sob a forma de razão prática[153].

O arbítrio movido por necessidades básicas, instintos primitivos ou por coação se reduz a desejos menores, indignos ao ser racional livre. Este seria próprio dos animais, movidos por

---

[148] DEJEANNE, Solange de Moraes. *A fundamentação da moral no limite da razão em Kant*. 2020. Tese (Doutorado em Filosofia) – Programa de Pós-Graduação em Filosofia. PUCRS, Porto Alegre, 2020, p. 53. Disponível em: https://doi.org/http://tede2.pucrs.br/tede2/handle/tede/2783. Acesso em: 15 dez. 20 às 00:42.
[149] DEJEANNE, 2008, p. 54.
[150] DEJEANNE, 2008, p. 54.
[151] KANT, I. *A metafísica dos costumes*. Petrópolis: Vozes, 2013; KANT, I. *Die Metaphysik der Sitten*. Frankfurt am Main: Suhrkamp, 1982, p. 317.
[152] WEBER, Thadeu. *Revista de Estudos Constitucionais, Hermenêutica e Teoria do Direito* (RECHTD), v. 5, n. 1, p. 38-47, jan.-jun. 2013, p. 38.
[153] WEBER, Thadeu. Autonomia e dignidade da pessoa humana em Kant. *Revista Brasileira de Direitos Fundamentais & Justiça*, v. 3, n. 9, p. 232-259, 2009, p. 233.

sensações e instintos. Abdicar da vontade pura ungida pela liberdade seria o mesmo que abdicar da própria dignidade de ser racional, seria abdicar da humanidade naquilo que possui de mais valoroso: a liberdade de escolher sem coação. Sair da "menoridade" significa aceitar o entendimento como senhor da própria ação[154].

A escolha moral, a capacidade da vontade de determinar uma vontade pura para si mesmo é o ápice da razão prática. Quando o ser humano se eleva em relação a bestas-feras e se coloca em posição de digna superioridade. Não se nega que a vontade possa ser afetada pelos estímulos, assim como não existe ser humano fora da natureza, mas ele não pode ter a sua vontade "determinada" por eles[155]. Não haveria vontade pura se ela fosse determinada por um terceiro ou pela necessidade. Tal situação implicaria uma contradição material.

A vontade pura, ou simples vontade, não pode estar alicerçada em qualquer critério empírico, tal como o de utilidade. Pensar diferentemente seria chamar de vontade uma mera "fantasmagoria" travestida de direção racional da vontade ou de pura vontade[156].

A vontade pura constitui uma pessoa em sujeito, a partir do momento em que as suas ações lhe podem ser imputadas. Assim, a pessoa moral é aquela que, livre e racionalmente, submete-se às leis morais que atribui a si própria[157]. A vontade, ao contrário, seria heterônoma, quando controlada por outra, de fora dela[158].

A vontade pode ser heterônoma em outro sentido, segundo *Kant*, por razões internas. Ao ser controlada por inclinações ou caprichos, o intelecto e a vontade cedem aos impulsos menores. Essa tensão e dicotomia entre animalidade e personalidade está presente na essência da moralidade kantiana[159]. Ele será o grande artífice da autonomia como conceito central da moralidade. Será categórico ao afirmar na "Fundamentação da Metafísica dos Costumes" (1785) que: "autonomia é o fundamento da dignidade da natureza humana e de toda a natureza racional"[160].

A abertura de sua obra já designa o seu objetivo manifesto, nada menos do que "a pesquisa e a determinação do princípio supremo da moralidade, o bastante para constituir um todo completo, separado e distinto de qualquer outra investigação moral" (grifos nossos)[161]. Esse princípio supremo somente pode ser derivado da razão pura, longe de toda a consideração

---

[154] WEBER, 2009, p. 232-259,
[155] WEBER, 2013, p. 39.
[156] KANT, 2013.
[157] KANT, 2013, p. 66.
[158] ZAGZEBSKI, 2013, p. 247.
[159] ZAGZEBSKI, 2013, p. 248.
[160] "Duas coisas me enchem a alma de crescente admiração e respeito, quanto mais intensa e frequentemente o pensamento delas se ocupa: O céu estrelado sobre mim e a lei moral dentro de mim". Immanuel Kant.
[161] WEBER, Thadeu. *Ética e Filosofia política*: Hegel e o formalismo kantiano. Porto Alegre: Edipucrs, 2009.

empírica, ou derivado da experiência. A razão para tanto decorrer que ele deve ser estender a todos os seres racionais, de modo universal e não depender da experiência contingente. Somente tal preceito permitirá o princípio no qual a humanidade e toda a natureza racional são fins em si mesmos e não meios.

Kant afirma que, no reino dos fins, tudo tem um *preço* ou uma *dignidade*. O que possui preço pode ser trocado e substituído por outra coisa intercambiável. O que possui dignidade está acima de todo preço. Tudo o que possui preço possui um valor externo e relativo, enquanto a dignidade possui um valor intrínseco[162]. Somente a moralidade pode fazer que um ser racional seja um fim em si mesmo. O trabalho e a habilidade possuem um preço de mercadoria, a imaginação, um preço de sentimento, mas a humanidade possui um valor em si mesma. Para Kant a "moralidade é, pois, a relação das ações com a autonomia da vontade, isto é, com a legislação universal que as máximas da vontade devem tornar possível"[163].

O princípio supremo da moralidade será a vontade, ou seja, a propriedade do ser racional em determinar para si mesmo a lei de seu agir, independentemente de toda a força externa e de considerações contingentes.

> A autonomia da vontade como princípio supremo da moralidade A autonomia da vontade é a propriedade que a vontade possui de ser lei para si mesma (independentemente da natureza dos objetos do querer). O princípio da autonomia é pois: escolher sempre de modo tal que as máximas de nossa escolha estejam compreendidas, ao mesmo tempo, como leis universais, no ato de querer.

Para *Kant* a *vontade* é a causalidade dos seres racionais e a *liberdade* é a propriedade de agir independentemente das forças externas. Assim a moralidade é a própria relação das ações com a autonomia da vontade. Como essa proposição é independente da experiência, denomina *Kant* que esta é uma proposição sintética, *a priori*, indemonstrável pela experiência empírica, livre de todo o contingente, mas obviamente se coaduna a leis imutáveis[164]. A liberdade é a "propriedade da vontade de todos os seres racionais"[165], ou seja, somente eles podem agir. Os demais seres, se intui, agem por necessidade e vontade externa.

A definição de liberdade em *Kant* será inicialmente negativa, como ausência de determinação externa, subordinação aos dados empíricos, da necessidade, das determinações externas, do contingente, do particular e dos desejos decorrentes das sensações. Somente então

[162] WEBER, 2009.
[163] WEBER, 2009.
[164] WEBER, 2009, p. 235.
[165] WEBER, 2009.

a vontade pura será livre para se constituir positivamente como autodeterminação, como legislador moral, capaz de formular os imperativos de sua ação com base no esclarecimento[166].

O princípio da moralidade somente pode se basear na ação livre, racional e autônoma, ou seja, na capacidade da pessoa ser sujeito de seus atos, responsável por suas consequências e comprometida na busca de seus fins[167]. Uma sociedade de ser humanos livres, comandados exclusivamente pela vontade livre, se constitui em uma comunidade moral. Dado que a autodeterminação da vontade não pode se orientar pelo particular, então, a legislação deve ser universal. Essa é a máxima do terceiro imperativo categórico: "age de tal maneira que a vontade pela sua máxima se possa considerar a si mesma, ao mesmo tempo, como legisladora universal"[168].

Assim *Kant* deriva duas máximas: da vontade como legislador universal e da dignidade. A primeira determina que se deve agir de tal modo que nossa ação possa ser tornada como lei universal. O segundo imperativo categórico estabelece que o homem deve agir de modo a considerar o outro como tendo finalidade em si mesmo.

O projeto kantiano de moralidade será duramente criticado pelos autores posteriores. *Hegel* será implacável, denominando o projeto de *Kant* como sendo formalista. Contudo, ele irá reconhecer a importância do conceito de vontade, como fundamento da moralidade, mesmo a considerando insuficiente. Assim afirmava: "Tão essencial é acentuar a determinação pura da vontade por si, sem condição, como raiz do dever, como é, por conseguinte, verdade dizer que o reconhecimento da vontade teve de esperar pela filosofia kantiana para obter um sólido fundamento do ponto de partida (§ 133°)"[169]. O que faltava a essa vontade era um direcionamento para um fim definido. A vontade por si é um sólido ponto de partida, mas não esgota em si todos os elementos para a determinação do agir, que será sempre concreto e referenciado a fins.

Para *Kant* o reconhecimento da autonomia como conceito central na moralidade implica igualmente o entendimento de que devemos respeitar a autonomia dos outros, ou seja, autonomia e princípio do respeito à autonomia estão necessariamente vinculados[170]. Esse entendimento kantiano, da vinculação necessária, será objeto de críticas de autores como *Stuart Mill* e será objeto de análise a seguir, na seção sobre as teorias éticas em confronto[171].

---

[166] WEBER, 2009, p. 237.
[167] WEBER, 2009, p. 237.
[168] KANT, I. *A metafísica dos costumes.* Petrópolis: Vozes, 2013, p. 76.
[169] HEGEL, G. W. F. *Princípios da Filosofia do Direito*. São Paulo: Martins Fontes, 1997. p. 118.
[170] GILLON, Raanan. Autonomy and The Principle of Respect for Autonomy. *British Medical Journal. Clinical Research Edition*, v. 290, n. 6.484, p. 1.806-1.808, jun. 15, 1985, p. 1807.
[171] Ver Seção 1.3.

Para os fins do presente trabalho, cabe determinar que o conceito de *sujeito* será fundamental para a determinação da moralidade e, portanto, toda e qualquer fundamentação de uma ética da inteligência artificial ou de agentes morais artificiais deverá ter por base essa noção.

O reconhecimento da centralidade da autonomia para a moralidade consagrou-se com as formulações kantianas, contudo, diversas foram as críticas dirigidas às limitações dessa proposta. Ela foi acusada de formalismo, de confundir autonomia com o princípio de respeito à autonomia, dentre tantas outras insuficiências. A contemporânea teoria sobre o sujeito moral surge sobre as bases da autonomia em *Kant*, mas avançando em pontos importantes, não examinados pelo autor.

Hegel será um dos primeiros autores a atacar o formalismo kantiano no famoso § 133 da obra "Princípios da Filosofia do Direito" (*Grundlinien der Philosophie des Rechts*). Inicialmente, o autor irá firmar a essência da moralidade kantiana do "*dever pelo dever*", com a seguinte formulação[172]:

> Para com o sujeito particular, oferece o Bem a relação de constituir o essencial da sua vontade, que nele encontra uma pura e simples obrigação. Na medida em que a singularidade é diferente do bem e permanece na vontade subjetiva, o Bem apenas possui o caráter de essência abstrata universal do dever e, por força de tal determinação, o dever tem de ser cumprido pelo dever, (grifo nosso).

Destaca-se, na formulação kantiana de moralidade, a essência abstrata do dever, longe dos particularismos da experiência empírica. O indivíduo deveria estar liberto das condicionantes particulares, do contingente.

Thadeu Weber esclarece que a crítica hegeliana dirige-se ao formalismo kantiano[173]. Para o autor, ele esquece que toda a forma possui uma matéria e um contexto empírico relevante. A defesa kantiana está em seu propósito de fixar o princípio supremo de toda a moralidade livre das amarras de condicionalidades e coações externas à vontade livre. *Hegel*, por sua vez, destaca que inexiste forma sem conteúdo, principalmente quando estamos a falar da Ética. Afinal, como poder-se-ia indicar o caminho a seguir sem uma apreciação dos fins a serem perseguidos?

No importante parágrafo 134 dos "Fundamentos da Filosofia do Direito"*, Hegel* irá afirmar literalmente:

> Como a ação exige para si um conteúdo particular e um fim definido, e como a abstração nada de semelhante comporta, surge a questão: o que é o dever? Para

[172] HEGEL, 1997, p. 118.
[173] WEBER, Thadeu. *Hegel*: Liberdade, Estado e História. Porto Alegre: Vozes, 1993. HEGEL, 1997, p. 118.

> responder, apenas dispomos de dois princípios: agirmos em conformidade com o direito e preocupar-nos com o Bem-estar que é, simultaneamente, bem-estar individual e bem-estar na sua determinação universal, a utilidade de todos, (grifo nosso).

Hegel irá atacar diretamente o caráter abstrato do dever ao questionar o sentido abstrato do dever, afinal, o questionamento de como devemos nos comportar exige uma apreciação de um conteúdo particular e de um fim definido[174]. E irá afirmar:

> No entanto, estas duas determinações não estão implicadas na mesma determinação do dever; mas como ambas estão condicionadas e limitadas, são elas que conduzem à esfera superior da incondicionalidade do dever. E na medida em que o próprio dever constitui, como consciência de si, a essência e o universal desta esfera, essência que fechada em si, só a si se refere, apenas contém ele a universalidade abstrata; É identidade sem conteúdo ou positividade abstrata; define-se por ausência de determinação[175], (grifo nosso).

Hegel irá atacar o resultado desse dever abstrato, ou seja, uma identidade sem conteúdo ou uma positividade abstrata. O agir ético exige determinações ao dever, não podendo se constituir de modo indeterminado. Assim, ele próprio será condicionado e limitado. O postulado central do pensamento kantiano encontra-se em debate nesse ponto, tendo em vista que toda a formulação kantiana é a de que a moralidade seja incondicionada para que se constitua em fundamento superior da moralidade.

A base da moralidade kantiana é fundada sobre a *ideia da determinação pura da vontade por si*, sem condição, coação externa, determinação empírica. O resultado é um positivismo abstrato, ancorado na ideia de realização do dever pelo puro dever. Esse é o sólido ponto de partida da moralidade, escolhido por *Kant*[176]. Assim:

> Tão essencial é acentuar a determinação pura da vontade por si, sem condição, como raiz do dever, como é, por conseguinte, verdade dizer que o reconhecimento da vontade teve de esperar pela filosofia kantiana para obter um sólido fundamento do ponto de partida (§133.°); a afirmação do ponto de vista simplesmente moral que se não transforma em conceito de moralidade objetiva reduz aquele progresso a um vão formalismo e a ciência moral a uma retórica sobre o dever pelo dever (FD, §135, nota).
>
> Deste ponto de vista, não é possível nenhuma doutrina imanente do dever[177].

O resultado dessa decisão filosófica é uma construção repleta de dificuldades. Se o dever é definido de modo indeterminado, sem conteúdo ou condição, então a sua formulação é

---

[174] HEGEL, 1997, p. 119.
[175] HEGEL, 1997, p. 119.
[176] WEBER, Thadeu. A Eticidade Hegeliana. *Pucrs.br*, 1993, p. 08. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/veritas/article/view/35935/18874. Acesso em: 18 dez. 2020 às 21:29.
[177] WEBER, Thadeu. *Hegel, Liberdade, Estado e História*. Porto Alegre: Editora Vozes, 1993, p. 94.

meramente formal, como mera ausência de contradição. Decorre que ele não pode ser guia orientador para a indicação de deveres particulares.

> Poder-se-á decerto recorrer a uma matéria exterior e assim chegar a deveres particulares, mas desta definição do dever como ausência de contradição ou como acordo formal consigo – que não é mais do que a afirmação da indeterminação abstrata – não se pode passar à definição dos deveres **particulares**, e quando um conteúdo particular de comportamento chega a ser considerado, aquele princípio não oferece o critério para saber se se trata ou não de um dever. Pelo contrário, permite ele justificar todo o comportamento injusto ou imoral. A mais rigorosa fórmula kantiana, a da capacidade de uma ação ser representada como máxima universal, introduz decerto a representação mais concreta de uma situação de fato mas não tem para si nenhum princípio novo, outro que não seja aquela ausência de contradição e a identidade formal[178] (FD, §135).

A solução de Hegel é a de que os fins são o conteúdo da lei moral, trata-se do reino dos fins, onde se erige a *eticidade*. Para Kant os fins não podem ser o fundamento da moralidade, dado que a lei moral estaria ancorada em determinações, limites e condições que decretariam a morte da "ideia da determinação pura da vontade por si"[179].

Hegel irá aceitar a ideia revolucionária de *Kant* de autodeterminação da vontade como princípio da moralidade, mas irá considerá-la insuficiente e meramente subjetiva. Seria necessário passar da *moralidade* para a *eticidade*, sob pena de cair-se em uma moralidade formal e vazia[180].

O *formalismo* resultante da formulação kantiana decorre de duas propriedades imputadas ao imperativo categórico: *necessidade* e *universalidade*. Necessário é aquilo que não pode ser de outra maneira. Assim, a ação moral não pode ser de outro modo, nem determinada ou limitada por condições, tal como emoções. O dever moral é independente de propósitos, motivações ou tendências. Pensar em sentido diverso poderia acarretar o risco de relativismo. E a ação moral dependeria de condições, determinações e outras forças externas à vontade livre[181].

A solução hegeliana ao relativismo é dada pela mediação histórica. É na coerência universal decorrente do processo dialético da história que se mantém a coerência. *Kant* resolve esse problema recorrendo à ausência de contradição formal entre as máximas de conduta. *Hegel* irá solucionar por meio da superação dialética, de tal forma que supera conservando[182].

Observa muito bem *Klein* que:

---

[178] HEGEL, 1997, p. 119.
[179] WEBER, 1993.
[180] WEBER, 1993, p. 94.
[181] WEBER, Thadeu. Autonomia e dignidade da pessoa humana em Kant. *Direitos Fundamentais &Justiça*, n. 9, p. 234, out./dez. 2009.
[182] WEBER, 1993, *op. cit.* p. 11.

> a) O imperativo categórico não possui um conteúdo próprio, ele é apenas o princípio de não-contradição aplicado ao âmbito prático. Logo, não se pode derivar unicamente a partir dele um conceito determinado de dever.
>
> b) O imperativo categórico só funciona como um princípio de universalização quando já existe a suposição de um conteúdo externo, mas, nesse caso, dependendo do conteúdo que é abarcado, pode-se derivar inclusive imoralidades e ilegalidades[183].

A não contradição da lei moral entre uma máxima e a lei universal é uma das bases da universalidade. Para que a lei seja universal, ela não pode entrar em contradição com uma máxima. Esta, para poder ser convertida em lei universal, deve valer para todos, mas, se um sujeito particular desejar algo somente para si, cairá em contradição[184].

Cabe, contudo, esclarecer que *Kant não* utiliza propriamente a lógica formal para determinar as máximas de condutas. O autor irá usar uma lógica transcendental, ou seja, as regras que permitem o conhecimento teórico ou prático de como conhecer o mundo. A lógica transcendental verifica as condições de possibilidade do julgamento do valor moral de um dever. O formalismo transcendental é distinto do formalismo da lógica formal, que determina as regras de raciocínio[185].

A crítica que permanece é a de que a lei moral formal não determina uma condução para a ação, ou seja, não esclarece qual conduta a ser tomada. Dado que a conduta universalizável pode respeitar o princípio da universalização, haveria a possibilidade de a vontade eleger o mal como lei universalizável. A solução hegeliana é dada pela superação da universalidade abstrata pela universalidade concreta. A coerência ética será encontrada na comunidade ética histórica, superando o formalismo vazio[186].

Para *Furrow* há um problema fundamental na posição kantiana. A escolha da lei moral como uma diretriz para a ação decorre de uma preocupação anterior. Determino a obediência ao imperativo categórico porque entendo que ele tenha um papel fundamental para mim[187].

Poderíamos afirmar que essa é a escolha racional, contudo, o seu pressuposto é deveras ideal. Ele parte do pressuposto de que um agente racional ideal em condições ideais agiria de modo ideal, contudo, o agente racional real não age em tais condições. Por que deveria agir de modo irreal em condições diversas? Seria irracional tentar ter um comportamento ideal em um mundo real? Trata-se de um paradoxo que a posição kantiana é incapaz de superar[188].

---

[183] KLEIN, Joel Thiago Klein. As críticas de Hegel à teoria moral de Kant: um debate a partir do §135 de linhas fundamentais da Filosofia do Direito. *Dissertatio*, n. 34, 2011, p. 367-396.
[184] WEBER, 1993, *op. cit.* p. 09.
[185] KLEIN, 2011, *op. cit.* p. 371.
[186] WEBER, 1993, *op. cit.* p. 98.
[187] FURROW, Dwight. *Ética*: conceitos-chave em filosofia. Porto Alegre: Artmed, 2007. p. 34.
[188] FURROW, 2007, p. 34.

Outro ponto questionável é conceber o agente como um ser racional puro, sem nenhum influxo decorrente de suas crenças, desejos, emoções ou inclinações. Além de ser uma posição irreal, é algo forçado. O sujeito racional somente será coagido por suas emoções e crenças quando se deixar dominar por estas e não apenas por simplesmente admitir que ele as possui. Somente haverá a perda da autonomia quando ele for dominado por seus instintos[189]. Pelo contrário, as crenças e emoções fazem parte do que é propriamente humano, afinal este não é uma máquina lógico-racional. As crenças revelam os valores profundos[190] ou, melhor dizendo, o "eu profundo".

Parece existir um intenso conflito interno na posição kantiana entre objetividade e autonomia. A autonomia exige respeito à posição individual, já a objetividade determina um julgamento moral independente de todas as circunstâncias, características[191] e as condições empíricas. A solução kantiana exige que a autonomia seja necessariamente delimitada por requisitos externos de objetividade racional, ou seja, a autonomia individual é extremamente limitada por requisitos externos. A escolha da lei moral é a renúncia à escolha individual, no final das contas.

Um sujeito é autônomo moralmente quando seus desejos e seus valores possuem uma consistência prática no contexto moral em que está inserido[192]. Um posicionamento puramente formal é limitado, incapaz de explicar com profundidade as deliberações morais tomadas pelo indivíduo.

Outro problema na formulação kantiana é que o agente toma decisões em um contexto, no âmbito de suas relações e circunstâncias. Nada impede, contudo, que estudos posteriores demonstrem o vigor e a consistência de uma formulação kantiana da inteligência artificial, superando as limitações ora apresentadas.

Apesar de todos esses dilemas, o caminho em direção à construção da autonomia como conceito central na moralidade estava irreversivelmente pavimentado. Uma teoria sobre agentes morais artificiais deverá necessariamente entender o sujeito artificial como dotado de autonomia moral, com todas as características e os elementos previamente definidos, ou seja, alguém dotado de vontade própria, autorregulado e direcionado à realização de fins morais.

Outro problema, que deverá necessariamente ser levado em conta por uma teoria dos agentes morais artificiais, é o de que um sujeito ético, ao definir seus fins, irá se deparar com

[189] FURROW, 2007, p. 34.
[190] FURROW, 2007, p. 35.
[191] FURROW, 2007, p. 35.
[192] FURROW, 2007, p. 36.

diferentes teorias morais, que tentam justificar as suas escolhas. A pergunta "como devo agir?" irá receber diferentes soluções conforme a teoria ética de base.

Vejamos, no próximo capítulo, três teorias éticas*:* contratualismo, teoria das virtudes e utilitarismo.

1.2.5 Dos limites ao conceito de autonomia

Diversas são as críticas dirigidas ao conceito kantiano de autonomia como elemento central da moralidade[193]. Existem aquelas derivadas do argumento hegeliano, tal como exposto, bem como outras mais recentes, oriundas de matrizes teóricas tão distintas quanto a sociologia, o marxismo, o incompatibilismo e tantas outras[194].

Dois ataques foram particularmente duros. De uma lado, por *Anscombe* e, de outro, por *MacIntyre*.

Anscombe irá tecer vigorosas críticas tanto ao utilitarismo quanto ao deontologismo kantiano. A principal crítica a *Kant* irá se dirigir ao seu conceito de autonomia do sujeito vinculada à noção de legislador universal. Para a autora, essa noção não possui sentido e somente o teria se fosse preenchida por um legislador superior físico (moralidade normativa) ou metafísico (moralidade divina). O iluminismo kantiano refuta a fundamentação da ética em uma fundamentação teológica. Assim, a autora irá defender uma "uma análise positiva da justiça não como um princípio ético, mas como uma virtude"[195], no seu sentido aristotélico.

MacIntyre irá reforçar seu argumento aceitando as vantagens de se admitir o reconhecimento da autonomia individual para a teoria moral moderna. A capacidade de se fazer escolhas independentes é um elemento importante da moralidade. Contudo, alerta o autor, não se pode negligenciar as "virtudes do reconhecimento da dependência" (*virtues of acknowledged dependence*)[196].

Há em *Anscombe* e *MacIntyre* uma ressignificação do conceito de autonomia.

Para *MacIntyre* um dos grandes problemas na teoria moral em *Kant* está na sua recusa em fundamentar a moralidade na natureza humana, o que por si só implica uma espécie de

---

[193] FREY, Jennifer A. Against autonomy: why practical reason cannot be pure. *Manuscrito* 41, n. 4, p. 159–93, December 2018. Disponível em: https://doi.org/10.1590/0100-6045.2018.v41n4.jf. Acesso em: 08.11.2020 às 01:21.
[194] O'SHEA, Tom. *The Essex Autonomy Project*. Critics of Autonomy. University of Essex. Disponível em: https://autonomy.essex.ac.uk/wp-content/uploads/2016/11/CriticsofAutonomyGPRJune2012.pdf. Acesso em: 08.11.2020 às 01:21.
[195] ASCOMBE. Modern Moral Philosophy. *Philosophy*, v. 33, n. 124, p. 1-19, January 1958, p. 1-19.
[196] MACINTYRE, Alasdair. *Dependent Rational Animals*. Illinois: Carus Publishing, 1999, p. 9.

incoerência interna. *Kant* fundamentaria a moralidade na natureza racional do indivíduo, mas negará, por outro lado, a formulação de uma certa antropologia[197].

O autor afirma que a ausência de fundamentos profundos da moralidade implica que a moralidade se torna, sob o Iluminismo, em mero instrumento do desejo e da vontade individuais[198]. *MacIntyre*, apesar de reconhecer o papel e a importância das emoções na moralidade, fará uma demolidora crítica ao emotivismo na teoria ética. Não se pode admitir o resultado diverso, retirar a racionalidade da ação moral e a substituir por meros desejos individuais.

Uma teoria ética deve admitir a capacidade de realização de escolhas independentes, mas igualmente a conexão com a existência de emoções, virtudes e responsabilidade.

1.3 DA POSSIBILIDADE DE MODELOS MORAIS EM INTELIGÊNCIA ARTIFICIAL

Um sujeito moral, livre, racional e autorregulado deve possuir ou elencar fins para conduzir a sua ação. Assim, qual deveria ser o modelo moral a conduzir as decisões de um sujeito artificial?

Um sistema artificial pode ter um algoritmo moral implementado originariamente de modo externo, e assim não poderíamos dizer que ele realmente é livre ou autônomo, no sentido que determina a teoria ética desde *Boécio*, muito menos em sentido kantiano ou hegeliano. Por qualquer teoria anteriormente vista, poderíamos dizer que o sistema realmente escolhe ou ele é *marionete* orientado por um controlador externo.

Digamos que ele possua um código inicial e que este evolua conforme as condições e circunstâncias vividas: nesse caso poderíamos aceitar que há espaços crescentes de autonomia. O código inicial pode ser extremamente simples, ao ponto de evoluir praticamente do nada e de regras muito básicas. Não se pretende verificar o risco para a humanidade de um desenvolvimento de um sujeito moral sem tutela, o que pode ser deveras perigoso.

A engenharia de sistemas passou a estudar esses dois casos de implementação de um sistema artificial moral. Afinal, qual dos dois seria factível com a possibilidade de um autômato moral?

---

[197] FELDHAUS, Charles. De Schopenhauer a ética de virtudes contemporânea. *Revista Guairacá*, v. 29, n. 2, p. 46, 2013.
[198] MACINTYRE, Alasdair. *Depois da virtude*: Um estudo em teoria moral. Trad. Jussara Simões. Bauru: Edusc, 2001. p. 115.

1.3.1. Teorias morais e inteligência artificial

A tentativa de construir uma máquina capaz de formalizar e computar todas as possibilidades razoáveis de uma determinada ação em qualquer circunstância foi pensada inicialmente por *Hobbes* (1588-1679). Para ele as ações eram resultados de cálculos sobre as paixões humanas[199]. Para este autor, as sensações eram recebidas pelo corpo e passariam pela imaginação, que iria ponderar e calcular se a ação seria realizada ou não[200].

As paixões humanas seriam divididas em pares se que afastam ou aproximam do objeto, tal como o desejo e a aversão[201]. A partir da dinâmica desses pares (esperança e medo) é que surgiriam as instituições, tal como o Estado Civil. A deliberação seria justamente o resultado do cálculo, de aproximação ou afastamento do objeto, pelo desejo ou aversão, considerando os benefícios deste para a autopreservação[202]. Para *Hobbes*: "Por esta imposição de nomes, uns mais amplos, outros de significação mais restrita, transformamos o cálculo das consequências de coisas imaginadas no espírito num cálculo das consequências de apelações"[203].

Para o autor, a razão teria um papel fundamental em calcular[204], por meio da adição ou subtração, qual a ação a tomar[205]. Falar nada mais seria do que calcular por palavras[206].

A ideia de que o cérebro humano realiza identificação de possíveis condutas éticas, consegue formalizar as escolhas por meio de pesos, determinando benefícios ou prejuízos na tomada de decisão e ao final delibera, considerando os riscos e as consequências, produz a indagação da possibilidade de mimetizar tal comportamento de deliberação moral, reproduzi-lo, aperfeiçoá-lo ou torná-lo autônomo.

O sonho de uma calculadora moral se tornou progressivamente mais desafiadora com a primeira máquina de calcular de *Blaise Pascal*, na intrigante máquina de *Leibniz* e, finalmente, com o computador de *Babbage*.

---

[199] ARAÚJO, Luana Broni de. A filosofia natural de Thomas Hobbes: a composição das paixões humanas. *Controvérsia*, São Leopoldo, v. 14, n. 3, p. 75-96, set.-dez. 2018, p. 84.
[200] ARAÚJO, 2018, p. 89.
[201] ARAÚJO, 2018, p. 92.
[202] ARAÚJO, 2018, p. 91.
[203] Cf. "*By this imposition of names, some of larger, some of stricter signification, we turn the reckoning of the consequences of things imagined in the mind, into a reckoning of the consequences of appellations*"; ver *in* HOBBES, Thomas. *Leviathan*. Oxford World's Classics. Oxford: Oxford University Press,1996. p. 45.
[204] LEIVAS, Cláudio Roberto Cogo. *Representação e Vontade em Hobbes*. 2005. Tese (Doutorado em Filosofia) – Instituto de Filosofia e Ciências Humanas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005. p. 254.
[205] ARAÚJO, 2018, p. 95.
[206] HOBBES, 1996, p. 45.

Afinal, seria possível uma *máquina de calcular moral*? Capaz de auxiliar na tomada de decisões deliberativa, escaneando todas as infinitas possibilidades e indicando o melhor caminho a seguir, tal como um *Oráculo de Delfos*[207] tecnológico?

A perspectiva de quantificar e reproduzir o raciocínio moral, em cálculos sobre o valor máximo de bem-estar, semelhante ao utilizado pelo cérebro tem instigado os engenheiros a pensarem em soluções de *design* moral[208]. Existem dois modelos alternativos: de topo ou de piso.

a. Design de topo

O primeiro modelo se chama de *topo* (*cima para baixo* ou *top-down*), ou seja, quando a arquitetura do sistema se estrutura de cima para baixo. São estruturados os princípios gerais e as regras de cimo e, a partir desse modelo, se constrói toda o design do sistema moral artificial. Nesse caso se pretende verificar as regras básicas que permitam uma ética formalizada e computável.

O modelo de regras de alto nível exige a escolha de qual conjunto de regras deve ser adotado, dentre os conjuntos rivais, tal como, por exemplo, entre o *consequencialismo* e a *deontologia.*

A *deontologia* é bem apresentada pelo modelo kantiano de cumprimento do dever pelo dever, tal como expresso na imperativo categórico. O objetivo kantiano é buscar um fundamento último de moralidade. E a busca de um fundamento autônomo da razão, não distorcida ou coagida por elemento externo, deve ser a base da liberdade. Kant tenta responder à pergunta: "o que devo fazer?", por meio da busca do princípio supremo da moralidade[209].

Paton enumera cinco formulações (*formulae*) do imperativo categórico[210], três tidas como principais e duas, derivadas. As cinco fórmulas são as seguintes:

- da lei universal: "Age somente segundo uma máxima por meio da qual possas querer ao mesmo tempo que ela se torne lei universal";

[207] VOLKER, Camila Bylaardt. *As palavras do Oráculo de Delfos*: um estudo sobre o *De Phytiae Oraculis* de Plutarco. Disponível em: https://repositorio.ufmg.br/bitstream/1843/ECAP-6ZFG54/1/microsoft_word___camila_bylaardt_volker.pdf. Acesso em: 15 jun. 2020 às 23:07.
[208] VALLE, Juan Ignacio del. Inteligencia artificial ética: Un enfoque metaético a la moralidad de sistemas autónomos (TFG). Bruxelas: Universidad Nacional de Educación a Distancia, 2019. Disponível em: https://www.researchgate.net/publication/337797495_Inteligencia_Artificial_Etica_-_Un_Enfoque_Metaetico_a_la_Moralidad_de_Sistemas_Autonomos_TFG. Acesso em: 14 jun. 2020 às 20:05.
[209] WEBER, Thadeu. *Ética e Filosofia política*: Hegel e o formalismo kantiano. Porto Alegre: Edipucrs, 2009. p. 31.
[210] PATON, H. J. *The Categorical Imperative*: A Study in Kant's Moral Philosophy. Philadelphia: University of Pennsylvania Press, 1971. p. 129.

- da lei da natureza: “Age segundo a máxima que, mesmo contrária à tua vontade, possa ser tomada como lei da natureza";
- do homem como fim em si mesmo: "Aja de tal forma que uses a humanidade, tanto na tua pessoa, como na pessoa de qualquer outro, sempre e ao mesmo tempo como fim e nunca simplesmente como meio";
- da autonomia da vontade: “Aja de tal maneira que tua vontade possa encarar a si mesma, ao mesmo tempo, como um legislador universal através de suas máximas";
- do “reino dos fins”: “Age como se fosses, através de suas máximas, sempre um membro legislador no reino universal dos fins".

Para *Paton*, essas *formulae* possuem um encadeamento progressivo, de tal modo que *Kant* está preparando o argumento geral para a adoção das duas últimas fórmulas. Poder-se-ia afirmar que a formulação da autonomia da vontade e a do “reino dos fins” são as mais importantes[211].

Hegel irá apresentar uma crítica ao imperativo categórico kantiano no famoso §133 de sua obra “Princípios da Filosofia do Direito”. Inicialmente, o autor irá firmar a essência da moralidade kantiana do “dever pelo dever”, com a seguinte formulação:

> Para com o sujeito particular, oferece o Bem a relação de constituir o essencial da sua vontade, que nele encontra uma pura e simples obrigação. Na medida em que a singularidade é diferente do bem e permanece na vontade subjetiva, o Bem apenas possui o **caráter de essência abstrata universal do dever** e, por força de tal determinação, **o dever tem de ser cumprido pelo dever** (FD, §133), (grifo nosso).

Destaca-se, na formulação kantiana de moralidade, a essência abstrata do dever, longe dos particularismos da experiência empírica. Kant irá buscar sempre os princípios mais gerais e elevados da moralidade, aplicáveis a qualquer doutrina[212]. A dificuldade principal desse modelo é o salto lógico entre os grandes princípios abstratos da moralidade e a experiência concreta.

As críticas de *Anscombe* são ainda mais ácidas. Para a autora, a ideia de um autolegislador é um “absurdo”, dado o fato de que teríamos sempre um resultado majoritário predefinido (na forma 1x0), fruto da reflexão do sujeito moral. A legislação, pondera a autora, exige sempre um poder superior ou, diríamos nós, exterior[213]. Sua regra de universalização seria inútil sem descrições adequadas da conduta a ser tomada.

---

[211] PATON, 1971, p. 130.
[212] PATON, 1971, p. 131.
[213] ASCOMBE, 1958, p. 2.

Anscombe é reconhecida pelo surgimento da denominação "consequencialismo" para designar a teoria moral inaugurada por *Sidgwick,* que superava o utilitarismo clássico[214].

O *consequencialismo* é a teoria moral que leva em consideração as consequências de cada decisão. Tal formulação tão ampla é, contudo, vazia e permitiria o preenchimento de qualquer significado, inclusive os absurdos. Uma formulação mais adequada seria: "o princípio segundo o qual uma ação (regra, prática ou instituição) é moralmente correta ou está justificada se, dentre as possibilidades, ela apresentar o maior saldo líquido de consequências desejáveis sobre aquelas indesejáveis"[215].

Essa definição acarreta duas implicações[216]:

- trata-se de uma teoria moral que leva em conta o resultado de sua ação para a definição da deliberação acerca de qual ação moral o agente deve seguir;
- o resultado da ação do indivíduo se constitui como critério (mais) relevante para controle da correção da escolha moral.

Três elementos participam dessa definição: fins, meios e critérios. O fim pretendido será o bem-estar, ou seja, o que é considerado bom para alguém.

O conceito de bem-estar exige algumas delimitações, dado que pode ser individual ou geral. A ideia de bem-estar estava historicamente ligada à escolha correta do prazer. Já afirmava *Sócrates* no "Protágoras de Platão" que a "a salvação de nossa vida se revelou como consistindo na escolha acertada de prazeres e de sofrimentos, conforme sejam mais ou menos numerosos, maiores ou menores, ou se encontrem afastados ou mais perto [...]"[217].

Esse critério será retomado por *Jeremy Bentham* na obra "Uma introdução aos princípios da moral e da legislação" (1789) e sob a égide do princípio da utilidade. Segundo o autor a ação humana é governada pelo binômio prazer e dor[218], ou nas suas palavras: "natureza colocou a humanidade sob o comando de dois mestres soberanos, o prazer e a dor"[219].

---

[214] ASCOMBE,1958, p. 12.
[215] PICOLI, Rogério Antonio. Utilitarismos, Bentham e a história da tradição. *Existência e Arte*, v. 2, p. 1-20, 2010, p. 4.
[216] PICOLI, 2010, p. 4.
[217] PLATÃO, 2002, p. 155.
[218] PICOLI, 2010, p. 11.
[219] Cf. "*I. Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think: every effort we can make to throw off our subjection, will serve but to demonstrate and confirm* it". BENTHAM, Jeremy. *An Introduction to the Principles of Morals and Legislation (1781).* Batoche Books. Kitchener, 2000. p. 14. Disponível em: https://socialsciences.mcmaster.ca/econ/ugcm/3ll3/bentham/morals.pdf. Acesso em: 20 jun. 2020 às 15:21.

A teoria moral de *Bentham* irá eleger o princípio da busca do prazer como critério fundamental da escolha moral. As ações dirigidas à satisfação desse fim seriam justificadas e desejadas; o contrário seria afastado. O autor denominou essa concepção de *utilitarismo*.

Os limites da utilização do critério do prazer logo se tornaram explícitos. A principal análise do consequencialismo, reformulando as bases iniciais do utilitarismo, deve-se a *Henry Sidgwick* na obra "The Methods of Ethics". Essa é considera a mais importante obra sobre ética moderna, que irá balizar muitos autores posteriores[220]. O objetivo explícito do autor é determinar um procedimento racional para determinar o que deve ser (*ought to*) ou é correto (*right to*) fazer, em determinada ação voluntária[221].

Para *Sidgwick*, seria contraditório, para um agente racional, eleger determinado fim a ser perseguido e não adotar todos os esforços para atingi-lo. Seria ainda mais inconsistente adotar um fim e se recusar a persegui-lo[222]. Ao analisar o binômio prazer e desprazer, o autor revela que a volição determinada por prazeres e sofrimentos é uma forma de hedonismo psicológico. Essa seria uma visão de que as leis éticas são governadas por princípios psicológicos, território em que a razão não ousaria governar.

As teorias consequencialistas irão ordenar de diferentes modos a composição entre a busca do bem individual e o bem-estar das pessoas sujeitas a uma ação moral[223]. Seria o consequencialismo uma teoria egoística, em que cada agente ao escolher o melhor para si produz o melhor resultado para todos (*Mandeville*[224]) ou seria uma escolha benevolente, onde as escolhas orientadas pelo bem-comum acarretam um melhor resultado para o bem individual (*Sidgwick*)?

A busca dos interesses individuais para *Mandeville* seria boa em si mesma, não importa se virtuosas ou viciosas[225]. Inclusive, os vícios privados poderiam gerar benefícios públicos, dado que a busca do prazer próprio irá produzir resultados gerais positivos[226].

---

[220] ASCOMBE 1958, p. 9.
[221] SIDGWICK, Henry. *The Methods of Ethics*. 2011, p. 3. Disponível em: https://www.earlymoderntexts.com/assets/pdfs/sidgwick1874.pdf. Acesso em: 20 jun. 2020 às 16:09.
[222] SIDGWICK, 2011, p. 11.
[223] PICOLI, 2010, p. 5.
[224] MANDEVILLE, Bernard Mandeville. The fable of the bees or private vices. *Public Benefits*, v. 1 [1732]. The Online Library of Liberty. Disponível em: http://oll-resources.s3.amazonaws.com/titles/846/Mandeville_0014-01_EBk_v6.0.pdf. Acesso em: 20 jun. 2020 às 10:05.
[225] FONSECA, Eduardo Giannetti da. *A Fábula das Abelhas*. Braudel Papers. The Tinker Foundation & Champion Papel e Celulose, 1994.
[226] BRITO, Ari Ricardo Tank. *As abelhas egoístas*: vício e virtude na obra de Bernard Mandeville. 2006. Tese (doutorado em Filosofia) – Programa de Pós-Graduação em Filosofia. Universidade de São Paulo, São Paulo, 2006. p. 128. Disponível em: http://livros01.livrosgratis.com.br/cp077816.pdf. Acesso em: 20 jun. 2020 às 18:03.

A proposta de *Sidgwick* indicaria um caminho mais sofisticado e complexo na determinação do bem a ser buscado. Não somente o prazer imediato deveria ser levado em conta, afinal, prazeres imediatos podem redundar em desprazeres futuros, muito mais relevantes. Na consideração "global" dos desejos atuais e futuros, com as suas diversas consequências, conforme uma das múltiplas possibilidades de conduta, dariam a resultado presente o melhor caminho a seguir. Seria do resultado agregado da composição hipotética (*hypothetical composition*) das múltiplas forças impulsivas que surgiria uma reflexão adequada sobre a deliberação a ser tomada em certas condições[227].

Para *Sidgwick* bem-estar individual está necessariamente conectado com o aumento da felicidade alheia. Se a sociedade estiver bem ordenada, com instituições bem estruturadas, será possível alcançar o melhor resultado líquido para cada um dos seus integrantes[228].

O meio a ser utilizado no consequencialismo será a maximização dos resultados práticos. A busca do bom indica que o indivíduo deverá buscar maximizar as ações que o aproximem do bem-estar individual e o afastem das ações que produzam resultado inverso. Será uma exigência de racionalidade prática, a deliberação orientada para otimizar os resultados pretendidos.

O cálculo dos desejos já havia sido escolhido como método deliberativo. *Platão* na obra *Protágoras* afirmava:

> 356b. É como se um homem bom em pesagens, somando prazeres com prazeres e somando dores com dores, depois de ajustar na balança a proximidade e a distância, disser quais são as maiores; porque se pesares prazeres com prazeres terás que aceitá-los sempre com dores em menor número e em menor tamanho. Agora, se forem prazeres com dores, se os prazeres as excederem, seja a proximidade menor que a distância ou a distância menor que a proximidade, terás que agir segundo o que estes ditarem. Se forem as dores a exceder os prazeres, não terás que o fazer[229].

O resultado da balança entre prazeres e desprazeres, benefícios *versus* prejuízos, definirá o caminho a seguir, mediante uma deliberação clara. O cálculo exigirá alguns componentes importantes. Quais serão aos prazeres e as dores confrontados? Qual o peso de cada um, ou terão pesos idênticos? Os prazeres de mesma classe se compensam com dores de mesma classe ou entram todos em conferência geral? A listagem exaustiva desses bens será minunciosamente tratada por *Bentham*, mas questionada por diversos autores, dada a sua dificuldade óbvia.

---

[227] Cf. "*He characterizes a person's future good on the whole as what he would now desire and seek if the consequences of all the various courses of conduct open to him were, at the present point of time, accurately foreseen by him and adequately realized in imagination*". RAWLS, John. *Theory of Justice*. Cambridge: Harvard University Press, 1999. p. 366.
[228] RAWLS, 1999, p. 20.
[229] PLATÃO, 2002, p. 155.

O consequencialismo caminhava para uma direção genérica. O cálculo deve prever as consequências da conduta do agente moral, de tal modo que o resultado líquido das expectativas desejadas seja positivo. De maneira geral, tem se aceito que o fim a ser alcançado é a satisfação das preferências, que podem ser as mais diversas, tais como: a saúde, educação, entre outros[230].

Dado que o fim almejado é um bem ou consequência esperada e que o método será o cálculo estimado do resultado líquido dos desejos alcançados, qual seria o critério definidor da ação do agente moral? Somente pode ser aquela escolha que *maximize* o resultado líquido, em nível individual ou social, de modo imediato, mas também mediato. Ou seja, se, na formulação kantiana o indivíduo governava suas ações por meio da escolha racional da norma universalizadora, em sentido abstrato. De outro lado, encontraremos a proposta consequencialista no exato oposto. A melhor deliberação moral será aquela que racionalmente conseguir maximizar os desejos do agente moral, levando em conta diversos elementos individuais e sociais; bem como os imediatos e mediatos.

Um sistema de inteligência artificial deveria ter a capacidade de processar em tempo real toda a informação necessária, com todos os dados disponíveis sobre as melhores consequências para dada ação. Um procedimento desse tipo foi imaginado por *James Gips* que sugeriu um algoritmo capaz de *scannear* todos os fatos do mundo, relevantes para a decisão e com capacidade de predizer todas as consequências para cada conduta imaginada.

Obviamente a máquina não poderia manter o processamento desses dados ao infinito, sendo que, em determinado momento, deve-se parar a cadeia de cálculos e proceder a uma decisão moral[231].

O consequencialismo demonstrar-se-á um modelo marcado por diversas divisões internas. Qual seria o bem objetivado? Para quem? Como? No fim as questões se somam, sem expectativa de solução à vista.

A falta de unanimidade no tipo de teoria moral a ser adotada pelo agente moral artificial acarretou uma dificuldade intransponível na construção de um modelo ético de topo (*top-down*). Desse modo, os teóricos em inteligência artificial passaram a considerar a possibilidade de um modelo de baixo para cima (*bottom-up*), por nós denominado de modelo de piso.

b. Design de piso

[230] FURROW, 2007, p. 53.
[231] VALLE, 2019.

Outro modelo proposto é o de *piso*, ou seja, uma moralidade que se constrói de *baixo para cima* (*bottom-up*)[232]. Nesse caso o sistema passa a adquirir capacidades morais. Essas características fazem que tal modelo se aproxima da ética aristotélica das virtudes.

Enquanto a deontologia se preocupa com uma formulação mais geral de deveres e máximas e o consequencialismo com a definição do desejável, a ética das virtudes estabelece as disposições do carácter do agente moral como a questão central da moralidade[233]. Talvez a resposta para a elaboração de um sistema ético artificial esteja na filosofia clássica aristotélica[234].

O interesse pela ética das virtudes decorre do acelerado e acentuado grau de autonomia dos agentes artificiais[235]. Não somente modelos de robôs cada vez mais sofisticados se sucedem, com novas e mais surpreendentes capacidades reais e possíveis. Novos dispositivos dotados de autonomia e sistemas inteligentes embarcados se multiplicam em formas, tamanhos e funcionalidades. Cada um deles como aumento exponencial de possibilidades e recursos. São drones, veículos autônomos, máquinas de cuidados, casas inteligentes, fábricas automatizadas ou armas inteligentes. A forma cada vez mais próxima ao humano é utilizada para romper a barreira da aversão às novas máquinas, que sorrateiramente passam a fazer parte do dia a dia da humanidade.

A multiplicidade de agentes artificiais, a sua rápida evolução e integração na vida social humana e a dificuldade de criar um mecanismo capaz de processar todos os dilemas morais em cérebro eletrônico, obrigou a diversos engenheiros cogitarem uma estratégia diversa para esses agentes morais artificiais. Ao invés de procurar o modelo completo para a um agente moral, a pergunta seria reformulada para: "quem sabe eles não devem ser governados por si mesmos?"[236].

---

[232] WALLACH, Wendell; ALLEN, Colin. *Moral Machines*: Teaching Robots Right from Wrong. 2008. Disponível em https://www.researchgate.net/publication/257931380_Moral_Machines_Teaching_Robots_Right_From_Wrong. Acesso em: 16 dez. 2020 às 21:05.

[233] Não serão estudados modelos de ética artificial de "piso" semelhantes, tais como o confucionismo. Cf. "*Compared to most Western ethical approaches that focus on moral reasoning and justification, Confucian ethics places more emphasis on moral practice and practical wisdom. What is central to Confucian ethics is the moral development model that consists of three interrelated components: observation, reflection, and practice*". ZHU, Qin; WILLIAMS, Tom; WEN, Ruchen. *Confucian Robot Ethics*, 2019. Disponível em: https://www.researchgate.net/publication/339815118_Confucian_Robot_Ethics. Acesso em: 09 ago. 2020 às 04:20.

[234] BERBERICH, Nicolas; DIEPOLD, Klaus. *The Virtuous Machine*: Old Ethics for New Technology? Munich: Munich Center for Technology in Society, 2018. p. 3. Disponível em: https://arxiv.org/pdf/1806.10322.pdf. Acesso em: 21 jun. 2020 às 21:50.

[235] BERBERICH; DIEPOLD, 2018, p. 1.

[236] Cf. "*Due to the inherent autonomy of these systems, the ethical considerations have to be conducted by themselves*". BERBERICH; DIEPOLD, 2018.

Será possível que a ética da virtude pode se constituir em um guia moral promissor para os sistemas de inteligência artificial e para os agentes morais em particular[237]? A análise aprofundada da possibilidade filosófica dessa estratégia será o tema da seção sobre os Agentes Morais Artificiais (AMA).

1.3.2 Conflitos morais e consistência moral

A resolução de conflitos morais é uma das grandes dificuldades práticas na implementação de uma ética artificial. O estabelecimento de um conjunto restrito de normas (leis morais) aplicável a toda e qualquer decisão prática se provou impraticável[238]. O exemplo do debate acerca das três leis da robótica é sintomático. Estas não sobreviveram aos testes de consistência. Afinal, o que fazer quando não se pode impedir um dano a um ser humano? Ou quando um ser humano provocaria um dano a outrem e a única forma de impedir seria machucar algum deles? Poderia a máquina agir como tutora de humanos? Ou substituir as escolhas destes para evitar danos futuros? E se eles derem ordens contraditórias? As perguntas tornam-se cada vez mais desafiadoras para que o design original ou inicial de um sistema artificial possa responder de modo satisfatório.

A possibilidade de que leis morais gerais possam conduzir adequadamente a conduta de agentes artificiais se demonstrou reduzida, em face dos conflitos morais concretos a que os agentes estão sujeitos.

O tema não era de todo desconhecido na escolástica, pelo contrário, era objeto de vívido debate. Afinal, a tradição medieval girava sobre a capacidade e vontade dos agentes em realizarem escolhas virtuosas. Existiam duas estratégias para enfrentar dilemas morais insolúveis (*irresolvable moral dilemas*)[239]. Na primeira estratégia, os agentes esperavam por uma intervenção divina milagrosa, o que não era de modo algum tranquilamente admitido. A aceitação de que uma iluminação ou revelação resolveria o dilema era muito incongruente, como uma análise racional do dilema.

São Tomás de Aquino apresenta o caso de um padre que descobre que seu cálice está envenenado e rejeita a ideia de que deveria tomar o líquido na esperança de proteção divina.

---

[237] Cf. "*theory for building moral machines is a promising approach to avoid the uncanny valley and to induce acceptance*". BERBERICH; DIEPOLD, 2018, p. 3.
[238] VALLE, 2019.
[239] Não será utilizada a distinção entre as denominações *dilemas insolúveis* e de *dilemas genuínos*, apontada por BRINK, 1994, p. 218 e adotada por DI NAPOLI, Ricardo Bins. O intuicionismo moral e os dilemas morais. *Dissertatio*, UFPel, v. 35, p. 79-98, 2012, p. 82.

O uso de estratégias de *intervenção externa* ao entendimento do agente continuou seguindo a tradição filosófica, principalmente a secular, com uma criatividade digna de nota. *Leibniz* relata o uso de meios bizarros para solução extrajurídica de casos judiciais, tal como o uso de sorteios (*drawing lots*) ou testar a sorte pelo jogo de "cara e coroa", atirando moedas ao ar (*flipping coins*)[240]. A literatura ainda refere casos similarmente curiosos, como decidir por privilégio a uma parte, inventar uma solução fictícia, dividir o bem em dois ou simplesmente abdicar do poder de julgar[241].

Leibniz, ao se deparar com o problema, encontrará dificuldades relevantes em superar um conflito moral entre escolhas com mesma força ou peso. O autor irá utilizar uma analogia com a geometria para resolver o problema. Se cada curso de ação fosse comparável a objetos de mesmo peso e mesma velocidade e viesse a se chocar, o resultado seria um movimento perpendicular equidistante dos objetos originais, assim, a melhor resposta possível seria dividir o objeto ou a pretensão entre os dois pretendentes[242]. As impropriedades do método de *Leibniz* se deveram principalmente à impossibilidade de computar todas as possibilidades jurídicas em fórmulas simples[243].

Um outro caminho, muito mais importante e relevante, será considerar o desafio de solução de um conflito moral como essencialmente um dilema assentando sobre a responsabilidade moral individual acerca das escolhas realizadas. Nenhuma chance externa ou elemento aleatório poderá salvar o agente moral da inexorável responsabilidade pela sua decisão. Nem o Oráculo de Delfos, nem Deus, nem a natureza, a sorte ou uma máquina artificial perfeita e profética salvará o indivíduo do peso da sua liberdade. A escolha é inexoravelmente individual.

Apesar de singular, a escolha pode ser afetada por informações externas que afetam a decisão do agente moral, ou seja, a deliberação pode ocorrer sob condições informacionais imperfeitas.

Duas ordens de limitações podem ocorrer, limitações do agente moral *per se* ou pelo meio em que se encontra. O indivíduo possui racionalidade limitada e, portanto, as suas

---

[240] LEIBNIZ, 1930, p. 231-256.

[241] Cf. "*Stéphan Geonget fait, dans un second temps, l'inventaire des méthodes de résolution des cas perplexes par le juge : 1/ interpréter et conci-lier les textes contradictoires ; 2/ s'en remettre à l'opinion commune ; 3/ préférer l'une des deux parties à l'autre ; 4/ renoncer à juger l'affaire et renvoyer les parties ; 5/ recourir aux dés ; 6/ en appeler à la fiction ; 7/ s'abandonner au jugement du prince ; 8/ espérer de Dieu un miracle*". FERRER, Véronique. Stéphan Geonget, La notion de perplexité à la Renaissance. *Revue de l'histoire des religions*, v. 3, 2008. Disponível em: http://journals.openedition.org/rhr/6763. Acesso em: 29 jun. 2020, à 23:36.

[242] ARTOSI, Alberto; PIERI, Bernardo; SARTOR, Giovanni. *Leibniz*: LogicoPhilosophical Puzzles in the Law. Philosophical Questions and Perplexing Cases in the Law. Heidelberg: Springer, 2013. p. xxiv.

[243] ARTOSI; PIERI; SARTOR, 2013, p. xxvi.

escolhas estão sujeitos a erro. De outro lado, agentes podem afetar as informações recebidas pelo indivíduo.

Um estratégia oferecida para auxiliar a decisão do agente moral poderia ser a "resolução menos danosa" (*Lesser Evil Resolution*) defendida por *Gratian, William of Auxerre* e autores da "Summa Halesiana"[244]. Os conflitos morais aparentes seriam muito mais obra de um agente tolo, incapaz de discernir corretamente qual a conduta moral a seguir[245], do que uma dificuldade lógica. Alegar a existência de um conflito insolúvel seria em verdade uma "muleta" para justificar os erros de sua falha de entendimento moral e de exclusão moral pela deliberação falha[246].

Capreolus, outro autor escolástico, irá considerar se um agente moral com boas intenções estará protegido de realizar o mal. O mesmo poder-se-ia dizer da situação em que a sua razão encontra-se obnubilado por um gênio maligno (*Deceiving Demon Dilemma*)[247]. Antecedia *Capreolus* séculos antes o problema do gênio maligno (*deceptor*) proposto por *Descartes*. A solução sugerida seria tentar encontrar o menor dos males ou a escolha menos danosa (*de duobus malis minus malum eligendum est*).

Mas a resposta do autor será negativa. Mesmo um agente bem-intencionado pode cometer o mal (*ad bonitatem actus voluntatis requiritur appetitus recti finis, non tamen sufficit*). Digamos, afirmava o autor medieval, que alguém furte para dar, aos pobres, o fruto do seu furto, estaria ele protegido moralmente? *Capreolus* afirma que não, mesmo coberto das melhores intenções, o ato ainda seria um furto e errado moralmente[248].

O silogismo apresentado pelo autor apresenta a seguinte formulação[249]:

- premissa maior: "o menor de dois males devem ser escolhido";
- premissa menor: "a blasfêmia ou ódio a Deus é o menor dos males";
- conclusão (ato de consciência): "blasfemar ou odiar a Deus deve ser escolhido".

Um silogismo elaborado dessa forma foi atacado pelos escolásticos, que enfatizavam que o uso de premissas menor em atos de raciocínio prático é errôneo[250]. A premissa maior da

[244] DOUGHERTY, M. V. *Moral dilemmas in medieval thought*: from Gratian to Aquinas. Cambridge: Cambridge University Press, 2011. p. 171.
[245] DOUGHERTY, 2011, p. 169-171.
[246] DOUGHERTY, 2011, p. 175.
[247] DOUGHERTY, 2011, p. 177.
[248] DOUGHERTY, 2011, p. 175.
[249] DOUGHERTY, 2011, p. 175.
[250] DOUGHERTY, 2011, p. 179-181.

"escolha do menor dos males" contém implicitamente a assunção de que esta se constitui em uma *verdade moral autoevidente* (*self-evident truth of moral reasoning*).

A fórmula correspondente determinava que a intenção de fins corretos perdoava eventuais males (*ex praedictis sequitur quod intentio recti finis excusat a peccato*). O pressuposto era o de que o agente deveria necessariamente escolher e que essa escolha praticamente era uma necessidade e não uma voluntariedade. Seria quase uma escolha por coação e não por intenção.

Essa concepção será realçada por *Statman* (1990) ao analisar o conceito de agente "moralmente admirável" para *Aristóteles*[251]. Segundo esse filósofo, um sujeito moralmente bom não é o que realiza muitos atos "bons", mas que possui um caráter bom. Mesmo que tenha de realizar uma ação ruim, ele sentir-se-á culpado, ainda que tal conduta seja justificada[252]. O papel das emoções na solução dos dilemas morais será retomado na teoria contemporânea, como veremos em seguida, mas primeiro vejamos a solução racionalista do problema, em *Kant*.

A solução de conflitos morais em um modelo deontológico foi objeto de análise por *Kant*. É importante analisar a resposta dada por este autor, mesmo considerando-se impossível manter-se, de modo consistente, modelos morais em modelos morais de topo (*top-down*) para sistemas de inteligência artificial. A ideia de que o imperativo categórico[253] possa se caracterizar como uma verdade inquestionável foi duramente questionada[254] e afastada na construção de designs de agentes morais artificiais. Outra crítica ainda mais profunda ao modelo kantiano decorre da já vista incapacidade de os imperativos categóricos indicarem a conduta correta em uma situação concreta. O dever pelo dever é vazio de conteúdo e incapaz de organizar adequadamente a deliberação sobre meios e fins.

Kant irá negar a possibilidade de existirem legítimos *conflitos insolúveis*. Para o autor "dever implica em poder" (*ought implies can – Sollen impliziert Können*)[255]. Os conflitos se originam em diferentes graus de fundamentação de deveres ou entre deveres e inclinações. Não existiriam conflitos diretos entre deveres de mesmo nível. A tentativa de *Kant* em construir um sistema de moralidade pura impede qualquer consideração empírica na construção teórica da

---

[251] STATMAN, 1990, *apud* NUNES, Lauren de Lacerda; TRINDADE, Gabriel Garmendia da. Conflitos morais insolúveis e sistemas racionalistas: uma abordagem sobre consistência moral. *Princípios*, Natal, v.18, n.30, p. 85-100 jul./dez. 2011, p. 88.

[252] NUNES; TRINDADE, 2011, p. 88.

[253] Considera-se que exista somente um imperativo categórico, mas cinco fórmulas ou formulações, no sentido de Paton.

[254] FOOT, Philippa. Morality as a System of Hypothetical Imperatives. *The Philosophical Review*, v. 81, n. 3, p. 305-316, jul. 1972.

[255] Diversas são as contraposições históricas a este conceito, desde o Direito Romano, com a sua célebre fórmula *Impossibilium nulla obligatio est* (*Digesta* 50,17,185).

lei moral. Nenhum princípio poderia assentar-se em qualquer matéria de fato de qualquer natureza (propósito, intenção ou valor substantivo)[256].

A *moralidade kantiana* não se assenta sobre a ideia de um único axioma moral a dirigir todos os dilemas morais e respondê-los, tampouco prevê a possibilidade da inexistência de preceitos morais distintos. Pelo contrário, admite a possibilidade de surgirem diversas prescrições alternativas de um imperativo. Desse modo, é intrigante que o autor não admita a possibilidade de conflitos morais ou, ao menos, os considere como um problema relevante da moralidade[257].

Kant trata do tema na obra "Metaphysics of Morals", em que elucida o seu raciocínio em quatro passos argumentativos. Primeiro, define o que significam deveres ou obrigações. A seguir afirma que esses conflitos são conceitualmente impossíveis. Logo, então afirma que podem existir conflitos entre razões de obrigações (*grounds of obligation*) e não exatamente entre obrigações. Por fim, conclui que nesse caso as razões de maior peso suplantam as de peso inferior[258]. Ele trata do tema no seguinte parágrafo[259]:

> Um conflito de deveres (*collisio officiorum, s. obligationum*) seria uma relação recíproca na qual um deles [dos deveres] cancelasse o outro (inteira ou parcialmente). Mas visto que dever e obrigação são conceitos que expressam a necessidade prática objetiva de certas ações, e duas regras mutuamente em oposição não podem ser necessárias ao mesmo tempo, se é um dever agir de acordo com uma regra, agir de acordo com a regra oposta não é um dever, mas mesmo contrário ao dever; por conseguinte, uma colisão deveres é inconcebível. Entretanto, um sujeito pode ter uma regra que prescreve para si mesmo dois fundamentos de obrigação (*rationes obligandi*), sendo que um ou outro desses fundamentos não é suficiente para submeter o sujeito à obrigação (*rationes obligandi non obligantes*), de sorte que um deles não é um dever[260].

Se dois deveres fossem devidos simultaneamente, para *Kant* um deles não seria uma obrigação objetiva. Não seria o caso da existência de regras em oposição (devo A e não devo A). Nesse caso, o agente moral estaria perante um dever mais fraco, denominado de *prima facie*, ainda incapaz de obrigar, mas suficiente para indicar uma conduta para o agente. Esse dever *prima facie* confrontado por uma obrigação cederia prevalência para o dever mais forte ou de maior peso normativo[261]. Não existiria um dilema genuíno, apenas um conflito aparente.

---

[256] TIMMERMANN, Jens. Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory. *AGPh*, v. 95, n. 1, De Gruyter, p. 36–64, 2013, p. 37.
[257] TIMMERMANN, 2013, p. 37.
[258] TIMMERMANN, 2013, p. 41.
[259] DI NAPOLI, Ricardo Bins. Conflitos de deveres e a casuística na filosofia moral de Kant. *Studia Kantiana*, v. 11, p. 178-200. Disponível em http://www.sociedadekant.org/studiakantiana/index.php/sk/article/download/96/47. Acesso em: 16 dez. 2020 às 21:20.
[260] KANT, 2003, p. 67.
[261] DI NAPOLI, 2011, p. 185.

O que *Kant* admite é a existência de conflitos entre razões para agir, denominados de *fundamentos de uma obrigação* (*rationes obligandi*)[262]. A obra de Kant é inconclusiva sobre quais seriam esses fundamentos em conflito e como seriam resolvidos. Em um momento afirma que cada dever possui tão somente um fundamento e, em outro momento, afirma que possa possuir mais de um. "Quando dois fundamentos tais conflitam entre si, a filosofia prática não diz que a obrigação do mais forte tem precedência (*fortior obligatio vincit*), mas que o fundamento de obrigação mais forte prevalece"[263].

Kant restringe o problema dos conflitos morais a dilemas aparentes, insuscetíveis de acarretarem um choque insolúvel, seja entre obrigações ou fundamentos de obrigações morais.

O tema dos conflitos morais irá ressurgir contemporaneamente no exemplo citado por *Jean Paul Sartre* sobre um rapaz que não consegue escolher entre aderir às forças francesas de resistência ao nazismo ou cuidar de sua querida mãe doente. No caso, o rapaz estabelecia o mesmo peso valorativo para cada decisão, tornando-se difícil decidir qual escolha tomar[264].

Outro exemplo semelhante bastante citado é o famoso caso da "escolha de Sofia" de *Styron* (1979). Nele, Sofia é obrigada escolher qual dos dois filhos irá encaminhar para a morte, na câmara de gás, caso contrário, os dois deveriam ser sacrificados[265]. A resposta de *Williams* (1965) para esse dilema foi a de que se tratava de uma experiência individual contraditória, incapaz de ser solucionada por algum recurso racional, ou seja, era um *dilema moral insolúvel*[266]. Qualquer curso de ação a ser tomado pelo agente acarretaria inexoravelmente um sentimento de remorso, de culpa ou arrependimento.

De modo diverso, poder-se-ia alegar que a existência de dilemas morais insolúveis viola dois princípios morais: de *agregação* (*axioma da lógica deôntica*) e do "dever que implica poder" (Princípio de *Kant*)[267].

O princípio da agregação é citado, inicialmente, por *Kant*, que afirma[268]:

> Impulsos da natureza, consequentemente, envolvem obstáculos na alma do ser humano ao seu cumprimento do dever e forças (por vezes poderosas) que a ele se opõem ao que ele precisa avaliar que é capaz de resistir e subjugar pela razão, não em alguma ocasião no futuro, mas imediatamente (no momento em que pensa no dever):

[262] DI NAPOLI, 2011, p. 181.
[263] KANT, 2003, p. 67.
[264] NUNES; TRINDADE, 2011, p. 86.
[265] WILLIAMS, B. A. O.; ATKINSON W. F. Ethical Consistency. Proceedings of the Aristotelian Society. *Supplementary Volumes*, v. 39, p. 103-138, 1965.
[266] NUNES; TRINDADE, 2011, p. 87.
[267] NUNES; TRINDADE, 2011, p. 87.
[268] NUNES; TRINDADE, 2011, p. 88.

> ele tem que considerar que pode fazer o que a lei lhe diz incondicionalmente que ele deve fazer[269].

Encontramos na "República" de *Platão* uma das primeiras menções filosóficas a um conflito moral. Nessa obra *Sócrates* questiona *Céfalo* sobre o que é a justiça, e este responde: "é pagar o que se deve", a que *Sócrates* replica com um exemplo hipotético. Digamos que alguém peça emprestado, a um amigo, uma arma, mas, no momento de devolvê-la, veja que esse mesmo amigo encontra-se em estado de perturbação mental que possa machucar-se a si mesmo. Deveria o amigo, mesmo sabendo desse risco, devolver a arma, cumprindo o dever moral de devolver ou deveria preservar a integridade física do seu amigo?[270]

Estaríamos perante o conflito de dois comandos contraditórios ou seria um conflito aparente? Se os deveres possuíssem pesos valorativos distintos, estaria afastado o dilema, caso contrário, haveria um genuíno conflito.

A análise lógica dos *conflitos morais* foi realizada por *Williams* (1965) que diferenciou duas situações distintas[271]:

a) *Devo fazer a* e *devo fazer b*, mas não posso fazer ambos $((Oa \wedge Ob) \wedge \neg\Diamond (a \wedge b))$. Nesse caso somente existirá uma inconsistência lógica se houver o acréscimo de uma outra condição. De que devo fazer a e b ao conjuntamente e ao mesmo tempo, denominado de *princípio da agregação*;

b) *Devo fazer* a e *Devo não fazer* a.

O argumento de *Williams*[272] pode ser apresentado da seguinte forma[273]:

1. Oa premissa
2. Ob premissa
3. $\neg \Diamond (a \wedge b)$ premissa
4. $Oa \wedge Ob$, conjunção de 1 e 2
5. $Oa \wedge Ob \rightarrow O (a \wedge b)$ princípio de aglomeração

---

[269] KANT, 2003, p. 224. Ver texto em alemão. Disponível em: https://korpora.zim.uni-duisburg-essen.de/kant/aa06/380.html. Acesso em: 30 jun. 2020 às 13:01. "*Die Antriebe der Natur enthalten also Hindernisse der Pflichtvollziehung im Gemüth des Menschen und (zum Theil mächtig) widerstrebende Kräfte, die also zu bekämpfen und durch die Vernunft nicht erst künftig, sondern gleich jetzt (zugleich mit dem Gedanken) zu besiegen er sich vermögend urtheilen muß: nämlich das zu können, was das Gesetz unbedingt befiehlt, daß er thun soll*".

[270] NUNES; TRINDADE, 2011, p. 88.

[271] WILLIAMS; ATKINSON., 1965, 103-108.

[272] O argumento original de *Williams* continha a seguinte estrutura: "*Using these, the conflict can be represented in the following form: (i) I ought to do a; (ii) I ought to do b; (iii) I cannot do a and b. From (i) and (ii), by agglomeration; (iv) I ought to do a and b; from (iii) by ' ought' implies ' can' used contrapositively, (v) It is not the case that I ought to do a and b*". WILLIAM; ATKINSON, 1965, p. 118.

[273] NUNES; TRINDADE, 2011, p. 91.

6. O (a∧b) *modus ponens* 4 e 5
7. ¬ ◊ (Oa∧Ob) → O (a∧b) contrapositiva do princípio de aglomeração
8. ¬ O (a∧b) *modus ponens* 3 e 7
9. Contradição de 6 e 8.

O conflito de tipo 1 não apresenta necessariamente uma inconsistência lógica, salvo se houver a inclusão de premissas extras. Uma estratégia seria a redução do caso de tipo 1 (*devo-devo não*), no caso de tipo 2 (*devo-não devo*). Tal situação é considerada complicada por *Williams*[274], que afirma que se teria de transformar o tipo 1 em uma exigência de ações diferentes e antagônicas, sem o uso de premissas adicionais, tais como o *princípio da agregação*.

Para *Willians* a tomada de decisão em conflitos morais envolveria um "resquício emocional" para o agente, frente a escolhas éticas contraditórias[275]. No fundo o autor nega a possibilidade de solução racional em conflito ético, face à impossibilidade de uma escolha fundada em razões morais.

A teoria de *Gowans* irá propor um equilíbrio reflexivo entre as crenças morais (certo e errado), mas também de valores pessoais, ou seja, sobre as relações dos sujeitos com outras pessoas[276].

Uma outra resposta para o problema dos dilemas morais seria o *recurso à intuição*, como por exemplo a compaixão. As emoções não seriam somente um dado estranho ao raciocínio moral, mas um elemento importante para a deliberação moral. A situação conflituosa irá gerar um estado subjetivo de aflição moral que impactará a escolha a ser tomada e pode estender os seus efeitos, mesmo depois da escolha tomada. Trata-se de uma resposta distinta dos racionalistas, que negam aos sentimentos ou às emoções o estatuto de *recurso deliberativo genuíno*. Pelo contrário, creditam a esses estados subjetivos da consciência um papel negativo, ao desvirtuar o entendimento e prejudicar a razão no seu protagonismo em decidir sem as amarras do contingente, do empírico ou do concreto[277]. As emoções deixam profundas raízes na natureza humana, vide as descobertas da neurociência sobre a relação entre a oxitocina (OXT) e o sentimento de compaixão[278].

---

[274] WILLIAM; ATKINSON, 1965, p. 103-138.
[275] NUNES; TRINDADE, 2011, p. 88.
[276] DI NAPOLI, 2012, p. 79.
[277] DI NAPOLI, 2012, p. 83.
[278] CHURCHLAND, Patricia S.; WINKIELMAN, Piotr. Modulating social behavior with oxytocin: How does it work? What does it mean? *Hormones and Behavior*, v. 61, n. 3, p. 392-399. March 2012. Disponível em:

Mas o que seria a intuição? Para *Audi* as intuições são respostas não inferenciais às experiências, ou seja, não sustentadas por uma premissa[279]. Os intuitivistas acreditam que, se uma intuição é verdadeira, então existe justificação *prima facie* para acreditar nela[280]. As intuições seriam uma espécie de crenças (*I have treated cognitive intuitions as a kind of belief*). Geralmente são consideradas *prima facie* as proposições morais autoevidentes, de caráter geral.

O intuitivismo parece duvidar da racionalidade como único critério de justificação para toda e qualquer proposição moral, algumas seriam "sólidas", incapazes de serem sindicalizadas pela razão. Para *Gowans* a racionalidade isoladamente não consegue solucionar dilemas morais, tampouco leis abstratas e racionais conseguem indicar a ação moral a ser tomada em um conflito moral[281].

Concordamos com esse entendimento, dado que, para que isso ocorresse, deveria existir um agente onisciente sobre todos os fatos e efeitos, diretos e indiretos, da ação moral, para que escolhesse a ação mais virtuosa. Nenhuma máquina seria capaz de processar todos os dados envolvidos em tal decisão, o que afasta a estratégia deontológica e utilitarista. Não se trata de uma estratégia cética ou irracionalista, dado que aceita e admite a capacidade de entender e conhecer a moralidade. O que se afasta é um fundacionalismo moral, que somente afirma que as crenças são justificáveis se forem certas e evidentes[282]. Se nossas crenças são conectadas com sentimentos, então podemos falar da existência de um conhecimento moral.

Esse entendimento confirma a tese de *Aristóteles* de que razões morais incompletas ou não racionais são guias adequadas para o agente performar ações que ele acredita serem corretas. Para o autor a virtude moral completa é composta de virtudes não racionais e *phronêsis* (prudência ou sabedoria prática)[283]. Esse seria o material racional para o agente responder a um dilema moral, em sua da busca da *eudaimonia* (felicidade, bem-estar ou plenitude)[284].

O problema dos conflitos morais, como observado, apresenta ainda uma vasta agenda de pesquisas, contudo, demonstra a importância do tema para o estudo dos Agentes Morais

---

https://www-sciencedirect.ez94.periodicos.capes.gov.br/science/article/pii/S0018506X11002807?via%3Dihub, Acesso em: 30 jun. 2020, às 00:51.

[279] Cf. "*I am taking intuitionism as an ethical theory to be, in outline and in a minimal version, the view that there is at least one moral principle that is non-inferentially and intuitively knowable*". AUDI, R. Intuitions, intuitionism, and moral judgment. *In:* AUDI, R. *Reasons, Rights, and Values*. Cambridge: Cambridge University Press, 129-159, 2015, p. 133.

[280] AUDI, 2015, p. 133.

[281] DI NAPOLI, 2012, p. 93.

[282] DI NAPOLI, 2012, p. 93.

[283] ANGIONI, Lucas. Phronesis e virtude do caráter em Aristóteles: Comentários à Ética a Nicômaco VI. *Dissertatio*, v. 34, p. 303-345, 2011.

[284] ENGBERG-PEDERSEN, Troels. Aristotle's Theory of Moral Insight. Review by Alfred R. Mele. *The Philosophical Review*, v. 94, n. 2, p. 273-275, Apr. 1985.

Autônomos (AMA). Estes, além de racionalidade, autonomia, vontade, crenças e responsabilidade, teriam intuições, emoções e sentimentos como a compaixão e empatia? Como se portarão perante conflitos morais?

2 SEGUNDA PARTE. AGENTES MORAIS ARTIFICIAIS (AMAS)

2.1 DA POSSIBILIDADE DE AGENTES MORAIS ARTIFICIAIS

Verificamos, na primeira parte deste trabalho, as condições de possibilidade para a existência de uma sujeito artificial, dotado de racionalidade, autonomia, vontade, responsabilidade e emoções. Verificou-se, igualmente, a possibilidade de modelos morais em inteligência artificial. A pergunta que se pretende investigar é sobre a possibilidade de existência de agentes morais artificiais (AMA) e não somente de máquinas dotadas de algoritmos morais.

No primeiro caso, o sistema artificial será autômato e responderá a padrões autônomos de decisão moral, com todos os dilemas e conflitos inerentes. No segundo caso, a máquina irá responder a um processamento prévio, conforme design e arquitetura alimentados por um programador, sob os limites previamente estabelecidos por este. Ou seja, **é possível existir um agente moral artificial autêntico, dotado de decisões morais próprias?**

Os receios da humanidade de que as máquinas morais evoluam para agentes morais fazem parte da contemporaneidade. Tornar-se-iam, essas máquinas, os nossos *Gollems* ou *Talos* modernos?

Iremos inicialmente verificar quais são as condições necessárias para que se possa considerar um agente moral artificial e, em seguida, o que significa o agir moral artificial.

2.1.1 Autonomia artificial: agentes morais implícitos e explícitos

A possibilidade da existência de agentes morais autônomos tem sido objeto de grande debate e curiosidade na literatura científica e filosófica. O tema tem sido tratado como uma decorrência natural do explosivo e exponencial desenvolvimento tecnológico atual. Quase como uma realidade inexorável. Da passagem de máquinas racionais, que imitam a racionalidade lógica humana, teríamos o surgimento de máquinas morais, que possuem racionalidade prática.

Muitas perguntas, ainda sem respostas, têm se somado ao problema. Afinal, há um caminho inexorável em direção ao surgimento de agentes morais artificiais? Ou, pelo contrário, essa é uma impossibilidade filosófica, limitada por restrições intrínsecas das máquinas, tais como pensadas por *John Searle* no teste da Sala Chinesa? Seria uma vã ilusão pensar que as máquinas possam algum dia verdadeiramente deliberar sobre escolhas morais?

O conceito de sujeito moral é o sujeito racional, autônomo, autoconsciente, dotado de vontade, livre e responsável. Podemos questionar se o conceito de agente moral artificial pode possuir os mesmos elementos necessários para a noção de sujeito.

A literatura recente tem diferenciado os agentes morais conforme o seu grau de autonomia: de um lado teríamos os agentes morais implícitos e, de outro, os agentes morais explícitos. A principal distinção entre os dois casos está em quem detém a capacidade deliberativa, a máquina autonomamente ou um agente externo (programador humano)[285].

Há autonomia moral no caso de agentes morais explícitos (*explicit ethical agente*), ou seja, o sistema é capaz de tomar decisões por si próprio, sem recorrer à deliberação externa, como se fosse um fantoche. O agente se autogoverna, autolegisla e decide qual escolha tomar sobre como deve agir.

Mas não basta o agente moral tomar decisões, ele deve ser capaz de *justificar* suas ações. Deve ser capaz de encontrar razões para agir nesse sentido. É a célebre distinção kantiana entre agir com um senso de dever (*acts from a sense of duty*), conforme um princípio ético, e meramente em acordo com um dever (*accordance with duty*)[286].

Os debates sobre agentes éticos explícitos têm afastado o modelo de princípios para a construção de sistemas éticos inteligentes, dado que o sujeito deve deliberar sobre uma infinidade de casos não previstos originariamente pelo programa. A capacidade de se deliberar sobre fatos novos ou em situações não previstas não pode ser diretamente derivada de um pequeno conjunto de princípios éticos. Contudo, os estudos têm demonstrado que os princípios cumprem uma função diversa, porém não menos importante, em determinar um padrão de revisão ou justificação para determinada ação concreta[287].

Existem limitações técnicas atuais que conduzem tal escolha. De modo geral, a utilização de redes neurais tende a privilegiar modelos de construção de soluções baseadas em modelos de programação com reforço de inferências que se reafirmam, ao estilo da programação de baixo para cima (*bottom-up*)[288].

A possibilidade de construírem-se agentes éticos artificiais fundamenta-se em duas premissas, conforme *Howard*[289]:

---

[285] ANDERSON, Michael; ANDERSON, Susan Leigh. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, v. 28, n. 4, p. 15-27, 2007. p. 17.
[286] ANDERSON; ANDERSON, 2007, p. 17.
[287] ANDERSON; ANDERSON, 2007, p. 17.
[288] HOWARD; MUNTEAN, 2016.
[289] HOWARD; MUNTEAN, 2016.

i) Há similaridade entre a *moralidade humana e a artificial*; e
ii) há similaridade entre a *cognição humana e moralidade*.

No primeiro caso, afirma o autor que é possível construir uma moralidade artificial com base no princípio da universalização. Seria possível pensar-se em uma teoria da agência ampla e não apenas humana? Haveria uma moralidade dos sujeitos autônomos, livres, conscientes, racionais e emocionais, responsáveis por seus atos perante a si e outros agentes?

Alguns autores defendem a impossibilidade de *replicação* da moralidade humana. A intencionalidade, a consciência ou a responsabilidade moral não seriam replicáveis artificialmente. Faltariam os elementos básicos para um agente moral artificial. Para *Johnson* "nem o comportamento da natureza nem o comportamento das máquinas são passíveis de explicações racionais, e a agência moral não é possível quando uma explicação racional não é possível"[290]. O problema tornar-se-ia ainda mais complexo se fosse adicionado o elemento emocional para a agência moral.

Essa impossibilidade lógica submeteria os agentes artificiais à condição de sistemas dependentes (*surrogate agents*)[291]. Não mereceriam nem mesmo a denominação de verdadeiros agentes, no sentido de sujeitos morais. Há, contudo, um certo entendimento majoritário de que o estatuto moral da humanidade não é uma condição excepcional da espécie humana[292].

Não há uma "agência moral excepcionalmente humana" (*essentially human agency*), estabelecida em bases ontológicas ou *a priori*[293]. A *distinção entre agência moral humana e artificial* deve ser procurada em outros fatores de delimitação (maior generalidade, grau de abstração ou complexidade). Exclui-se a *tese da excepcionalidade* da possibilidade da agência moral, como algo somente ou demasiadamente humano. Seríamos uma espécie de agentes

---

[290] Cf. "*Neither the behavior of nature nor the behavior of machines is amenable to reason explanations, and moral agency is not possible when a reason-explanation is not possible*". Ver JOHNSON, Deborah G. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, v. 8, n. 4, p. 195-204, 2006.
[291] MUNTEAN, Ioan; HOWARD, Don. A Minimalist Model of the Artificial Autonomous Moral Agent (AAMA). *Philpapers.org*, 2016. Disponível em: https://philpapers.org/rec/MUNAMM.
[292] DANIELSON, P. *Artificial morality virtuous robots for virtual games*. London: New York: Routledge, 1992; DANIELSON, P. (ed.). *Modeling rationality, morality, and evolution*. New York: Oxford University Press, 1998; ALLEN, C.; VARNER, G.; ZINSER, J. Prolegomena to anyfuture artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 12, n. 3, p. 251–261, 2000; ABNEY, K.; LIN, P.; BEKEY, G. A. (ed.). *Robot Ethics*: The Ethical and Social Implications of Robotics. The MIT Press, 2011; ANDERSON, M.; ANDERSON, S. L. (ed.). *Machine Ethics*. Cambridge University Press, 2011; WALLACH, W. *A Dangerous Master*: How to Keep Technology from Slipping Beyond Our Control. Basic Books, 2015; WALLACH, W.; FRANKLIN, S.; ALLEN, C. A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Topics in Cognitive Science*, v. 2, n. 3, p. 454-485, 2010.
[293] MUNTEAN; HOWARD, 2016.

morais, o que demonstraria a *tese da universalização* dos princípios morais, que podem ser instanciados[294] por mais de um modo.

Ou, por outro lado, dada essa similaridade, a construção de um agente moral artificial seria um ato de replicação da agência moral humana? Neste último caso, teríamos de aceitar que os sistemas inteligentes complexos possuiriam, em seu código de base, os vieses e as predisposições de seu programador. Apesar de a teoria ética ser uma preocupação constante na história da humanidade, não podemos considerar a humanidade como um referencial ético absoluto ou isento de contradições. Talvez a moralidade humana esteja condicionada por suas características contingentes. Os seres humanos são, para *Eric Dietrich*, seres biológicos em constante competição com outros. Assim, a moralidade humana é afetada claramente por um mecanismo genético que privilegia o mecanismo de sobrevivência (*survival mechanism*)[295].

Replicar a moralidade humana pode reproduzir, mesmo inconscientemente, nossos piores defeitos. Uma criatura assim surgida pode assemelhar-se ao personagem do jovem Frankenstein, nem totalmente humana nem totalmente artificial, um ente com uma cisão interna irreconciliável. Talvez fosse a reedição do mito de Prometeu, ressurgindo novamente com os riscos inerentes à ambição (i)legítima de superação das limitações naturais, de domínio completo da natureza e do "proibido" controle da criação[296].

Poderia, ao contrário, para *Dietrich*, ser a possibilidade histórica de superação das limitações éticas do ser humano, de seu mecanismo genético de sobrevivência, permitindo o surgimento de um agente moral artificial livre das condicionantes antiéticas dos humanos. Seriam sistemas surgidos sobre as boas características morais humanas, uma espécie de versão melhorada dos humanos, uma "humanidade 2.0".

Haveria a possibilidade de grandes descobertas na teoria ética, livre das amarras do *comportamento humano antiético* (*unethical human behavior*)[297]. Esse novo horizonte otimista enxerga um progresso histórico linear, com uma racionalidade histórica interna consistente. No princípio encontramos o indivíduo lutando desesperadamente para alcançar a superação de sua condição miserável, de submisso às implacáveis ações da natureza. A seguir aperfeiçoando o

---

[294] Em linguagem de programação, "instanciar" uma classe significa adicionar um objeto àquela classe (*class instance*).
[295] ANDERSON; ANDERSON, 2007, p. 17.
[296] CARDOZO CIACCO, Felippe. Sobre o monstro, a natureza e a origem: uma releitura de Frankenstein ou o Prometeu moderno. *Outra Travessia*, Florianópolis, n. 22, p. 161-174, ago. 2016. ISSN 2176-8552. Disponível em: https://periodicos.ufsc.br/index.php/Outra/article/view/2176-8552.2016n22p161/34652. Acesso em: 02 jul. 2020; SHELLEY, Mary. *Frankenstein ou o Prometeu moderno*. Rio de Janeiro: Nova Fronteira, 2011; SHELLEY, Mary. *Frankenstein, Or the Modern Prometheus*. Engage Books, 2008; PARK, Katharine; DASTON, Lorraine J. Unnatural conceptions: the study of monsters in sixteenth-and seventeenth-century France and England. *Past & Present*, n. 92, p. 20-54, 1981.
[297] ANDERSON; ANDERSON, 2007, p. 17.

seu domínio sobre a natureza pela técnica e, finalmente, controlando a natureza e o ato de criação por meio da tecnologia (bioética ou inteligência artificial). Mas não seria esse caminho demasiado perigoso sem os cuidados necessários? Por outro lado, seria possível o surgimento de uma nova teoria moral artificial distinta das imaginadas até o momento, talvez até incompreensível para a racionalidade atual?

Experimentos recentes demonstraram a possibilidade do surgimento de uma linguagem artificial a partir da linguagem natural fornecida pelo sistema. Também foram registrados casos em que uma linguagem artificial surgiu espontaneamente a partir do uso *default* de uma linguagem natural. Tal situação, não controlada, ficou famosa no controverso caso do "chatbot" de negociação do Facebook, que teria sido corrigido ou "desligado" [298], por ter fugido ao escopo inicial da programação pretendida. Nessa situação um "robô" de negociação passou a se comunicar com outro "robô" em uma linguagem desconhecida pelos programadores, o que exigiu uma intervenção.

Estudos recentes têm destacado a possibilidade do surgimento de uma linguagem artificial desde o início ou "do zero", especialmente com o uso de sistemas inteligentes de ambiente multiagente e com métodos de aprendizado de máquina, sem nenhum contato inicial com a linguagem natural. Trata-se de experimentos muito recentes, mas que demonstram as grandes possibilidades de desenvolvimento futuro[299].

Se pode surgir, de modo controlado ou espontâneo, um linguagem artificial compreensível ou não à racionalidade humana, será que não poderiam surgir, igualmente, regras morais próprias desses agentes artificiais, compreensíveis ou não, para os programadores

---

[298] Há controvérsias sobre o surgimento dessa linguagem opaca aos programadores do Facebook, bem como se esta seria uma linguagem mais eficiente do que a linguagem natural. Sobre um histórico do assunto, veja-se os seguintes artigos: KUCERA, Roman. *The truth behind Facebook AI inventing a new language*. Disponível em: http://errancesenlinguistique.fr/02-Journal/14/MachineLanguage.pdf. Acesso em: 02 jul. 2020 às 12:56; ALEXANDER Sneha. *How the story of Facebook "shutting" its ai after bots invent own language unfolded*. Disponível em: https://www.boomlive.in/how-the-story-of-facebook-shutting-its-ai-after-bots-invent-own-language-unfolded/. Acesso em: 02 jul. 2020 às 12:24. Disponível em: https://www.fastcompany.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it. Acesso em: 02 jul. 2020 às 12:26.

[299] Cf. "*We have presented a multi-agent environment and learning methods that brings about emergence of an abstract compositional language from grounded experience. This abstract language is formed without any exposure to human language use. We investigated how variation in environment configuration and physical capabilities of agents affect the communication strategies that arise*". MORDATCH, Igor; ABBEEL, Pieter. Emergence of grounded compositional language in multi-agent populations. *In*: THE THIRTY-SECOND AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-18), 37., 2018, New Orleans, p. 1495-1502, p. 1502. Disponível em: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17007/15846. Acesso em: 02 jul. 2020 às 14:54.

humanos? Somente por essa razão e preocupação, já bastaria incluir no programa de pesquisas científicas[300] o problema da possibilidade filosófica de agentes morais artificiais explícitos.

De modo geral, podemos afirmar que é possível falar-se em agência moral artificial, seja pela *tese da universalidade* ou da *replicação por similaridade*. Trata-se de uma demonstração por inferência ou dedução. Uma estratégia distinta para determinar a existência de uma legítima agência moral artificial seria aplicar uma versão do Teste de Turing a uma máquina artificial e determinar se ela é capaz de reproduzir com competência uma racionalidade prática, indistinguível de um ser humano.

2.1.2 Teste de Turing Moral

O Teste de Turing foi utilizado para determinar as situações em que um agente artificial agiria de modo competente a tal ponto de se tornar indistinguível de um agente humano. A questão que surge é sobre a possibilidade de utilização de um teste semelhante para verificar a presença de um agente moral artificial explícito.

Allen será o primeiro autor a realizar um tratamento substantivo do Teste de Turing para sistemas inteligentes[301]. Nesse caso, o teste implicaria conversações entre uma máquina e interrogadores humanos, se estes não identificarem o sistema artificial, então este seria um agente moral[302]. A tese é a de que haveria uma correlação entre o Teste de Turing para verificação de inteligência artificial e para verificação de um agente moral artificial.

Allen e *Wallach* serão os primeiros autores a designarem essa espécie similar de jogo de imitação moral de Teste de Turing Moral (*Moral Turing Test* – MTT)[303]. Uma das dúvidas iniciais era qual *modelo ético* a ser utilizado nesse teste: deontológico, utilitarista ou da virtude.

Beavers elenca os parâmetros necessários para um teste nesse sentido: consciência (*consciousness*), intencionalidade (*intentionality*), livre-arbítrio (*free will*), reponsabilidade moral (*moral responsibility*) e (*moral accountability*)[304].

---

[300] LAKATOS, I. O falseamento e a metodologia dos programas de pesquisa científica. *In*: LAKATOS, I.; MUSGRAVE, A. (org.) *A crítica e o desenvolvimento do conhecimento*. São Paulo: Cultrix, 1979; LAKATOS, I. History of science and its rational reconstructions. *In*: HACKING, I. (org.). *Scientific revolutions*. Hong-Kong: Oxford University, 1983.

[301] ARNOLD, Thomas; SCHEUTZ, Matthias. Against the Moral Turing Test: accountable design and the moral reasoning of autonomous systems. *In*: SCHEUTZ, Matthias (dir.). *Hrilab*. Medford, 2016. Disponível em: https://hrilab.tufts.edu/publications/arnoldscheutz16mtt.pdf. Acesso em: 02 jul. 2020 às 22:48.

[302] ARNOLD; SCHEUTZ, 2016.

[303] ARNOLD; SCHEUTZ, 2016.

[304] Cf. "*Though this might sound innocuous at first, excluded with this list of inessentials are not only consciousness, intentionality, and free will, but also anything intrinsically tied to them, such as conscience, (moral) responsibility, and (moral) accountability*". BEAVERS, 2011, p. 340.

Floridi irá utilizar parâmetros totalmente diferentes para distinguir um objeto moral (*moral patient*) de um agente moral (*moral agent*). Para o autor, não se pode exigir que um agente moral artificial seja livre, consciente e responsável. Bastaria que possuísse interatividade (*interactivity*), autonomia (*autonomy*) e adaptabilidade (*adaptability*). A interatividade é a capacidade de responder a estímulos decorrentes da mudança de estados. A autonomia seria a capacidade de alterar seu status sem precisar de estímulos externos e a adaptabilidade significa possuir regras de transição para outros estados[305].

Apesar de sedutora, a tese de *Floridi* reduz muito o conceito de agente moral artificial a um conceito fraco e limitado, não condizente com a tradição moral. Seus critérios seriam mais adequados à assunção de um senciente artificial do que de um agente artificial. A tese rompe igualmente com a noção de similaridade entre a moralidade humana e a artificial. Reduzindo a agência moral artificial a uma sombra da agência humana, de modo a praticamente negar a possibilidade de existência de uma verdadeira agência moral artificial. Desse modo, iremos adotar os parâmetros apresentados por *Beavers*.

Allen irá sugerir que o MTT compare ações morais em vez de respostas verbais[306]. Em vez de um interrogador, haveria um sujeito que tentaria distinguir se o agente que realizou a ação era humano ou artificial. Se fossem indistinguíveis, o agente moral artificial passaria no MTT[307].

Nesse teste há uma mudança em relação ao TT clássico. Não se trata de uma verificação de performance linguística ou conversacional, mas de habilidade para agir em determinada situação moralmente relevante. Se a máquina não for identificada como o agente "menos moral" (*less moral member*), então ela passaria no teste. Trata-se de um método de resultado comparativo, por isso denominado de Teste de Turing Moral Comparativo (*comparative* MTT – cMTT)[308].

A principal razão para a mudança de estratégia do MTT decorre do profundo desacordo entre os filósofos morais sobre as teorias morais[309]. Qual seria a resposta correta para determinada situação? Deve-se mentir para se salvar um inocente ou a mentira por si só viola a lei moral? Desse modo, se optou por um outro caminho, a comparação de comportamentos em

[305] FLORIDI, L.; SANDERS, J. On the Morality of Artificial Agents. *Minds and Machines*, v. 14, p. 349-379, 2004.
[306] ALLEN, Colin; VARNER, Gary; ZINSER, Jason. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 12, n. 3, p. 251-261, 2000.
[307] ARNOLD; SCHEUTZ, 2016.
[308] ALLEN; VARNER; ZINSER, 2000, p. 255.

[309] STAHL, B. C. Information, ethics, and computers: the problem of autonomous moral agents. *Minds and Machines*, v. 14, p. 67-83, 2004.

determinada situação moralmente relevante[310]. Bastaria verificar como um agente artificial se comporta. Se ele agir de modo similar a um humano consciente e moralmente responsável, em uma situação relevante, então ele seria considerado um agente moral artificial[311].

O uso do Teste de Turing Moral (*Moral Turing Test* – MTT) para verificar a performance moral de sistemas artificiais inteligentes não é evidente *a priori*. Para *Arnold* e *Scheutz* o jogo de imitação de MTT sustenta-se em pressupostos frágeis, mesmo que se utilize de cenários morais computáveis. As principais falhas listadas pelos autores são a vulnerabilidade a erros (*vulnerable to deception*), raciocínios inadequados (*inadequate reasoning*) e performance moral ineficiente (*inferior moral performance*)[312].

Dentre as diversas fraquezas do MTT, segundo *Allen*, podemos citar o baixo nível de comparação entre máquinas e humanos. Afinal, o design do sistema poderia ter como *standard* a moralidade de uma criança ou similar, dado que a comparação deve prever que a máquina não se sairia moralmente pior do que um ser humano em qualquer situação similar. Por outro lado, existe uma certa tolerância a que seres humanos realizem escolhas morais errôneas, o que de modo algum é claro quando se trata de máquinas. Haveria a mesma tolerância para que elas agissem de modo imoral? *Allen* alerta que provavelmente iríamos exigir mais das máquinas do que de outros seres humanos. E os seus erros seriam menos tolerados[313].

Outro problema seriam as pequenas falhas morais. Não se admite de um lado ou de outro, para humanos e máquinas, que matem ou pratiquem grandes males, mas o que se dirá em relação às pequenas mentiras, às mentiras inofensivas, às mentirinhas brancas ou boas[314]. Geralmente, admitimos um certo grau de tolerância a estas, mas não se pode dizer o mesmo em relação às máquinas[315].

---

[310] Cf. "*A Moral Turing Test (MTT) might […] be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human* "*interrogators*" *cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent*". ALLEN; VARNER; ZINSER, 2000, p. 254.

[311] CROCKETT, Larry. AI Ethics: the thin line between computer simulation and deception. *In:* GRIFFITHS, Paul; NOWSHADE, Mitt Kabir. *Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics*. Oxford: ACPI, 2019. p. 83.

[312] ARNOLD; SCHEUTZ, 2016.

[313] ALLEN; VARNER; ZINSER, 2000, p. 255.

[314] Cf. "*Apesar da valoração negativa do fenômeno, é possível extrair alguns de seus aspectos positivos. No campo profissional, a mentira pode ser vista como uma habilidade importante no processo de comunicação, na resolução de problemas com os chefes, companheiros e clientes, e na resolução de negociações complexas*". MATIAS, Danilo Wágner de Souza; LEIME, Jamila Leão; WALENTINA, Carmem, TORRO-ALVES, Nelson; BEZERRA, Amorim Gaudêncio. Mentira: Aspectos Sociais e Neurobiológicos. *Psicologia: Teoria e Pesquisa*, v. 31, n. 3, p. 397-401, jul.-set. 2015, p. 397.

[315] ALLEN; VARNER; ZINSER, 2000, p. 255.

Talvez a vulnerabilidade a erros e outros defeitos do MTT sejam apenas uma limitação tecnológica atual[316], a ser superada pelo avanço "explosivo" da tecnologia. Nesse meio tempo, o uso do MTT deveria ter um escopo reduzido. Poderia ser utilizado tão somente como um objetivo geral a ser alcançado; como um teste de capacidade de um agente moral artificial[317], mas não de sua moralidade em si ou mesmo como modelo "enfraquecido" a ser utilizado para agentes morais artificiais dotados de inteligência artificial fraca (*weak* AI)[318].

Talvez o MTT jamais seja atingido pela *strong* AI ou mesmo pela *weak* AI[319], por mais vigoroso que seja o desenvolvimento tecnológico. Haveria um limite intransponível que impediria que um sistema artificial passasse no MTT. O julgamento moral humano seria tão robusto que jamais seria atingido por um agente artificial[320]. Esta última tese colide com o entendimento de que podem surgir agentes morais artificiais. Talvez o teste em si seja inadequado para verificar a presença de um agente moral artificial (AMM).

Arnold e *Scheutz* irão defender uma outra estratégia distinta do MMT, denominada de "verificacionismo" ("*verification*"). O principal erro desse teste está na sua opacidade, na sua falta de transparência sobre as justificativas para a escolha moral artificial[321]. Afinal, não basta o agente moral deliberar ou agir, ele deve saber por quais razões deliberou desta ou daquela forma.

Para *Allen* e *Wallach,* dever-se-ia substituir um MTT completo por um teste mínimo de MTT (*Moral Turing Test*) para se determinar a performance de *agente moral artificial explícito* (*explicit* AAMA).

Como se pode notar, a tese de que há similaridade entre a *moralidade humana e a artificial* admite teoricamente a possibilidade de condutas morais comparáveis, entre humanos e agentes artificiais, em situações relevantes. O teste comparativo (cMTT) entre os agentes esbarra ainda hoje em dificuldades tecnológicas intransponíveis, que, talvez, sejam no futuro superadas pelo desenvolvimento exponencial dos sistemas autônomos.

Outra crítica que poderia ser formulada à tese do agentes morais autônomos é a de que eles poderiam *imitar atuar* como agentes morais humanos, mas jamais *atuariam verdadeiramente*. Eles não disporiam de intencionalidade ou vontade própria, mas

---

316 ALLEN; VARNER; ZINSER, 2000, p. 259.
317 GERDES, A.; ØHRSTRØM, P. Issues in robot ethics seen through the lens of a moral Turing Test. *Journal of Information, Communication and Ethics in Society*, v. 13, n. 2, p. 98-109, 2015. Disponível em: https://portal.findresearcher.sdu.dk/da/publications/issues-in-robot-ethics-seen-through-the-lens-of-a-moral-turing-te. Acesso em: 04 jul. 2020 às 00:59.
318 ARNOLD; SCHEUTZ, 2016.
319 ARNOLD; SCHEUTZ, 2016.
320 ARNOLD; SCHEUTZ, 2016.
321 ARNOLD; SCHEUTZ, 2016.

responderiam a comandos predeterminados. Eles jamais compreenderiam o conteúdo de suas escolhas, ou seja, jamais passariam no teste da Sala Chinesa de moralidade.

2.1.3 Da objeção de consciência e intencionalidade: ausência de vontade própria

Turing irá expor o seu argumento contra a objeção de que as máquinas possam possuir consciência e intencionalidade. O argumento teria sido exposto pelo Professor Jefferson, em 1949. No seu discurso, teria descrito as intransponíveis barreiras para uma máquina[322]. Ela poderia talvez organizar palavras com engenhosidade ímpar. Equipar-se em métricas bem construídas, elaborar sonetos e sinfonias arrebatadoras, mas sem nunca entender o sentido desses símbolos, ter consciência de seus escritos ou sentir o tremor da alma na leitura. Sua condição inata jamais permitiria sentir a miséria da condição humana ou as delícias dos pequenos detalhes da vida humana. Ela jamais teria consciência ou visão em primeira pessoa.

Turing irá responder que a dificuldade em se responder ou encontrar ou localizar a consciência é semelhante em seres humanos. Para ele os defensores da objeção de consciência poderiam ser persuadidos a abandoná-la, preferencialmente, e adotar uma posição solipsista. Aceitar a armadilha de procurar a consciência em estados mentais interiores poderia incorrer em graves dificuldades epistemológicas. Assim, alerta o autor, seria mais adequado aceitar o seu teste como medida de verificação[323].

A objeção de consciência seria aprimorada e reorientada por *Searle*, sob a forma do Argumento da Sala Chinesa (*Chinese's Room Argument*). Este foi elaborado pelo autor para se opor à tese de que bastaria um sistema passar pelo Teste de Turing para que pudesse ser considerado inteligente[324].

Allen e *Wallach* apresentaram o problema dos agentes morais artificiais de modo muito claro em duas questões distintas[325]:

- pode um robô ser considerado um agente moral (questão ontológica)?; e

---

322 "*Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain-that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants*". Ver TURING, 1950, p. 443.

323 Cf. "*In short then, I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test*". TURING, 1950, p. 443.

324 Cf. Disponível em: https://moral-robots.com/philosophy/briefing-the-chinese-room-argument/. Acesso em: 04 jul. 2020 às 22:09.

325 ALLEN; WALLACH, 2008, p. 58.

- como podemos saber se um robô é um agente moral (questão epistemológica)?

Trata-se de questões distintas, porém conexas. Digamos que seja possível afirmar a existência filosófica e prática de um agente moral artificial, como podemos determinar se estamos perante um agente assim? As consequências dessa determinação são tremendas. Teriam direitos ou responsabilidades? O Direito poderia considerá-los como pessoas e não como objetos? Ou mesmo como seres sencientes? Seriam seres conscientes?

Vamos iniciar a análise destas questões pelo difícil problema da consciência. Para *Searle* os *modelos computacionais de consciência não são suficientes para constituir verdadeiramente um ser consciente*. O autor apresenta um exemplo ilustrativo ao afirmar que que ninguém supõe que o processamento de um modelo matemático de tempestades em Londres deixariam alguém molhado[326].

Podemos concordar com Searle que os modelos matemáticos são aproximações imperfeitas, pelo menos até o momento, da realidade. Contudo, é possível discordar que o processamento computacional não possa reproduzir sensações reais[327]. Afinal, o sentir-se molhado pode ser considerado como uma reação elétrica e bioquímica no cérebro[328], que pode artificialmente ser simulado. Estudos recentes demonstraram a possibilidade de modelagem matemática do "sentir-se molhado"[329]. Logo podemos considerar possível reproduzir artificialmente tal sensação.

Estudos atuais sobre a simulação computacional de estados cerebrais demonstraram tanto as limitações quanto as possibilidades desse programa de pesquisas. A *compreensão mecanicista* dos modelos matemáticos, dos neurônios e de seu funcionamento garantiu somente um esqueleto para o entendimento do funcionamento do cérebro. Essa limitação induziu a uma nova agenda de pesquisas denominada de "jogo de imitação biológico" (*biological imitation game*). O objetivo é a reprodução do funcionamento do comportamento real do cérebro, sob o

---

326 SEARLE, J. R. *Consciousness and Language*. Berkeley: Cambridge University Press, 2002, p. 16.

327 Cf. "[...] *sensory feelings are not properties of molecules or events in the external world; they are the evolved adaptive illusions of a conscious mind*". JOHNSTON, V. S. *Why we feel*: The science of human emotions. Reading, MA: Perseus Books, 1999, p. 7.

328 Cf. "[...] *sensations and feelings, are a product of the physical and chemical organization of the brain*". JOHNSTON, 1999, p. 7.

329 Cf. "*This model supports the hypothesis that the brain infers about the perception of wetness in a rational fashion, taking into account the variance associated with thermal afferents and mechanoafferents evoked by the contact with wet stimuli, and comparing this with a potential neural representation of a* "*typical wet stimulus*", *which is based on prior sensory experience*". FILINGERI, Davide; FOURNET, Damien; HODDER, Simon; HAVENITH, George. Why wet feels wet? A neurophysiological model of human cutaneous wetness sensitivity. *J Neurophysiol.*, v. 112, Issue 6, p. 1457-1469, September 2014.

lema cunhado por *Feynman*, "o que não pode ser criado, não pode ser compreendido" (*What I cannot create, I do not understand*)[330].

Outra limitação do Argumento da Sala Chinesa está em considerar o cérebro humano como um mecanismo unitário, com um ponto de entrada de informações, um *locus* de processamento e um ponto de saída. Nesse caso, alega *Searle* que a mera troca de informações seria insuficiente para a demonstração da intencionalidade ou consciência.

Cabe observar, contudo, que os estudos atuais em neurociência demonstram um funcionamento significativamente menos mecanicista. Essa nova agenda de pesquisas se denomina "*Wet Mind*" e possui alguns princípios muito importantes: divisão de trabalho (*division of labor*); modularidade fraca (*weak modularity*); restrições de satisfação (*constraint satisfaction*); processamento concorrente (*concurrent processing*) e oportunismo (*opportunism*)[331]:

- a *divisão de trabalho* (*division of labor*) designa o fenômeno em que o cérebro divide determinada tarefa em grupos de neurônios, sendo que um apenas pode ser insuficiente para que uma função se torne significativa;
- a *modularidade fraca* (*weak modularity*) apresenta funções cerebrais amplas que não podem ser localizadas somente em um local do cérebro e são identificadas no cérebro como um todo;
- as *restrições de satisfação* (*constraint satisfaction*) demonstram que o cérebro é capaz de realizar tarefas simultâneas;
- o *processamento concorrente* (*concurrent processing*) aponta o fato de que as redes neurais funcionam em paralelo e de modo serial; e
- o *oportunismo* (*opportunism*) traz o fato de que o cérebro utiliza a informação disponível, mesmo que ela não seja diretamente aplicável ao caso.

Harris advoga que, apesar da singular complexidade do cérebro humano, é possível e mesmo necessário realizar a modelagem matemática deste, com novos mas sofisticados padrões de análise (*each function has been modeled in neural networks and many combinations of networks have been assembled to study overall large functions*)[332]. Tais estudos de neurociência

---

[330] EINEVOL, T.; DESTEXHE, Alain; DIESMANN, Markus *et al.* The Scientific Case for Brain Simulations. *Neuron.,* v 102, Issue 4, pp. 735-744, 22 May 2019.
[331] HARRIS, Paul. *Wet Mind, a New Cognitive Neuroscience and its Implications for Behavioral Optometry*. 2020. Disponível em: https://www.oepf.org/sites/default/files/referencearticles/WET_MIND_A_NEW_COGNITIVE_N.pdf. Acesso em: 05 jul. 2020 às 00:07.
[332] HARRIS, 2020, p. 10.

seriam muito importantes para a compreensão de lesões traumáticas no cérebro (*traumatic brain injury*) e seu efeito na visão, por exemplo.

Outros estudos demonstraram existir consciência em pessoas em estado vegetativo. Foi solicitada a uma mulher em estado vegetativo, no teste registrado por ressonância por imagem, a se imaginar jogando tênis. Nesse estudo se detectou a ativação das áreas do cérebro (*premotor córtex*) correspondentes à inicialização e imagem de movimentos[333]. Tais estudos indicam um novo caminho de pesquisas, alternativo ao uso de *respostas comportamentais* que não são plenamente confiáveis (*unreliable behavioral responses*).

Pode-se afirmar que o desenvolvimento da modelagem matemática, da compreensão neuromatemática do cérebro, dos avanços em instrumentos de análise por imagem demonstra a possibilidade de que os *modelos computacionais de consciência podem ser suficientes para se compreender o mecanismo da consciência* e assim poder-se-ia inverter o lema de *Feymann* ao afirmar "o que pode ser compreendido pode ser criado".

A possibilidade de criação de uma *consciência artificial* (*artificial consciousness*) foi objeto de diversos estudos e pioneiramente defendido por *Chalmers*[334]. O autor sumarizou as principais teses contra a existência de uma consciência artificial em duas dimensões distintas. A primeira tese foi denominada de "suficiência computacional" (*computational sufficiency*), que afirma que a possibilidade da correta engenharia computacional seria suficiente para a reprodução da consciência. A segunda tese é denominada de "explanação computacional" (*computational explanation*), que designa a possiblidade de a computação providenciar uma adequada compreensão dos estados cognitivos.

A primeira tese foi questionada, conforme *Chalmers*, por *Dreyfus* (1974) e *Penrose* (1989), que negaram a possibilidade de certas habilidades cognitivas serem duplicadas computacionalmente. De outro lado, mesmo que fosse possível sua duplicação, a sua instanciação não seria suficiente para apresentar a existência de uma mente computacional[335].

A segunda tese teria sido negada por *Edelman* (1989) e *Gibson* (1979), que questionaram a possibilidade de a computação fornecer um desenho adequado (*inappropriate framework*) para a explicação dos processos cognitivos. E, mesmo se pudessem ser explicados, a descrição restaria vazia (*vacuous*), conforme *Searle* (1990, 1991).

---

[333] OWEN, Adrian M. The Search for Consciousness. *NeuroView. Neuron*., v. 102, Issue 3, p. 526-528, 8 May 2019.
[334] CHALMERS, David. A Computational Foundation for the Study of Cognition. *Journal of Cognitive Science*, Seoul Republic of Korea, 2011, p. 323-357.
[335] SEARLE, 1980.

Chalmers irá defender que a tese de que a "explanação computacional" (*computational explanation*) nos permite uma linguagem adequada (*perfect language*) para a compreensão da organização causal dos processos cognitivos. Por sua vez, a tese da "suficiência computacional" (*computational sufficiency*) se sustenta, dado que todas as implementações computacionais conseguem replicar adequadamente a estrutura da mente (computational sufficiency holds because in all implementations of the appropriate computations, the causal structure of mentality is replicated)[336].

Comprovada a possibilidade teórica de *consciência artificial*, cabe questionar a possibilidade de *consciência moral em agentes artificiais* (*artificially conscious moral agents*). *W. Wallach, C. Allen* e *S. Franklin* irão solidamente defender que é possível a sua existência, como base nas seguintes proposições[337]:

1. a consciência é especialmente importante para decisões morais volitivas (*volitional moral decisions*);
2. cognição moral é suportada (*supported*) por processos cognitivos gerais; e
3. A capacidade de tomar decisões morais é um atributo essencial dos agentes conscientes.

Os autores chegam à conclusão de que um agente artificial completo deve ser também um agente artificial com consciência moral (*artificial conscious moral agent*).

A *capacidade de tomar decisões morais* (*volitional decision making*) é um processo cognitivo de ordem superior para a escolha de ações a serem tomadas. Trata-se de uma categoria muito distinta de outras formas de seleções de ações, tais como: escolhas mediadas, automatizadas ou mera execução de comandos[338]. A capacidade de tomar decisões de modo autônomo com base na vontade artificial é um tema que tem merecido uma recente atenção da literatura. Há o entendimento da possibilidade de escolhas morais artificiais, com estatura similar às escolhas humanas.

As consequências sociais e, mesmo, existenciais para humanidade, dessa radical possibilidade são desafiadoras. Por outro lado, cabe destacar que escolher o curso de ação moral a ser tomado não exige somente escolhas racionais, mas igualmente emocionais, o que nos remete a outro questionamento: seria possível as máquinas possuírem emoções artificiais?

---

[336] CHALMERS, 2011, p. 354.
[337] WALLACH, Wendell; ALLEN, Colin; FRANKLIN, Stan. Consciousness and ethics: artificially conscious moral agents. *International Journal of Machine Consciousness*, v. 03, n. 01, p. 177-192, 2011, p. 189-190.
[338] WALLACH, Wendell; FRANKLIN, Stan; ALLEN, Colin. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, v. 2, p. 454-485, 2010, p. 469.

2.1.4 Da objeção biológica e das incapacidades: ausência de emoções

Turing irá destacar as *incapacidades* (*disabilities*) das máquinas, a sua natural e intrínseca limitada condição. Fatos triviais e frívolos estariam distantes da mais potente das máquinas atuais. O mero apreciar de sorvete, talvez de um canto de pássaros ou um suave brisa primaveril. Essa incapacidade, talvez fútil, conduziria a incapacidades de maior nível, tal como a inaptidão para a empatia com outros seres humanos ou entre máquinas e seres humanos[339]. Uma das mais importantes capacidades humanas é a de possuir sentimentos morais. Um autêntico agente moral deveria ter a capacidade de sentir.

A capacidade de sentir é uma das linhas demarcatórias que separam os objetos dos seres vivos. E possuir sentimentos é o que caracteriza os seres *sencientes*. Esse entendimento remontava aos escolásticos, que desde *Santo Agostinho*, diferenciavam os seres conforme a capacidade de sentir e de raciocinar. Dizia o autor: "e, entre os viventes, os sencientes são superiores aos não-sencientes, como às árvores os animais. Entre os sencientes, os que têm inteligência são superiores aos que não a têm, como aos animais os homens"[340]. Haveria apenas uma diferença de grau entre seres inteligentes e seres sencientes, mas seríamos todos portadores da mesma condição de possuir sentimentos.

Bentham irá retomar modernamente o tema, ao tratar do conceito de agente e afirmar que a linha demarcatória para o reconhecimento moral de um agente não deveria ser dada pela razão, mas pela capacidade de sentir, de sofrer.

O autor irá erigir o *princípio da igual capacidade de consideração* (*principle of equal consideration*) com base na teoria dos sentimentos morais. Afirmava o autor: "o que mais deve delimitar a linha insuperável? É a faculdade da razão, ou talvez a faculdade do discurso? [...] a questão não é: eles podem raciocinar? nem, eles podem falar? mas eles podem sofrer?"[341]. É a capacidade de sofrer e não a capacidade falar ou de raciocinar que estabelece a linha básica para

---

[339] "*There are, however, special remarks to be made about many of the disabilities that have been mentioned. The inability to enjoy strawberries and cream may have struck the reader as frivolous. Possibly a machine might be made to enjoy this delicious dish, but any attempt to make one do so would be idiotic. What is important about this disability is that it contributes to some of the other disabilities, e.g., to the difficulty of the same kind of friendliness occurring between man and machine as between white man and white man, or between black man and black man*". TURING, 1950, p. 444.

[340] AGOSTINHO *apud* BRANDÃO, Ricardo Evangelista; COSTA, Marcos Roberto Nunes. Agostinismo político: a apropriação dos textos agostinianos no *De ecclesiastica potestate*. *Perspectiva Filosófica*, v. 2, n. 40, p. 111, 2013.

[341] Cf. "*What else is it that should trace the insuperable line? Is it the faculty of reason, or perhaps, the faculty for discourse? [...] the question is not, can they reason? nor, Can they talk? but, Can they suffer*?". BENTHAM, [1781], 2000.

a consideração moral de ser. O *atormentamento* com o sofrimento alheio é uma demanda moral, que afasta qualquer justificativa para o desprezo com a trágica condição de outrem.

Assim, se a *senciência* é a capacidade de sentir sofrimento ou prazer, então poderiam surgir considerações sobre a possibilidade de os agentes artificiais sofrerem[342]. Existiria a possibilidade de agentes morais artificiais serem considerados sencientes? E, nesse caso, as máquinas inteligentes *sencientes* deveriam ser consideradas como seres protegidos pelo princípio da igual consideração? Qual seria o seu estatuto jurídico? Seriam pessoas? Essas e tantas outras questões serão fundamentais no futuro. O que nos interessa no presente trabalho, contudo, é a possibilidade de os agentes morais artificiais possuírem emoções.

Partindo da premissa de que há similaridade entre a moralidade humana e a artificial, então poderíamos questionar se há similaridade entre a agência moral humana e a artificial. Se existir tal similaridade, seria possível que um agente moral artificial completo (*full AMA*) possa ter a presença de sentimentos morais artificiais?

A primeira tese é negativa e afirma que não existe conexão entre racionalidade e emoções. Aquela poderia ser modelada computacionalmente, esta não. Poderíamos formalizar raciocínios, a cognição e mesmo deliberações morais racionais. As emoções seriam inefáveis e impossíveis de serem modeladas.

Aaron Sloman entende de modo distinto. Não é correto o entendimento de que cognição e emoções sejam completamente distintas[343]. Se isso for correto, então a computação possui o grande desafio de modelar algo tão difuso quanto os sentimentos. Nota-se de imediato a dimensão assombrosa dessa tarefa. Afinal, a própria humanidade possui uma dificuldade gigantesca em tratar dos sentimentos. Muitas vezes a racionalidade não consegue abarcar toda a carga de sentido de uma emoção. Somente a literatura consegue expressar de modo pleno o profundo e inefável sentido das dores e alegrias humanas. Mesmo assim, nos questionamos sobre como modelar algo tão inexprimível quanto um sentimento?

Os *sentimentos artificiais*, para a modelagem computacional, são considerados como motivadores primitivos para a seleção de ações (*primitive motivators*)[344]. Eles são representados

---

[342] HÅKANSSON, Simon. *The Chinese Room and Turing's Wager*: moral status in the age of artificial intelligence. 2016. Disponível em: https://www.researchgate.net/profile/Simon_Hakansson/publication/309634694_The_Chinese_Room_and_Turing%27s_Wager_Moral_Status_in_the_Age_of_Artificial_Intelligence/links/581aecf308ae30a2c01d53b5/The-Chinese-Room-and-Turings-Wager-Moral-Status-in-the-Age-of-Artificial-Intelligence.pdf?origin=publication_detail. Acesso em: 05 jul. 2020 às 15:16.

[343] SLOMAN, Aaron. Why robots will have emotions. Proceedings IJCAI. *Cognitive Science Research Paper,* Sussex University Vancouver, v. 176, p. 1, 1981.

[344] WALLACH; FRANKLIN; ALLEN, 2010, p. 466.

como nós (*nodes*) de memória perceptual, em que cada nó representa sua própria valência, positiva ou negativa, e segundo uma determinada graduação.

As *emoções* seriam tratadas como sentimentos com conteúdo cognitivo (*feelings with cognitive content*), tal como a alegria de encontrar um amigo ou o embaraço de cometer uma gafe[345]. *Johnston* apresenta ainda os sentimentos como parte de um mecanismo de premiações e desincentivos a determinadas escolhas[346]. Eles auxiliariam na seleção de ações ao indicarem quais os incentivos e desincentivos de determinada escolha.

Apesar desses importantes esforços para a modelagem dos sentimentos, ainda permanece a pergunta: poderiam os *agentes artificiais morais* apresentarem algo tão inefável e singular quanto uma emoção? Ou seria apenas um mero "jogo de imitação emocional"?

A tese restritiva afirma que não. Os sistemas inteligentes poderiam apenas ser capazes de compreender ou racionalizar sobre as emoções, mas não necessariamente possuí-las (*AI systems must be able to reason about emotions*)[347].

Esse argumento é denominado por *Turing* de objeção biológica ou de *continuidade do sistema nervoso*. As máquinas são sistemas de estados discretos, já o cérebro humano não. Pequenas ou ínfimas variações de sinais nervosos podem provocar resultados díspares e consideráveis. *Turing* irá responder que máquinas poderiam computar valores não discretos e predizer a resposta de máquinas de análise diferencial (*differential analyzer*)[348]. Tudo seria uma questão de computabilidade que poderia ser superada.

O desafio, e mesmo a possibilidade filosófica de os agentes artificiais possuírem emoções, é algo muito mais complexo e desafiador. Nem se trata do questionamento sobre se eles deveriam ter emoções ou se estas deveriam ser limitadas, controladas ou ajustadas ao convívio com humanos, mas algo muito mais profundo e radical: poderiam os agentes artificiais possuírem autênticas emoções?

Esse questionamento é fundamental porque, para que existam Agentes Morais Artificiais completos, eles deverão ser dotados de razão e emoção. Ao se depararem com conflitos éticos, necessitarão estar armados mais do que com apenas a razão. Deverão utilizar o bom senso, o senso de justiça, a empatia e tantos outros elementos puramente subjetivos.

Vejamos o que é necessário para que existam autênticas emoções artificiais, não apenas de sua simulação artificial, sua modelagem racional ou computacional. Para que possam ter

---

[345] WALLACH; FRANKLIN; ALLEN, 2010, p. 466; e JOHNSTON, 1999.
[346] JOHNSTON, 1999, p. 17.
[347] PICARD, Rosalind. *Affective computing*. Cambridge, MA: MIT Press, 1997. p. 195.
[348] TURING, 1950, p. 445.

emoções é necessária a capacidade de sintetizar e gerar emoções. Para *Rosalind Picard*, as emoções possuem cinco componentes descritivos[349]:

1. comportamento emocional (*emotional behavior*);
2. emoções primárias rápidas (*fast primary emotions*);
3. emoções geradas cognitivamente (*cognitively generated emotions*);
4. experiência emocional (*emotional experience: cognitive awareness, physiological awareness, and subjective feelings*);
5. interações mente-corpo (*body-mind interactions*).

A autora afirma que nem todos os elementos serão necessários ao mesmo tempo, sendo que nem todos os animais os possuem. Essa observação não afasta, contudo, a complexa e difícil tarefa de entendimento sobre os sentimentos, tampouco sobre os elementos necessários à sintetização artificial das emoções.

Tome-se por exemplo a formalização do sentimento de *alegria* (*joy*) proposto por *Ortony*, *Clore* e *Collins* na sua obra "The Cognitive Structure of Emotions" (1988). Para os autores, essa fórmula teria como elementos a desejabilidade de um evento, por uma dada pessoa, em um tempo *t*. Essa função retornaria valores positivos, se o evento esperado tivesse consequências benéficas, e negativos em caso contrário[350]. Nesse modelo, a regra ativa a emoção de alegria, a partir do momento em que o limite de intensidade zero é superado. As emoções seriam consideradas resultados de situações que incluem eventos, objetos e agentes[351].

Existe ainda a dúvida se toda a carga, complexa e multifacetada, do sentimento primário de alegria pode ser reduzida em fórmulas. Talvez ela expressasse no máximo um sentimento de satisfação, mas não exatamente de alegria.

Herbert Simon será um dos primeiros a tratar das emoções artificiais e contraditar a tese de *Ulric Neisser*[352]. Este último autor afirmava que as máquinas somente poderia possuir uma

---

[349] PICARD, 1997, op. cit., p. 193.

[350] "*Then an example rule for joy is:*
IF D (p, e, t) 0
THEN set Pi (p, e, t) = fi (D(p, e, t), Ig(p, e, t))
where fi() is a function specific to joy" (PICARD, 1997, p. 195).

[351] PICARD, 1997, p. 198.

[352] SIMON, H. A. A Theory of Emotional Behavior. Carnegie Mellon University Complex Information Processing (CIP) Working Paper #55, June 1, 1963. Disponível em http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=346072. Acesso dia 27.12.2020 às 00:31. A crítica de *Neisser* era fundamentada em três pontos: "*Three fundamental and interrelated characteristics of human thought . , . are conspicuously absent from existing or contemplated computer programs: 1) human thinking always takes place in, and contributes to, a cumulative process of growth aftd development; - 2) human thinking begins in an intimate {association with emotions and feeling* which is never-entirely lost; 3) almost alt human activity, includinthinking, serves not one but a multiplicity of motives at the same time*"; p. 02.

"*cognição fria*" (*cold cognition*), mas jamais uma "*cognição quente*". A primeira seria própria da racionalidade, do raciocínio e decisão e a segunda relacionada às emoções e sentimentos. Simon irá explicar a teoria das emoções por meio de uma teoria do processamento de informações (*information processing behavior*)[353].

Outro ponto importante é se a geração de emoções possui estreita relação com o controle delas. Para a autora, se o agente artificial não puder controlar suas emoções, talvez não seja capaz de sintetizá-las de modo apropriado. Como saberia gerar emoções se não for capaz de as reconhecer de modo cuidadoso, de expressá-las de modo fluente? A tese subjacente da autora é a de que, para gerar emoções, o agente artificial deveria primeiro possuí-las. A sintetização de emoções exige *inteligência emocional*.

O aprendizado das emoções exigiria, além da inteligência emocional, uma *inteligência social*, sobre como agir e se comportar em interações sociais[354]. Os agentes morais artificiais devem não apenas "ter" emoções (gerar ou sintetizar), saber reconhecê-las, controlá-las e expressá-las de modo competente, mas, igualmente, se comportar emocionalmente em relacionamento com outras pessoas. Naturalmente, o contato com os seres humanos irá gerar uma gama bastante distinta de reações e sentimentos (temor, alegria, esperança, medo, etc.). Como agentes artificiais emocionais lidariam com situações como aversão, ódio ou repugna?

Para *Azeem et alii*, a conclusão a que se chega é que as emoções são diretivas para as decisões humanas, que se originam em decorrência de interação com o ambiente, com os outros e em virtude de estados mentais internos da memória. São as emoções que tornam os seres humanos únicos e autônomos na sua capacidade decisória (*in fact, emotions make human beings* "*autonomous*" *in their decision-making*[355]).

As emoções podem ser consideradas como diretrizes, conforme *Nico Frijda*, para auxiliar os seres humanos a superarem preocupações (*concerns*) ou temores. Dentre as várias preocupações listadas, poderíamos citar a sobrevivência ou a segurança, dentre outros exemplos[356]. Talvez os agentes morais artificiais nunca desenvolvam emoções complexas como

---

[353] SIMON, 1963, p. 27.
[354] PICARD, 1997, p. 194.
[355] AZEEM, M. M. *et al.* Emotions in Robots. *In*: CHOWDHRY B.S. *et al.* (ed.). *Emerging trends and applications in information communication technologies*. Communications in Computer and Information Science. Berlin: Springer, 2012. v. 281.
[356] FRIJDA, Nico. (2016). The evolutionary emergence of what we call "emotions". *Cognition & Emotion*, v. 30, p. 1-12, 2010. Disponível em: https://www.researchgate.net/publication/297583225_The_evolutionary_emergence_of_what_we_call_emotions/citation/download. Acesso em: 16 dez. 2020 às 23:16.

as humanas[357], talvez desenvolvam[358]. Afinal, não há um argumento definitivo que demonstre indubitavelmente a existência de um impeditivo ontológico para afirmar-se o contrário[359].

Considerando que existe similaridade entre compreender emoções e poder senti-las, é possível afirmar que elas possam se desenvolver. Afinal, seres menos complexos podem igualmente sentir o sofrimento. A igualdade em sofrer permitiria a possibilidade do surgimento de autômatos racionais, morais e emocionais[360]. Agora, se isso é desejável, bem, se trata de outro problema.

A possibilidade de a excepcionalidade humana ser superada pela tecnologia é algo real. Dos argumentos apresentados, não há como afirmar ou negar *a priori* sobre a impossibilidade lógica ou ontológica do surgimento de agentes artificiais morais dotados de emoções.

Alguns autores afirmarão categoricamente, como *Sloman*, que os autômatos possuirão emoções e estas poderão ser modeladas artificialmente[361]. Se esta tese for verdadeira, de que possuirão emoções e racionalidade prática, então que sejam virtuosos, que procurem o bem e se afastem do mal.

2.1.5 Da objeção teológica

O último argumento contra a possibilidade filosófica da existência de agentes morais artificiais completos (*full AMA*) é o de que eles não possuem ou jamais possuirão uma alma imortal. Mesmo que a tecnologia alcance níveis inimagináveis para os padrões atuais, jamais terão essa propriedade absolutamente exclusiva aos humanos.

Não há nenhuma prova, contudo, de que a possibilidade de avanço tecnológico seja limitada, a ponto de impedir o surgimento de máquinas extremamente avançadas.

Turing irá chamar essa oposição à ideia de máquinas inteligentes de "objeção teológica" (*the theological objection*). O autor não pretende apresentar uma objeção externa a esse argumento, afirmando, por exemplo, que Deus não existe. Pelo contrário, irá assumir o desafio

---

[357] Contra moral machines veja-se WYNSBERGHE, Aimee van; ROBBINS, Scott. Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, v. 25, n. 3, p. 719-35, 2018; MOSAKAS, Kestutis, 2020, p. 33-48.

[358] SLOMAN, Aaron. *What Are Emotion Theories About?* Disponível em: https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aaai04-emotions.pdf. Acesso em: 17 dez. 2020.

[359] PICARD, R.W.; VYZAS, E.; HEALEY, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, n. 10, p. 1175-1191, 2001.

[360] VELASQUEZ, J. A computational framework for emotion-based control. *In*: PROCEEDINGS OF THE WORKSHOP ON GROUNDING EMOTIONS IN ADAPTIVE SYSTEMS, INTERNATIONAL CONFERENCE ON SAB, University of Zurich, Switzerland August 21, 1998.

[361] SLOMAN, Aaron; CROUCHER, Monica. You don't need a soft skin to have a warm heart: towards a computational analysis of motives and emotions. *Cognitive Science Research Paper*, Sussex University, p. 1, 1981.

de refutar essa objeção com argumentos teológicos, partindo da assunção de que Deus existe[362]. Inicia a sua exposição afirmando que o fará assim, apesar de não aceitar nenhum desses pressupostos[363].

A capacidade de pensar seria própria de um indivíduo detentor de uma alma imortal. Deus teria dado uma alma imortal para cada homem e mulher na Terra, mas não para os animais e as máquinas. Assim, nenhum animal ou máquina poderiam pensar[364].

Considere-se que os atos criativos de Deus são de duas espécies: materiais e imateriais. A tecnologia somente poderia criar coisas materiais, mas estariam reservados a Deus os componentes imateriais da criação. A tecnologia jamais poderia criar agentes morais artificiais completos. Ela no máximo nos legaria imitações externas de seres com alma.

Russell C. Bjork irá tratar da possiblidade de conflito teológico entre a inteligência artificial e o surgimento da alma humana. Afinal, se os seres humanos foram criados à imagem e semelhança de Deus (*Imago Dei*), haveria uma excepcionalidade humana a excluir a criação de seres artificiais com alma?[365].

O autor irá listar três questões, fundamentais e provocativas, sobre os problemas teológicos confrontados pela IA, como[366]:

> 1. Existe um conflito entre a Inteligência Artificial e a doutrina bíblica sobre a origem da alma humana?
>
> 2. Existe um conflito entre a IA e o ensino bíblico sobre o valor humano e a nossa criação em à imagem de Deus?
>
> 3. O ensino bíblico sobre personalidade tem implicações para o nosso trabalho em IA?

Russell C. Bjork irá apresentar uma leitura ligeiramente diferente daquela exposta por *Turing*. Poder-se-ia dizer que os seres humanos não "recebem" uma alma imortal de Deus, mas eles se "tornam" uma alma imortal.

Seria um erro pressupor que Deus "concederia" uma alma a qualquer coisa ou objeto, tal como um pedaço de madeira ou pedra. De igual modo, a alma não seria recebida por um

---

[362] BRINGSJ, Selmer. God, souls, and Turing: in defense of the theological objection to the Turing test. *Kybernetes*, v. 39, n. 03, p. 414-422, 2010, p. 417. Disponível em: http://kryten.mm.rpi.edu/SB_theo_obj_tt_offprint.pdf. Acesso em: 18 jul. 2020 às 20:41.

[363] Cf. "[...] *unable to accept any part of this but will attempt to reply in theological terms*". TURING, 1950, p. 433-460.

[364] "*Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think*". TURING, 1950, p. 433-460.

[365] BJORK, Russell C. Artificial Intelligence and the Soul. *Perspectives on Science and Christian Faith*, v. 60, n. 2, p. 95-102, June 2008, p. 98.

[366] SCHUURMAN, D. C. *Artificial Intelligence*: discerning a Christian response. Ontario/Canada: Canadian Scientific & Christian Affiliation, 2020. Disponível em: https://www.csca.ca/uploads/18Jan20SchuurmanDiscerningAI.pdf. Acesso em: 18 jul. 2020 às 21:12.

conjunto mecânico qualquer. A *relação especial entre corpo e alma* tem sido reafirmada a partir de *Santo Agostino*, *Santo Tomás de Aquino*, *Descartes* ou *Leibniz*. Esse é um dos pontos teológicos mais importantes na cristandade, esquecido na objeção teológica de *Turing*.

O aspecto imaterial dos humanos (personalidade) "emerge" da interação de sua condição biológica, dos neurônios no cérebro[367]. As propriedades mentais seriam emergentes quando localizadas em condições biológicas altamente complexas, não surgindo em formas de vida mais simples. Essa abordagem seria teologicamente coerente.

Poder-se-ia alegar, contra o argumento teológico da singularidade do ser humano na Criação, que a ciência demonstrou os limites de nossa pequenez. *Copérnico* provou que nosso local físico no Universo não é especial. *Darwin* provou que nossas evolução não teria sido especial, mas seguiu rumos naturais. A prova de que os animais são sencientes e inteligentes desbancou o mito de nossa excepcionalidade também nesse campo. Os avanços da inteligência artificial não trariam um ataque inédito à tese da exclusividade humana no campo da Criação[368].

O que nos faria especiais não seria a nossa constituição singular e única, mas o nosso propósito e o nosso relacionamento especial com Deus[369].

A criação de um agente artificial inteligente seria o equivalente à criação de um ser à imagem de Deus? Seria essa afirmação herética? Para *Bjork*, não há uma implicação necessária na afirmação de que ser racional é ser criado à imagem divina. Outros problemas irão surgir, tais como a noção de redenção ou revelação. O autor afirma que os seres humanos não deveriam se sentir ameaçados pela emergência de uma inteligência artificial, mesmo que ela seja dotada de um espaço na Criação e divida o domínio da Terra. Pelo contrário, nossa salvação dependeria ainda mais Dele para a realização de nossos valores e propósitos finais[370].

Bjork ainda se questiona o papel dos teístas perante o desafio da inteligência artificial. Deveriam se limitar às pesquisas em IA fraca ou se aproximar da IA forte? A sua resposta é não, com base em argumentos teológicos. Avançar os estudos em IA forte não entraria em conflito com a Teologia. Afinal, seria mais uma forma de contemplar as maravilhas e os mistérios da Criação[371].

---

[367] BJORK, 2008, p. 97.
[368] BJORK, 2008, p. 99.
[369] BJORK, 2008, p. 99.
[370] BJORK, 2008, p. 100.
[371] BJORK, 2008, p. 101.

Os estudos em inteligência artificial teriam fundamento inclusive nos estudos teológicos de Doutores da Igreja, como *São Tomás de Aquino*. *Walter Freeman*[372] irá afirmar que a compatibilidade da doutrina se dá pela centralidade de dois conceitos principais: *intencionalidade* e *imaginação*. A intencionalidade ocorre pela unidade entre mente e corpo, em ação no mundo. A imaginação pelo apelo à criação de cada indivíduo, por meio de escolhas construtivas (*constructive choice*).

Para *Walter Freeman*, a compatibilidade desce às minúcias, especialmente, quando trata da percepção ativa até a intelecção. O autor demonstra, em um quadro ilustrativo, a coerência entre os conceitos filosóficos em *Santo Tomás de Aquino* e a neurociência computacional: (i) *sensatio* e percepção sensorial (*sensory perception*); (ii) *phantasmata* e mecanismos neurais em grupo (*hebbian nerve cell assembly*[373]); (iii) *abstractio* e cortéx sensorial e reconhecimento de padrões (*sensory cortex* e *AM patterns*); (iv) *sensus communis* e sistema límbico (*limbic system*); (v) *imaginatio* e ondas cerebrais (*wave packet e neocortex*) e, por fim, (v) *intellectio* e a cognição simbólica (*symbolic cognition*).

Diversos estudos em neurociência computacional demonstraram a capacidade de formalização em modelos matemáticos do funcionamento do cérebro e da intelecção[374]. Não há um fosso ou inconsistências profundas entre conceitos filosóficos consagrados na escolástica, capazes de impedir *prima facie* uma agenda de pesquisas comum entre IA e teologia.

Richard Swinburne irá adotar uma estratégia diferente para investigar os limites da ciência em explicar a alma humana. Em vez de se utilizar de argumentos dedutivistas, irá preferir uma abordagem probabilística, ou seja, argumentos que partem de dadas evidências para confirmar a sua veracidade[375]. No lugar de defender o monismo da unidade entre mente e corpo, irá defender um dualismo substancial e, ao mesmo tempo, tentar comprovar a unidade entre uma dada alma e um corpo[376].

Para o autor a evidência da existência de memórias conectadas com a consciência de eventos anteriores no mesmo cérebro seria uma primeira evidência de que uma mente e um cérebro são da mesma pessoa. De outro lado, a continuidade de crenças e desejos (*beliefs and*

---

[372] FREEMAN, Walter J. *Nonlinear Brain Dynamics and Intention According to Aquinas*, Seattle, AI2, p. 232. Disponível em: https://www.semanticscholar.org/paper/Nonlinear-Brain-Dynamics-and-Intention-According-to-Freeman/1058e99a0036f6f9b9a76a9b7dc59e6b16cf736a. Acesso em: 31 jul. 2020 às 00:51.

[373] GERSTNER, Wulfram. *Hebbian learning and plasticity*. AI2, Seattle. Disponível em: https://pdfs.semanticscholar.org/f9fc/99a5c52aa5df1b530dfdeb25dfb6b10bdecf.pdf. Acesso em: 31 jul. 2020 às 00:52.

[374] MCCULLOCH, Warren; STURGIS; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, v. 52, p. 99-115, 1990.

[375] SWINBURNE, Richard. *Are we bodies or souls*? New York: Oxford University Press, USA, 2019.

[376] Cf. "*I argued at the beginning of this chapter that only substance dualism is thus compatible*". SWINBURNE, 2019, p. 170.

desires) dessa mesma pessoa poderiam fundamentar que são a mesma pessoa[377]. Pensar que o mesmo corpo de conecta com diversas mentes ao mesmo tempo ou em diferentes momentos trairia as evidências anteriores.

Alguns elementos raros podem contrariar a afirmação acima, conforme *Swinburne*, nos casos de *desordens de múltipla personalidade* (*multiple personality disorder*) e na *divisão de cérebros* (*split brains*), por meio de procedimento cirúrgico (*corpus callosum*)[378]. Poder-se-ia questionar se, no caso da divisão do cérebro, estaríamos perante dois cérebros, duas consciências e duas almas. Teriam, as duas almas, os mesmos eventos mentais após essa separação? Para o autor, as evidências sugerem o inverso, que não existem dois sujeitos (diríamos duas intencionalidades) disputando o controle do mesmo corpo. Assim, para o autor, é preferível a explicação mais simples[379] de que a divisão de cérebro não conduz a tese das "duas almas" em um corpo[380].

A ideia de conexão entre corpo, mente e alma explicaria outro fenômeno importante, como o inusitado experimento denominado de "*download* de um cérebro". Para o autor, caso fosse possível realizar o registro computacional de todos os eventos mentais de um dado indivíduo para outro cérebro, indivíduo ou simulacro, seria extremamente improvável que esse procedimento preservaria a alma original[381].

Dificilmente será possível formular uma prova científica da alma imortal ou uma teoria científica completa desta, o que não impede a sua consideração em termos científicos, conforme *Swinburne*[382]. Tampouco, impede o debate teológico sobre o futuro da ciência.

Assim, seja por uma abordagem teológica dedutivista ou indutivista, monista ou dualista da alma, não há como negar que pode ocorrer o fenômeno da *emergência* da consciência em determinado corpo artificial. O que parece estar descartada é a possibilidade de se *enxertar* uma consciência e uma alma em um corpo natural ou artificial.

Afastando-se a objeção teológica, demonstra-se não descartada ou refutada a possibilidade da existência de agentes morais artificiais completos (*full* AMAs). Sendo assim,

---

[377] Cf. "[...] *the existence of very many later a-memories of one's own earlier conscious events connected to the same (brain and so) body, and the continuity of the beliefs and desires of the later person with those of the earlier person with the same (brain and so) body* [...]". SWINBURNE, 2019.
[378] SWINBURNE, 2019, p. 149-150.
[379] SWINBURNE, Richard. *Simplicity as evidence of truth (Aquinas lecture).* Milwaukee: Marquette University Press, 1997.
[380] Cf. "*and that might well lead us to prefer one of the interpretations of the split-brain cases which does not lead to the 'two souls' view*". SWINBURNE, 2019, p. 152.
[381] Cf. "*But it is extremely improbable that 'downloading' a person's brain onto another brain or other system (for example, by 'tele transporting' it into the brain of some person on another planet) would preserve the original person and so their soul*". SWINBURNE, 2019, p. 153.
[382] SWINBURNE, 2019, p. 161.

poderíamos questionar a possibilidade de modelagem da agência moral em algoritmos computacionais. Talvez essa seja a última importante objeção.

Uma objeção teológica distinta parte da teoria do "desenho inteligente" (*intelligent design*). A teoria do *intelligent design* e da inteligência artificial possuem algo em comum: a inteligência. O *intelligent design* se dirige a explicar as informações existentes na natureza, para além da aleatoriedade. A inteligência artificial pretende mimetizar a inteligência humana[383].

Marks et alli pretendem demonstrar a impossibilidade lógica de reprodução da criatividade humana em algoritmos. Haveria uma limitação intrínseca na criatividade computacional (*computer creativity*), devido a duas leis naturais: a lei da conservação de informações (*law of conservation of information*) e a teoria da informação algorítmica (*Algorithmic Information Theory* – AIT). Essas seriam as limitações absolutas aos modelos de criatividade e inteligência artificial. Haveria algo de inexplicável, inatingível ou inefável na natureza da mente humana que jamais seria possível de ser capturada ou computável por meio de algoritmos[384].

Os processos evolucionários não podem criar informação, para o autores, o que afastaria a possibilidade de infundir criatividade em uma máquina inteligente (*infused into the program by the computer programmer*)[385]. Essa limitação seria denominada, por eles, de *Lovelace Test*, em homenagem à *Lady Lovelace*, que afirmava: "Os computadores não podem criar nada. Pois a criação requer, minimamente, algo originário. Mas os computadores não originam nada; eles simplesmente fazem o que nós ordenamos que eles, por meio de programas, façam"[386].

O *Teste de Lovelace* poderia explicado da seguinte forma: a IA forte será demonstrada quando a criatividade artificial *(machine's creativity)* está além da explanação de seu criador[387]. E o lampejo do gênio criativo (*flash of creative genius*) não seria computável ou mimetizável[388].

A objeção teológica dos autores possui três partes. A primeira ataca diretamente a possibilidade da teoria da evolução, o que foge ao propósito do presente trabalho. A segunda parte nega a possibilidade de emergência de inteligência de modo natural, o que não encontra fundamento teológico unânime. Por último, nega a possibilidade de superação do *Teste de*

---

[383] MARKS, Robert; DEMBSKI, William; EWERT, Winston. *Introduction to Evolutionary Informatics*. Hackensack: World Scientific, 2017, p. 281.
[384] MARKS; DEMBSKI; EWERT, 2017, p. 281.
[385] MARKS; DEMBSKI; EWERT, 2017, p. 282.
[386] Cf. "*Computers can't create anything. For creation requires, minimally, originating something. But computers originate nothing; they merely do that which we order them, via programs, to do*". LOVELACE, *op. cit.*
[387] MARKS; DEMBSKI; EWERT, 2017, p. 284.
[388] MARKS; DEMBSKI; EWERT, 2017, p. 288.

Lovelace, de que um algoritmo possa permitir a *emergência* de uma inteligência artificial forte. Sobre este último aspecto, iremos nos deter a seguir.

2.1.6 Requisitos para uma ética artificial virtuosa

Se for possível a existência de agentes morais artificiais, então é muito provável que ajam, moral e racionalmente, conforme um modelo de virtudes artificiais (*artificial virtue*). Não parece ser factível pensar que devam seguir um conjunto de regras morais altamente abstratas.

Esse tem sido o entendimento de diversos autores, tais como *Berberich* e *Diepold* (2018); *Howard* e *Muntean* (2016); *Howard* e *Muntean* (2017) e *Govindarajulu* (2019). Todos eles pensaram em modos de modelagem de ética das virtudes para agentes morais artificiais[389]. Um dos caminhos mais promissores para essa difícil tarefa está no uso de aprendizado de máquina (*machine learning*) e algoritmos evolucionários (*evolutionary algorithms*).

Nem *Kant*, nem *Bentham* têm sido utilizados como referencial teórico para essa revolucionária tarefa, mas justamente a tradição aristotélica tem sido redescoberta pelos inovadores em engenharia da inteligência artificial. Não existe, contudo, uma única abordagem filosófica no campo das virtudes artificiais.

Outras tradições têm sido igualmente exploradas, tais como confucionismo, taoísmo e budismo[390]. Apesar de não terem sido citados, os estudos em filosofia escolástica medieval parecem ser muito promissores para a compreensão profunda da ética das virtudes.

Os autores parecem não adotar referencial filosófico em sua integralidade, restringindo-se a adotar os princípios gerais aplicáveis à compreensão do tema. O ponto principal da abordagem pela ética das virtudes está na análise do caráter do agente moral em ação. Não se trata de uma abordagem centrada na avaliação moral das deliberações do agente, tal como na deontologia; nem de uma consideração dos estados de coisas resultantes das escolhas morais, tal como no consequencialismo.

O modelo teleologicamente estruturado da ética das virtudes em Aristóteles parece ter recebido o atento interesse dos estudiosos em inteligência artificial. A preocupação com a ação do sujeito moral e a sua busca pela excelência no caráter e nas ações determinaram uma

---

[389] GAMEZ, P. *et al.* Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY*, Springer, 2020.

[390] KEOWN, D. *Buddhist ethics*: a very short introduction. New York: Oxford University Press, 2005; VALLOR, S. *Technology and the virtues*: a philosophical guide to a future worth wanting. Oxford: Oxford University Press, 2016; e SIM, M. Confucian and daoist virtue ethics. *In*: CARR, D.; ARTHUR, J.; KRISTJÁNSSON, K. (ed.). *Varieties of virtue ethics*. London: Palgrave Macmillan, 2017. p 105-121.

modelagem ao mesmo tempo flexível ao ambiente e rigorosa na escolha das deliberações morais.

Há a compreensão de que agentes humanos e artificiais atuam de modo absolutamente distinto em situações normais ou críticas. Geralmente, há a percepção de que os humanos se apresentam mais virtuosos do que as máquinas, mas, igualmente, mais viciosos[391]. Talvez, essa diferença se explique pelo fato de que as virtudes possuam uma base racional, mas não se limitam a esse aspecto. Para *Kraut*, a ética das virtudes é incodificável ou, melhor dizendo, não é totalmente modelável. Sendo incapaz de permitir a elaboração de um procedimento decisório unicamente racional e abstrato. Esse fato, por si só, já afasta a pretensão de um modelo teórico estilo "de baixo para cima" (*top-down*) para programação da agência moral artificial[392].

A virtude moral é aprendida, exercitada, observada e estudada por meio da ação continuada. O agente moral realiza suas deliberações observando exemplos morais e aplicando em situações similares, em um processo de treino e correção. Esse modelo teórico se ajustou admiravelmente, bem como os avanços em aprendizado de máquina e nos algoritmos evolucionários.

Novamente, o dilema sobre a possibilidade de autênticos agentes morais artificiais irá ressurgir. A ação genuinamente virtuosa, como expressão do caráter, é obviamente racional, mas de uma racionalidade prática, distinta dos modelos de racionalidade teorética.

Até que ponto estaremos perante uma agência aparente ou autêntica (*full agency*)? A presença de agentes morais artificiais que atuam realizando escolhas éticas no dia a dia obscurecem a clara distinção entre agentes éticos explícitos e implícitos. Conforme *Gamez*, a distinção entre agentes e pacientes morais torna-se muito tênue. O critério da capacidade de sofrer como relevante à aplicação do princípio da mesma consideração (*equal moral consideration*) longamente utilizado para crianças e animais poderia ser estendido para outros seres não humanos, tais como os agentes morais artificiais? Deveriam possuir direitos e exigir-se-iam deveres humanos para com eles?

Gamez et alii compararam o desempenho virtuoso de humanos e de AMAs em cinco diferente domínios: *verdade, justiça, riqueza, medo e honra*. De modo geral, não há grande distinções entre o comportamento de humanos e os de agentes artificiais, conforme a simulação realizada pelos autores.

Os estudos sobre IA mudaram o foco da modelagem de regras morais gerais e abstratas, aplicáveis a todas as situações, para o aprendizado pela experiência. Para *Berberich* a imitação

[391] GAMEZ *et al.*, 2020.
[392] GAMEZ *et al.*, 2020.

a partir de exemplos morais deve ser o conceito central em ética da virtude artificial[393]. A ideia de *self-improvement* é primordial nesse conceito, superando a noção de agente dotado de corpo de regras pronto, acabado e aplicável a qualquer situação.

Essa nova abordagem tem despertado o interesse dos engenheiros em computação, em razão da rápida expansão de novos dispositivos autônomos, das mais variadas espécies, desde veículos autônomos e robôs cuidadores a armas inteligentes. A dificuldade de atribuir a responsabilidade por todas as difíceis decisões concretas ao desenvolvedor, fabricante ou usuário desse sistema tem induzido uma nova abordagem, em que o sistema, de modo autônomo, decida qual a melhor escolha no caso concreto. Nesses casos, a engenharia não está preocupada se o sistema realmente decide ou aparenta escolher, imitando comportamentos humanos.

O modelo, denominado de teleológico, da ética aristotélica tem sido bem recebido pela ciência da computação. A obra dos fundadores da moderna cibernética, intitulada *Behavior, purpose and teleology*, de *Rosenblueth*, *Wiener* e *Bigelow*, de 1943, demonstrou exatamente a preferência da engenharia pela teoria moral aristotélica[394].

A escolha teórica foi confirmada nas obras posteriores "*The Human Use of Human Beings*", novamente do pioneiro *Wiener* (1950) e de *Terrell W. Bynum* (2005). O ponto principal dessas teorias não está exatamente na mera busca de fins, mas no "alinhamento de valores morais" (*moral alignment*). A importância do alinhamento de valores é fundamental para agentes artificiais por razões fundamentalmente de engenharia. A escolha teórica não decorre de questões filosóficas de fundo, mas de relevância prática. A máquina não pode processar todas as alternativas decisórias instantaneamente, ela deve escolher com base em fins e não em procedimentos.

Os agentes morais artificiais são descritos como possuidores de Crenças, Desejos e Intenções (*Beliefs, Desires and Intentions* – BDI), em um modelo de raciocínio prático proposto por *Bratman*[395]. Nesse modelo, as crenças representam o conhecimento do mundo, e os desejos são os objetivos para um determinado fim pretendido. As intenções seriam os planos do agente para alcançar esse desejo[396]. Alguns modelos computacionais são baseados no Modelo BDI,

[393] BERBERICH; DIEPOLD, 2018.
[394] BERBERICH; DIEPOLD, 2018, p. 5.
[395] BRATMAN, M. E. *Intentions, Plans, and Practical Reason*. CSLI, 1987.
[396] DIGNUM, Virginia. *Responsible artificial intelligence*: How to develop and use ai in a responsible way. Cham: Springer, 2019. p. 18.

tais como os elaborados pelo *Prof. Bordini*, no interpretador de programação orientada à agentes (*AgentSpeak*), denominado de *Jason*[397].

O conceito de autonomia é reconhecido como o mais característico de uma agência artificial, contudo, nenhum agente será capaz de agir de modo completamente autônomo em relação ao ambiente, em todas as situações e em todos os conflitos. A situação será ainda mais *problemática* nos casos de interação com outros agentes.

A locução do pai da cibernética, *Wiener*, é sintomática:

> Resultados desastrosos são esperados não apenas no mundo dos contos de fadas, mas no mundo real, sempre que duas agências essencialmente estrangeiras sejam acopladas na tentativa de alcançar um objetivo comum. Se a comunicação entre essas duas agências quanto à natureza dessa finalidade for incompleta, é de se esperar que os resultados dessa cooperação sejam insatisfatórios.
>
> Se usarmos, para alcançar nossos propósitos, uma agência mecânica cuja operação não podemos interferir eficientemente uma vez iniciada, porque a ação é tão rápida e irrevogável que não temos dados para intervir antes que a ação seja concluída, então tivemos é melhor ter certeza de que o objetivo colocado na máquina é o objetivo que realmente desejamos e não apenas uma imitação colorida dela[398].

A estratégia inicial denominada de Gofai (*Good Old-Fashioned AI*), a velha e tradicional inteligência artificial cognitivista, no modelo *top-down*, foi superada por novas abordagens. Contribuíram para esse novo caminho, o uso de redes neurais, uma visão *conectivista*, o uso de *big data* e de *deep learning*. Dentre tantos outros avanços em programação, podemos citar o uso do aprendizado de máquina baseado em reforço de aprendizado (*reinforcement learning*), conduzido com base em objetivos determinados (*goal-driven*) na condução de comportamentos de agentes[399].

A utilização de modelos de aprendizado de máquina se inserem em uma das características fundamentais dos agentes artificiais autônomos: a *adaptabilidade*[400]. Para

---

[397] BORDINI, R. H., HÜBNER, J. F.; WOOLDRIDGE, M. *Programming Multi-Agent Systems in Agent Speak Using Jason.* John Wiley & Sons, 2007; BORDINI, Rafael H.; MOREIRA Álvaro F. Proving BDI Properties of Agent-Oriented Programming Languages. *Ann. Math. Artif. Intell*., v. 42, n. 1-3, p. 197-226, 2004; VIEIRA, Renata *et al.* On the formal semantics of speech-act based communication in an agent-oriented programming language. *J. Artif. Intell. Res.* (JAIR), v. 29, p. 221-267, 2007.

[398] Cf. "*Disastrous results are to be expected not merely in the world of fairy tales but in the real world wherever two agencies essentially foreign to each other are coupled in the attempt to achieve a common purpose. If the communication between these two agencies as to the nature of this purpose is incomplete, it must only be expected that the results of this cooperation will be unsatisfactory. If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it*." WIENER, Norbert. *Some moral and technical consequences of automation science*, v. 131, Issue 3410, p. 1355-1358.

[399] BERBERICH; DIEPOLD, 2018, p. 5.

[400] FLORIDI, L. *The ethics of information*. Oxford University Press, 2013.

Floridi, é a característica de o agente artificial aprender com suas próprias experiências, sensações e interações, com capacidade de reagir ao ambiente[401].

Os mecanismos de aprendizado de máquina por reforço (*reinforcement learning*) são uma das três espécies de *machine learning*, os outros são o aprendizado não supervisionado e supervisionado.

O aprendizado de máquina (*machine learning*), ao contrário do que possa parecer, possui diferentes estratégias, objetivos, técnicas e modelos de treinamentos. Uma visão ilustrativa das abordagens principais pode ser vista no quadro a seguir, sobre os modelos de aprendizado não supervisionado, supervisionado e de reforço (*reinforcement learning*).

Quadro 1 – Quick Guide to Machine Learning

Quick Guide to Machine Learning				
Approach		**Unsupervised Learning**	**Supervised Learning**	**Reinforcement Learning**
Objective		Discover structures	Make predictions	Make decisions
Possible Techniques	**Simple domains**	• Clustering	• Regression • Classification	• Markov Decision Processes • Q-Learning
	Complex domains	**Deep Learning** (many-layered neural networks and large datasets)		
Training requirements			Labelled data	Reward function
Example Application		Customer segmentation	Identify spam	Playing a game (e.g. Go)

Fonte: BERBERICH; DIEPOLD, 2018.

O reforço no aprendizado de máquina (*reinforcement learning*) funda-se em três mecanismos de informações: pela reação do ambiente a condutas imorais, pela autorreflexão sobre a conduta pessoal e pela observação de comportamentos morais exemplares[402].

O modelo de *baixo para cima* (*bottom-up*) ou de *piso* é bastante compatível com uma ética das virtudes. A engenharia de programação demonstrou ser mais eficiente o estabelecimento de agência artificial e a definição de um valor de função a ser maximizado por consequências adequadas de uma ação correta. As ideias de aprendizado, reforço e habituação são centrais na ética das virtudes, mas não possuem o papel central nem na deontologia nem no consequencialismo[403].

---

[401] DIGNUM, *Responsible artificial intelligence*…, 2019, p. 17.
[402] BERBERICH; DIEPOLD, 2018, p. 6.
[403] BERBERICH; DIEPOLD, 2018, p. 6.

Aristóteles apresentava de modo claro esse ponto de vista ao afirmar:

> não se acredita que exista um jovem dotado de sabedoria prática. O motivo é que essa espécie de sabedoria diz respeito não só aos universais mas também aos particulares, que se tornam conhecidos pela experiência. Ora, um jovem carece de experiência, que só o tempo pode dar[404].

A conclusão que diversos cientistas de programação chegaram é a de que um modelo de agência artificial fundado em *machine learning*, combinado com a *ética da virtude*, é o caminho mais natural, coeso, coerente, integrado e "bem costurado" (*seamless*) do que as outras teorias morais[405].

O *modelo computacional da ética das virtudes* é significativamente mais complexo do que outras teorias morais (deontologia e consequencialismo). O agente deve ser capaz de racionalizar sobre seus motivos, suas ações e consequências[406]; e, ainda, aprender com estas, aprimorando a sua escala de virtudes morais e sua biblioteca de exemplos virtuosos.

Shannon Vallor será capaz de descrever adequadamente a natureza complexa do raciocínio prático na sua obra "Technology and the Virtues" da seguinte forma:

> A sabedoria prática é frequentemente classificada como uma virtude intelectual porque envolve cognição e julgamento; contudo, opera no âmbito moral, unindo capacidades cognitivas, perceptivas, afetivas e motoras em expressões refinadas e fluidas de excelência moral que respondem de maneira adequada e inteligente às exigências éticas de situações particulares[407].

O modelo aristotélico será o mais referenciado dentre as teorias morais. Contudo, ainda existem outros modelos que mereciam uma importante atenção, tal como as teorias morais da escolástica medieval e colonial. Autores que começou a explorar esse caminho são *Berberich et alii*[408]. Os autores passaram a debater as virtudes cardinais em *São Tomás de Aquino*, que seriam de quatro tipos: *prudência, coragem, temperança* e *justiça.* A possibilidade de modelar computacionalmente esses valores virtuosos desperta a curiosidade desses estudiosos.

Outro ponto importante é o de que a ética das virtudes possui natureza relacional, ou seja, exige responsabilidade do agente. O que significa um agente moral artificial ser responsável?

---

[404] ARISTÓTELES. *Ética a Nicômaco*. Tradução de Leonel Vallandro e Gerd Bornheim da versão inglesa de W. D. Ross. São Paulo: Nova Cultural, 1991. Livro VI, 8.
[405] BERBERICH; DIEPOLD, 2018, p. 6.
[406] DIGNUM, *Responsible Artificial Intelligence*…, 2019, p. 44.
[407] Cf. "*Practical wisdom is often classified as an intellectual virtue because it involves cognition and judgment; yet it operates within the moral realm, uniting cognitive, perceptual, affective, and motor capacities in refined and fluid expressions of moral excellence that respond appropriately and intelligently to the ethical calls of particular situations*". VALLOR, Shannon. *Technology and the virtues*: a philosophical guide to a future worth wanting. New York: Oxford University Press, 2016. p. 99.
[408] BERBERICH; DIEPOLD, 2018, p. 5.

2.1.7 Máquinas responsáveis

Uma teoria moral artificial está ligada à noção de responsabilidade moral. Deve existir um sentimento de empatia pelo sofrimento do outro, de dever de cuidado, de compaixão pela situação alheia. A virtude da *empatia* cumprirá um papel fundamental em um sistema artificial[409]. Muitas das principais relações humanas se fundam na virtude da empatia por outros, pelo simples amor ao servir[410]. A compaixão é desinteressada. O amor fraterno, filial, conjugal ou religioso pode ser abençoado pelo desprendimento. Por entregas sem contrapartida, pelo simples bem alheio.

Uma inteligência artificial responsiva (*responsible articial intelligence*) é comprometida com a noção de que as ações de agentes autônomos devem ser eticamente responsáveis pelas consequências de seus atos[411].

A noção de uma *IA Responsável* (*Responsible* AI) permeia não somente os agentes autônomos, mas todos os sistemas inteligentes, desde o momento da pesquisa (*Responsible Research and Innovation* – RRI) até as ações de agentes morais artificiais[412].

Os sistemas inteligentes são divididos em três níveis de autonomia[413]. Primeiro podem ser considerados como meras ferramentas, que auxiliam os humanos a realizarem suas tarefas. A responsabilidade ética, nesse caso, é *operacional* e se dirige aos agentes humanos, que programam, fabricam ou utilizam máquinas inteligentes. Em segundo lugar, se encontram os sistemas inteligentes assistentes. Estes não são completamente autônomos, mas possuem uma "consciência" do ambiente com o qual interagem. A responsabilidade ética nesse caso será *funcional*, ou seja, as máquinas artificiais poderão ajustar/adaptar a sua conduta conforme o ambiente. Por último, teremos os agentes morais artificiais completos, capazes de reflexão, adaptabilidade ao ambiente e tomada de decisões éticas com responsabilidade completa (*full ethical behaviour*).

Três são os princípios norteadores da responsabilidade ética dos sistemas inteligentes artificiais: adaptabilidade, responsabilidade e transparência.

Dignum irá acrescentar mais um princípio, o da participação, na formatação do desenho dos sistemas inteligentes. Um sistema responsável deve levar em consideração todos os

---

409 VALLOR, 2016, p. 139.
410 VALLOR, 2016, p. 139.
411 DIGNUM, *Responsible Artificial Intelligence*…, 2019, p. 48.
412 DIGNUM, *Responsible Artificial Intelligence*…, 2019, p. 48.
413 DIGNUM, *Responsible Artificial Intelligence*…, 2019, p. 88.

aspectos éticos e sociais, compartilhados em sociedade. Indivíduos, grupos e sociedades possuem diferentes visões morais e valorativas, que devem ser levadas em consideração. Afinal de contas, diferentes valores implicam diferentes decisões[414].

Fishkin defende que uma escolha moral legítima deve respeitar cinco características essenciais:

i. informações (*informatione*): devem acuradas, relevantes e acessíveis para todos os participantes;
ii. balanço substantivo (*substantive balance*): diferentes posições podem ser comparadas, com base em suas evidências de suporte (*supporting evidence*);
iii. diversidade (*diversity*): todas as principais posições relevantes estão disponíveis para todos os participantes;
iv. conscientização (*conscientiousness*): os participantes ponderam todos os argumentos;
v. igual consideração (*equal consideration*): as visões são baseadas em evidências e não em uma visão particular.

Dignum irá ainda acrescentar o princípio fundamental da transparência (*openness*), que determina que todas as opções e escolhas se encontram claras e acessíveis[415].

A ideia de uma inteligência artificial responsável é muito ampla. De um lado, envolve a exigência de que os sistemas artificiais inteligentes irão interagir com os seres humanos em uma miríade de situações[416]. Algumas rotineiras, como a assistência em tarefas domésticas; outras muito delicadas, como em cirurgias, em drones militares ou na área jurídica. Reconhecer a importância desses sistemas em comportarem-se de modo responsável perante seres humanos é uma tarefas desafiadora. Quais a serão os níveis de responsabilidade a serem exigidos desses autômatos? Como organizar a arquitetura de escolhas, dados e conhecimento, processo e colaboradores, de modo a termos um sistema artificial racional, virtuoso e responsável?

O primeiro aspecto da inteligência artificial responsável está no plano da responsabilidade humana[417] em assumir a inarredável tarefa de preocupar-se com o uso e desenvolvimento responsável de agentes morais artificiais. Talvez o surgimento de agentes morais artificiais completos não seja nunca alcançado. Talvez a técnica falhe miseravelmente em criar sistemas robóticos similares aos seres humanos. Muitas são as possibilidades de

[414] DIGNUM, *Responsible Artificial Intelligence*…, 2019, p. 84.
[415] DIGNUM, *Responsible Artificial Intelligence*…, 2019, p. 85.
[416] DIGNUM, *Responsible Artificial intelligence*…, 2019, p. 119.
[417] HARARI, Y. N. *Homo Deus*: a brief history of tomorrow. Random House, 2016; WALSH, T. *Machines that think*: the future of artificial intelligence. Prometheus Books, 2018.

desenvolvimento tecnológico. Contudo, não seria responsável não analisar com cuidado tal situação e o impacto que teria ou terá sobre a condição humana.

Os sistemas artificiais inteligentes terão um impacto sobre todas as esferas da vida humana futura. Sejam eles operativos, assistivos ou agentes morais completos. A economia, a política, o direito, a cultura e a sociedade serão impactados de uma forma ainda não compreendida em toda a sua profundidade. Assim, diversas perguntas se encontram em aberto. Qual será o futuro dos empregos? Da democracia? Como os agentes artificiais inteligentes irão impactar a economia? Quais serão os deveres desses autômatos? Terão direitos? As perguntas são tão variadas e complexas que é impossível delimitar todas em uma sistematização, sem o risco de reduzir a complexidade inerente ao tema.

Os desafios são impactantes, complexos, importantes, mas exigirão respostas cada vez mais rápidas da humanidade. No princípio o indivíduo se deparava com os desafios naturais ou dos deuses. Os poucos autômatos na cena ficcional eram peças de ornamentação, frente aos grandes rivais (natureza e desígnios divinos). Com o domínio crescente da natureza pela técnica e a dessacralização da vida moderna, o grande desafio é ser humano em um mundo de seres (mais) inteligentes artificiais.

As máquinas podem, teoricamente, possuir o raciocínio mais rápido, mais resiliente, talvez sejam mais éticas. Mas talvez jamais tenham o mais importante: *alma.*

2.2. DA POSSIBILIDADE DE ALGORITMOS MORAIS

O objetivo do presente tópico está na análise da objeção algorítmica, ou seja, da tese da impossibilidade de que um algoritmo possa permitir a *emergência* de uma inteligência artificial forte. Esse argumento se caracteriza como (im)possibilidade de superação do *Teste de Lovelace.* Poderia uma máquina artificial criar algo?

2.2.1 Algoritmos morais

A origem da palavra algoritmo é antiga e objeto de controvérsias. Segundo a informação mais aceita, ela decorre dos trabalhos do matemático nascido na Pérsia *Muhammad ibn Musa al-Khwarizmi*, em aproximadamente 783–850 d.C. Esse matemático famoso é reconhecido como o fundador desta bela área da matemática, a álgebra. Ele teria escrito o primeiro manual de Álgebra e um “Manual de Cálculo com Algorismos”. A tradução da obra seria feita para o

latim como *Algoritmi de numero Indorum*, também conhecido como *Dixit Algorismi* (*Algorismos têm dito*)[418]. A partir de então, o termo *algorismo* passou a significar a contagem com a ajuda de números arábicos.

O conceito de algoritmos trouxe uma revolução no pensamento matemático do século XX. Afinal quais seriam os limites do pensamento matemático? Poderia existir um método que determinasse quais problemas poderiam ser solucionados e quais ficariam sem solução? Esse era o famoso problema de *Hilbert*[419], que recebeu a genial resposta de *Turing*, sob a forma da famosa *Máquina de Turing* (1936). Essa não era verdadeiramente uma máquina real, como vimos, mas um modelo ideal capaz de modelar qualquer computador digital. A sua importância foi ímpar para a teoria da computação, dado que permitiu o surgimento de modelos gerais que comandariam a revolução informática neste século. E na base de toda essa revolução estavam os algoritmos.

O significado atual de algoritmos é "um conjunto de passos, passível de repetição, que resolve um problema"[420]. Esse sentido mais amplo pode ser reduzido a um sentido mais restrito, como um conjunto de rotinas automatizadas, que seguem um procedimento preestabelecido. Os algoritmos paulatinamente assumiram uma posição de destaque em ciência da computação, em que a sua principal tarefa é a possibilidade de resolver um problema e ser capaz de repetir indefinidamente essa operação.

De forma muito simplificada, vejamos a estrutura desses modelos computacionais. O algoritmo somente pode estar bem estruturado se seguir uma determinada *lógica*, ou seja, deve ser formado por *sentenças* que se expressam conforme uma certa *sintaxe*. Esta por sua vez garante que as sentenças estejam bem formadas. A sintaxe utilizada irá garantir a produção de raciocínios lógicos com sentenças consistentes, sendo que uma lógica deve igualmente possuir uma semântica, ou seja, um sentido para as sentenças. Elas devem possuir um determinado valor de verdade em relação a cada *mundo possível*. Nas lógicas clássicas, os valores de verdades ocorrem de modo excludente, ou a sentença é verdadeira ou falsa, não podendo ser simultaneamente uma e outra. A ciência da computação passou a utilizar a expressão *modelo* para designar um mundo possível. Assim a afirmação "m é modelo para a sentença α".

---

[418] Al-Kwarizmi. n.d. Disponível em: http://jnsilva.ludicum.org/hm2008_9/Livro9.pdf. Acesso em: 31 jul. 2020 às 00:56.
[419] TURING, Alan. On computable numbers, with an application to the Entscheidungs problem. *Proceedings of the London Mathematical Society*, Series 2, v. 42, 1936-7, p. 230-265. Disponível em: https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf. Acesso em: 31 jul. 2020 às 00:56.
[420] SOFFNER, Renato. *Algoritmos e programação em linguagem C*. São Paulo: Saraiva, 2013. p. 21.

O raciocínio computacional exige sentenças bem formadas, conforme uma determinada sintaxe e semântica, bem como, conforme determinado modelo, com seus respectivos valores de verdade. Mas é fundamental, para que ocorra um raciocínio válido, que exista uma *implicação* lógica entre as sentenças, ou seja, que, dada uma sentença *p,* se siga logicamente outra sentença. A implicação lógica toma a forma $p \rightarrow q$ (lê-se: *se p então q*). Assim, se a sentença *p* é verdadeira, segue-se que a sentença *q* também é.

Outro conceito relevante será o de inferência lógica, que é o processo lógico pelo qual, a partir de certos dados, se chega a determinadas conclusões. Um algoritmo de inferência será aquele do qual se derivam sentenças válidas, a partir de determinadas sentenças. As regras de inferência são *standards* de inferência, que podem derivar de cadeias de conclusões, que nos levam a resultados desejados, tal como o *modus ponens* (*se a sentença p implica p e q, então p deve ser inferida*).

A base epistemológica para a estruturação dos algoritmos é dada pelas noções de sentença, *sintaxe, implicação e inferências*, mas a representação do mundo por meio de algoritmos exige a estruturação do conhecimento pela forma ordenada de uma ontologia[421]. *Neches* foi o primeiro a definir uma ontologia como "os termos básicos e as relações que definem um vocabulário"[422]. Tem sido entendida como uma especificação formal de conceptualizações compartilhadas, isto é, de modelos abstratos de certos fenômenos[423]. Assim, por exemplo, a construção de ontologias legais é uma parte fundamental dos algoritmos jurídicos, como forma de conceituação abstrata do fenômeno normativo. Não há, contudo, acordo sobre a melhor forma de construção de determinada ontologia específica ou especializada, de tal modo que os resultados serão distintos, conforme o modo de construção. Se a programação for na área médica, militar ou financeira.

Outra dificuldade possui uma natureza técnica, mas com consequências éticas profundas. *Nick Bostrom* e *Eliezer Yudkowsky* relataram a importância da escolha do modelo de geração de algoritmos. Existem fundamentalmente dois modelos, em árvore de decisão (IA simbólica) e redes neurais ou algoritmos genéticos (IA conexionista).

---

[421] PEREZ, A.G.; RODRIGUEZ, F.O.; TERRAZAS, B.V. Legal ontologies for the Spanish e-government. *In*: CAEPIA. *Researchgate.net/*, 2006. p. 301-310. Disponível em: https://www.researchgate.net/profile/Asuncion_Gomez-Perez/publication/221275037_Legal_Ontologies_for_the_Spanish_e-Government/links/0fcfd50b23ad68a223000000/Legal-Ontologies-for-the-Spanish-e-Government.pdf. Acesso em: 17 dez. 2020 às 00:58.

[422] "*Ontology defines the basic terms and the relations that include the vocabulary of a specific area, in addition to the rules to combine terms and relations to define extensions to the vocabulary*". NECHES, R. *et al.* Enabling Technology for Knowledge Sharing. *AI Magazine*, v. 12, n. 3, p. 36-56, 1991.

[423] PEREZ, A.G.; RODRIGUEZ, F.O.; TERRAZAS, B.V, 2006, p. 302.

Dois modelos de inteligência artificial se destacaram deste o início: a IA simbólica e a IA conexionista. O primeiro utiliza essencialmente o modelo estruturado em árvores de decisão, enquanto o segundo procura reproduzir o funcionamento do cérebro por meio de redes neurais. Apesar de possuírem origem praticamente no mesmo período, os dois modelos seguiram desenvolvimentos distintos, em face da capacidade computacional disponível e dos avanços em linguagem de programação. O uso de IA *conexionista*, apesar de ser mais promissora, exigia maior quantidade do uso de processamento de máquina. Por outro lado, a dificuldade do uso de redes neurais está na sua opacidade, ou seja, de sua abertura à transparência e previsibilidade nas decisões. Atualmente tem se trabalhado com modelos que utilizam de modo misto tanto a IA simbólica, quanto a conexionista.

Aceita a possibilidade teórica de agentes artificiais com racionalidade, consciência, perspectiva de primeira pessoa, senciência, emoções e diretrizes morais, podemos questionar se seria possível ocorrer a emergência de intencionalidade moral. Assim, a questão que se impõe é a de que se existe a possibilidade de um algoritmo permitir a emergência de um agente moral artificial ou se regras morais somente poderiam ser introjetadas externamente.

Estruturar algoritmos morais não é uma tarefa simples, fácil ou clara. Diversos são os desafios, obstáculos e dificuldades.[424] *Derek* ressalta algumas dificuldades, tais como a presença de vieses e tendências do programador no algoritmo moral. Assim, digamos que o programador possua dados valores pessoais, preconceitos, preferências e interesses. Ele poderá incluir, de modo transparente ou oculto, suas orientações subjetivas. Os efeitos dessas escolhas podem ser desastrosas. Diversos estudos relatam casos de algoritmos discriminatórios e medidas antivieses.

Outro problema citado pelo autor decorre da escolha da teoria moral de preferência do programador. Assim, caso o código incorpore determinado modelo teórico (*realismo* ou *antirrealismo moral*), este terá efeitos importantes nas escolhas do agente moral artificial. Digamos que agentes artificiais adotem em seu algoritmo teorias morais distintas[425], será possível ocorrer um conflito ético?

Diversos estudos apresentaram desenhos de mecanismos computacionais sobre regras morais utilizando as modelagens morais *top-down* e *bottom-up*, bem como a solução de conflitos morais por meio de algoritmos de assistência moral. Outros estudos se dedicaram à responsabilidade na construção de algoritmos morais sem vieses ou com gatilhos de proteção

[424] LEBEN, Derek. *Ethics for robots*: how to design a moral algorithm. Oxon/NewYork, NY: Routledge, 2018. p. 4.
[425] LEBEN, 2018, p. 5.

anti-bias. O propósito de nosso estudo não é verificar os desafios da formulação responsável ou dos controles aos algoritmos morais, mas a possibilidade de estes permitirem a *emergência de agentes morais artificiais*.

2.2.2 Da possibilidade de emergência de agentes morais artificiais

A possibilidade de *emergência de agentes morais artificiais* é dos temas mais instigantes e difíceis? A dificuldade principia com a compreensão e aceitação do conceito filosófico de *emergência*.

A sabedoria romana já observava esse fenômeno no famoso brocardo que dizia *senatores boni viri, senatus mala bestia* (os senadores são bons homens, mas o Senado é mau). De entidades isoladas pode surgir um todo completamente diferente, que não pode simplesmente ser deduzido das qualidades dos seus componentes individualmente considerados[426].

A noção filosófica de emergência remonta à *Stuart Mill*, que irá propor a ideia na obra "*System of Logic*" (1843). O autor irá diferenciar dois modos de "ação conjunta de causas" (*the conjoint action of causes*), as mecânicas e as químicas. No modo mecânico, o efeito conjunto de causas nada mais é do que a soma do efeito das causas tomadas isoladamente. Como exemplo, citava como o modo químico reagia de forma distinta. Nesse caso, os efeitos não são meramente aditivos. O efeito conjunto de diferentes causas é diferente da mera soma dos efeitos das causas tomadas isoladamente[427].

Lloyd Morgan irá defender essa noção nas obras "*In Emergent Evolution*" (1923), "*Life, Spirit and Mind*" (1926) e "*The Emergence of Novelty*" (1933). O autor irá introduzir a noção de emergência no processo de evolução. Para ele as propriedades emergentes são causalmente autônomas e possuem poderes causais descendentes (*Thus emergent properties are causally autonomous and have downward causal powers*)[428]. Ou seja, o curso de novas propriedades não pode ser derivado ou previsto a partir de entidades anteriores.

---

[426] NEGROTTI, Massimo. Naturoids — on the nature of the artificial. London/Singapore: World Scientific Publishing, 2002, p. 48.
[427] VINTIADIS, Elly. *Emergence.* Internet Encyclopedia of Philosophy. Disponível em: https://www.iep.utm.edu/emergenc/. Acesso em: 24 ago. 2020 às 01:52.
[428] *Idem, ibidem*.

O conceito de poderes causais descendentes (*downward causal powers*) de um nível superior para um nível inferior pode receber três sentidos distintos: forte, médio ou fraco[429]. No *conceito forte* (*strong emergence*), a alteração em um nível superior implica mudanças no nível inferior. Não há uma causação direta nos níveis inferiores por alterações nos níveis superiores, na *causação descendente média*. Por último, na *causação descendente fraca*, os níveis superiores somente possuem a função organizacional, da estrutura dos elementos constituintes. Assim, uma alteração nesse nível possui somente a potencialidade de alteração na dinâmica inferior[430].

C. D. Broad's, em sua obra "*Mind and Its Place in Nature*" (1925), tratou da questão sobre se as propriedades de um sistema complexo são diretamente relacionadas às propriedades de suas partes. Os emergentistas defendiam que o comportamento do todo não pode ser deduzido do conhecimento do comportamento das partes. Esse seria um fenômeno ontológico e não epistemológico, ou seja, decorreria da estrutura metafísica do mundo. O "arcanjo matemático" (*mathematical archangel*) não teria previsto as propriedades emergentes porque estas são fatos brutos e não seriam redutivamente explicáveis[431].

O *emergentismo* surge hodiernamente com a teoria dos sistemas complexos, a neurociência e a filosofia da mente. *David Chalmers* (2006) afirmava que o *emergentismo fraco* (*weak emergence*) é comum e compatível com as noções de auto-organização, complexidade e não linearidade. Trata-se de uma noção epistemológica e não metafísica, sendo definida em termos de *imprevisibilidade* (*unpredictability or unexpectedness*). Assim, dadas as características e propriedades das partes de baixo nível ou fundamentais, podem emergir propriedades imprevisíveis. Esse é o caso tanto de padrões emergentes no *automata* celular ou em redes conexionistas ou em transições de fase, como congestionamentos, voos de bandos de pássaros, etc.[432]

A emergência fraca é compatível com a redução, no sentido de que ela é imprevisível, porém reduzível. As partes podem ser compreendidas conforme leis determinísticas, mas os seus resultados são imprevisíveis devido às consequências decorrentes das condições iniciais.

A questão que se põe é sobre a possibilidade da emergência de agentes morais artificiais. Seria essa uma possibilidade real? Uma resposta positiva para essa pergunta foi dada por *Di*

---

[429] EMMECHE, Claus; KØPPE, Simo; STJERNFELT, Frederik. *Levels, Emergence, and Three Versions of Downward Causation*. Disponível em: http://www.nbi.dk/~emmeche/coPubl/2000d.le3DC.v4b.html. Acesso em: 24 ago. 2020 às 15:04.
[430] CHALMERS, D. Strong and Weak Emergence. *In*: DAVIES, P.; CLAYTON (ed.). *The re-emergence of emergence*. Oxford University Press, 2006. p. 1-03.
[431] BROAD, C.D. *The mind and its place in nature*. London: Routledge and Kegan Paul, 1925. p. 71-72.
[432] CHALMERS, 2006, p. 1-03.

Marzo et alii e avalizada por outros autores na análise de sistemas multiagentes (*Multi-Agent Systems* – MAS) coordenados por auto-organização (*self-organization*) e mecanismos de emergência (*emergence mechanisms*)[433].

Os sistemas de auto-organização são aqueles que funcionam sem um controle central e operam com base em interações contextuais. A particularidade desses sistemas está na espontaneidade, em face de mudanças no ambiente. Segundo os autores, a auto-organização pode permitir o surgimento de comportamentos emergentes. Tal situação tem sido particularmente explorada em ocasiões em que a ação centralizada não é possível, pela dificuldade de supervisão[434]. Assim, por exemplo, no caso de redes de sensores, controle de veículos aeroespaciais ou em zonas perigosas.

Os algoritmos possuem um papel fundamental na auto-organização de entidades autônomas para se organizarem. Diversos softwares têm sido utilizados para não apenas simularem a auto-organização, mas para permitirem a realização de funcionalidades emergentes.

2.2.3 Da possibilidade de algoritmos que possuam mecanismos de emergência

Dado que é possível a emergência de agentes morais artificiais, cabe questionar se é possível estruturar algoritmos capazes de permitir a *emergência* de *agentes morais artificiais completamente autônomos.*

Uma saída apresentada por *Leben* está no uso de modelos da teoria dos jogos morais. Como fundamento desse entendimento, tem-se que os algoritmos morais pretendem alcançar determinados fins. As regras morais seguiriam uma arquitetura racional capaz de solucionar os mais difíceis dilemas éticos por meio de modelos de cooperação e não cooperação, denominados de *Maximin*[435].

Um dos setores de inteligência artificial mais avançados se relaciona ao *aprendizado de máquina por reforço de multiagentes* (*multi-agent reinforcement learning* – MARL), que se caracteriza pelo estudo da dinâmica cooperativa e competitiva entre agentes artificiais inteligentes.

---

[433] DI MARZO, Giovanna; GLEIZES, Marie-Pierre; KARAGEORGOS, Anthony. Self-Organisation and Emergence. *MAS: An Overview Informatica*, v. 30, p. 45-54, 2006.
[434] DI MARZO; GLEIZES; KARAGEORGOS, 2006, p. 45-54.
[435] LEBEN, 2018, p. 5.

Jaderberg et alii comprovaram a possibilidade prática de emergência espontânea de um comportamento, que nunca havia sido explicitamente treinado. Foi utilizado com sucesso, em um jogo, o uso de um algoritmo de aprendizado por reforço (*RL-based training*). Nesse caso, surgiram espontaneamente comportamentos eficientes na busca de melhores resultados[436]. O jogo estabelecia, como objetivo, que equipes multiagentes deveriam capturar uma bandeira. Tratava-se de uma tarefa que não havia sido programada, e as equipes deveriam aprender como alcançar esse desafio. Um dos pontos importantes do programa é que ele gerava, de modo aleatório, todos os mapas e as informações espaciais das bandeiras. Essa variável aumentava consideravelmente o esforço de coordenação por aprendizado das equipes artificiais, que deveriam se coordenar em uma rica, múltipla e variada representação de ambientes.

Um ponto merece muito destaque e atenção. Afinal, a inteligência artificial poderia ser tão eficiente quanto a humana? Para isso ela deveria dominar o grande segredo dos *sapiens*. Algo que permitiu à espécie que se instalasse em hábitats distantes e inóspitos, totalmente desconhecidos. Sem mapas prévios e lidando com outras espécies completamente desconhecidas, lidaram com outras espécies humanas e as superaram – ou, melhor, as massacraram. Os *sapiens* conquistaram o mundo graças à sua linguagem única[437]. O desenvolvimento de agentes morais completamente autônomos deve prever o uso ou a emergência de uma linguagem artificial eficiente.

Mas seria possível a emergência de uma linguagem artificial singular e eficiente? Poderiam agentes artificiais criarem tal linguagem[438]? Talvez essa seja a ferramenta mais importante rumo à superação da *Lei de Lovelace*.

O surgimento emergente de comportamentos multiagentes foi comprovado em outras situações[439]. Foi constatada a possibilidade de emergência de linguagem composicional fundamentada (*grounded compositional language*) para atingir finalidades e objetivos em populações de multiagentes[440].

Nesse teste, cada agente tinha objetivos a realizar, especificados por vetores e não observáveis externamente pelos demais agentes. Dentre os objetivos estavam as tarefas de movimentação e posicionamento espacial. Estes poderiam exigir algum grau de coordenação

---

[436] JADERBERG Max *et alii*. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science31*, p. 859-886, May 2019.
[437] HARARI, Yuval. *Sapiens*: Uma breve História da humanidade. São Paulo: L&PM Editores, 2015, p. 27-28.
[438] KIRBY, Simon. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. IEEE Trans. *Evolutionary Computation*, v. 5, p. 102-110, 2001; KIRBY, Simon. Natural language from artificial life. *Artificial Life*, 2002; CHRISTIANSEN, Morten H.; KIRBY, Simon. Language evolution: consensus and controversies. *Trends in cognitive sciences*, v. 7, n. 7, p. 300-307, 2003.
[439] MORDATCH, 2018.
[440] MORDATCH, 2018, p. 1.495.

ou comunicação entre agentes. O teste revelou não somente o uso de ferramentas verbais, mas a emergência do uso de sinais não verbais que não tinham sido ensinados, bem como outras estratégias não comunicativas (*noncommunicative strategies*)[441]. O teste relatou a emergência de uma linguagem composicional abstrata a partir de uma experiência fundamentada (*emergence of an abstract compositional language from grounded experience*)[442].

O objetivo de criar agentes artificiais dialogais orientados para alcançar objetivos (*goal-driven dialog agents*), capazes de perceber o ambiente, por meio da visão, audição ou sensores e interagir com humanos ou outros agentes, mediante comunicação, possui limites técnicos ainda.

Os estudos demonstram, porém, que os agentes não dominam o significado funcional da linguagem, tais como *grounding* (mapeamento de palavras para conceitos físicos), *composicionalidade* (combinação de conhecimento de conceitos mais simples para descrever conceitos mais ricos) ou *aspectos do planejamento* (entendendo o objetivo da conversa). Mais grave, a linguagem natural não emerge *naturalmente* no diálogo multiagentes, apesar de relatos técnicos nesse sentido[443]. Esses resultados provisórios não provam, contudo, a impossibilidade técnica, apenas a limitação técnica atual.

Desse modo, não se pode concluir pela impossibilidade de emergência de comportamentos comunicacionais em algoritmos que utilizem o aprendizado de máquina por reforço em multiagentes (*multi-agent reinforcement learning* – MARL). Pelo contrário, os estudos tendem a avançar em complexidade, profundidade e ousadia nas possibilidades computacionais[444], expandindo o caminho para a superação do *Teste de Lovelace*.

Obviamente a possibilidade de emergência de uma linguagem não basta para a singularidade artificial. Insetos possuem linguagem. Formigas e abelhas se comunicam e cooperam para alcançarem objetivos comuns. Tampouco a comunicação humana se destaca por ser a única linguagem verbal ou vocal. Os símios possuem alguma modalidade desta. Os *sapiens* dominaram porque possuíam algo muito diferenciado, uma linguagem extremamente versátil[445].

---

[441] MORDATCH, 2018, p. 1.497.
[442] MORDATCH, 2018, p. 1.501.
[443] KOTTUR, Satwik *et al.* Natural language does not emerge 'naturally' in multi-agent dialog. *In*: EMNLP, 2017. Disponível em: https://arxiv.org/pdf/1706.08502.pdf. Acesso em: 26 jul. 2020 às 00:37.
[444] LAZARIDOU, A.; PEYSAKHOVICH, A.; BARONI, M. Multi-agent cooperation and the emergence of (natural) language. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 2017; LAZARIDOU, A.; HERMANN, K. M.; TUYLS, K.; CLARK, S. Emergence of linguistic communication from referential games with symbolic and pixel input. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), Vancouver, 2018; LEE, Jason D. *et al.* Emergent translation in multiagent communication. *CoRR*, abs/1710.06922, 2018.
[445] HARARI, 2015, p. 27-28.

Eles se comunicavam para alcançar objetivos comuns, mas, principalmente, conseguiram trabalhar e transmitir a sua imaginação. São a única espécie a inventar a ficção. Graças a ela foram possíveis os mitos, as coisas que não existem, as abstrações que estão na base da economia (dinheiro), política (governo) e direito (pessoa jurídica)[446].

Os experimentos atuais demonstram a possibilidade de emergência de comunicação artificial, mas nada provam sobre a possibilidade de uma *imaginação artificial*. Tampouco a *Lei de Lovelace* se detém nesse aspecto. O seu teste é limitado a declarar a impossibilidade de algoritmos criarem algo. O que está se provando superado.

Diversos estudos têm sido conduzidos justamente nessa preocupação. Seria possível a emergência de uma criatividade artificial?[447]. Inicialmente, o uso de algoritmos de geração de linguagem natural (*Natural Language Generation* – NLG) serviam a um único propósito, assistir os seres humanos em suas tarefas de redação de textos, tais como e-mails, no jornalismo digital e mesmo auxiliando autores de ficção na produção de conteúdo. Igualmente, houve uma expansão do uso técnico da geração de textos em áreas técnicas como no direito, em finanças e medicina[448].

Os programadores são testados a apresentarem uma nova geração de algoritmos ainda mais ousados. O objetivo é produzir um modelo autônomo na geração de textos criativos, que sejam instigantes. Trata-se se um desafio gigantesco. A escrita criativa, contudo, ainda, tem sido considerada impregnável para as máquinas.

Uma nova geração de algoritmos criativos tem surpreendido pelo seu poder de geração de conteúdo. Os resultados têm se estendido de textos em não ficção, artigos de notícias até ficção, como dramas e poesias[449]. Os textos demonstram as qualidades técnicas da gramática e semântica, bem como combinam uma qualidade preocupante. Os potenciais conflitos éticos decorrentes assustam. A manipulação de informação ou desinformação é real. Mas o mais importante é que as portas em direção à criatividade artificial foram abertas e não podem ser simplesmente negligenciadas. Alegar uma teórica impossibilidade técnica não condiz com o debate no setor tecnológico. Desse modo, cabe considerar a realidade de um cenário de emergência de uma linguagem artificial de alto nível, versátil e criativa, como algo possível,

---

[446] HARARI, 2015, p. 30.
[447] KOBIS, Nils; MOSSINK, Luca. Creative artificial intelligence – Algorithms vs. humans in an incentivized writing competition. *ResearchGate.* 2020. Disponível em: https://www.researchgate.net/publication/338689473_Creative_Artificial_IntelligenceAlgorithms_vs_humans_in_an_incentivized_writing_competition. Acesso em: 25 jul. 2020 às 10:11.
[448] KOBIS; MOSSINK, 2020, p. 2.
[449] KOBIS; MOSSINK, 2020, p. 3.

mesmo que improvável a curto prazo. A superação do Teste de Turing (fraco) parece próxima nesse campo.

2.2.4 Algoritmos evolucionários morais

A necessidade do debate sobre os algoritmos morais possui uma relevância existencial para humanidade. Não se trata de mero problema técnico. Dado que podem surgir agentes artificiais dotados de intencionalidade, autonomia e perspectiva de primeira pessoa, então os debates sobre a possibilidade de uma moralidade artificial se tornam prementes. Conforme *Picard* (1997): "quanto maior for a Liberdade de uma máquina, maior será a necessidade de padrões morais" (*The greater the freedom of a machine, the more it will need moral standards*)[450].

A possibilidade de geração de algoritmos morais é algo claro para a doutrina[451]. Os algoritmos morais têm sido estudados como assistentes em decisões, como protocolo para ação de drones, robôs e softwares[452]. A questão que se impõe é outra: é possível existir um verdadeiro agente moral artificial? Este poderia emergir da máquina ou seria simplesmente um aparato construído e alimentado por dados e comandos externos? Afinal, se trata de um problema de ordem conceitual ou ontológico[453]? Estas questões não serão respondidas no presente trabalho, mas logo se percebe a sua importância fundamental para avançar nos questionamentos sobre a natureza e características de algoritmos morais autênticos.

Três estratégias têm sido vislumbradas para alcançar o objetivo de um verdadeiro agente moral artificial, conforme verificado anteriormente, o modelo *top-down* (*de cume*), o *bottom-up* (*de baixo para cima*) e a combinação híbrida. A dificuldade da estratégia de cume está na sua impossibilidade técnica em prever todos os casos de dilemas morais.

Gerdes e *Øhrstrøm* defendem a combinação híbrida entre a estratégia *top-down*, em termos de teoria ética, *bottom-up* no uso de redes neurais e *machine learning*[454]. Outra alternativa seria a utilização direta de uma estratégia de baixo para cima (*bottom-up*) e questionar se os agentes artificiais podem aprender espontaneamente a diferença entre o certo e o errado, sem depender de uma teoria moral de cume (estratégia *top-down*).

---

[450] PICARD, R. *Affective Computing Cambridge*, MA: MIT Press, 1997.
[451] KEARNS, Michael; ROTH, Aaron. *The ethical algorithm*: the science of socially aware algorithm design Publisher. Oxford University Press, 2019.
[452] *Ibidem*.
[453] ALLEN; VARNER; ZINSER, 2000, p. 252.
[454] ARNOLD; SCHEUTZ, 2016.

Seria possível, no entanto, emergir *crenças morais artificiais* sobre o certo e o errado?

Wendell Wallach e *Colin Allen* trataram da primeira etapa desse desafio. Seu questionamento se dirige a ensinar as diferenças entre certo e errado para os robôs. A experiência do autor tem demonstrado que essa tentativa tem indicado aos humanos diversas lacunas no raciocínio moral[455].

Outro caminho analisado são os algoritmos evolucionários em teoria dos jogos (*evolutionary game-theory*), aplicados à ética. Essa estratégia desvia do problema de criar um sistema complexo, de larga escala, entre agentes racionais, plenamente informados e dotados de regras morais claras. A moralidade é entendida como um efeito não intencional da interação de agentes. Mas, principalmente, nesse modelo evolucionário, a moralidade *emerge*[456] de uma série de interações repetidas entre pequenos grupos de agentes para a solução de problemas repetidos[457].

O uso de algoritmos evolucionários permite compreender a emergência de normas morais em situações de cooperação e não cooperação para a solução de dilemas[458]. Contudo, ela não auxilia a entender o que é certo ou errado, mas tão somente o resultado intencional ou não intencional das interações entre agentes racionais[459]. Os estudos têm comprovado claramente a possibilidade da emergência de normas morais, porém não quais normas morais emergem ou deveriam emergir. A ideia de que os sistemas morais devem ser considerados como sistemas evolucionários[460] é partilhada por outros autores[461]. E, assim,

---

[455] WALLACH, Wendell; ALLEN, Colin. *Moral Machines*: teaching robots right from wrong. Oxford: Oxford Scholarship Online, 2009.

[456] Cf. "*First, the evolutionary approach provides a genuine explanation of the emergence and persistence of moral norms. Norms are the unintended side-effect of the actions of (boundedly) rational agents and emerge in the process of repeated interactions. On the evolutionary approach, the 'function' of a moral norm is to select a stable equilibrium, in a situation in which there is more than one*". VERBEEK, Bruno; MORRIS, Christopher. *Game Theory and Ethics*. 2009. Disponível em: https://stanford.library.sydney.edu.au/archives/spr2009/entries/game-ethics/#7. Acesso em: 26 jul. 2020 às 21:38.

[457] BRAITHWAITE, Richard Bevan. *Theory of games as a tool for the moral philosopher*. Cambridge: Cambridge University Press, 1955; KUHN, Steven T. Reflections on Ethics and Game Theory. *Synthese*, v. 141, n. 1, p. 1-44, 2004; SMITH, John Maynard. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

[458] KRÜGER, L. Ethics according to nature in the age of evolutionary thinking. *Grazer Philosophische Studien*, v. 30, p. 25-42, 1987.

[459] VERBEEK; MORRIS, 2009.

[460] LEYHAUSEN. The biological basis of ethics and morality. *Science, Medicine & Man*, v. 1, p. 215-235, 1974. RICHARDS, R. A defense of evolutionary ethics. *Biology & Philosophy*, v. 1, p. 165-293, 1986.

[461] ALEXANDER, R. D. *The biology of moral systems*. New York: Aldine de Gruyter, 1987; AYALA, F. J. The biological roots of morality. *Biology & Philosophy*, v. 2, p. 235-252, 1987; BISCHOF, N. On the phylogeny of human morality. *In*: STENT, G. S. (ed.). Morality as a biological phenomenon. Berlin: Springer, 1978. p. 53-73; CAMPBELL, D. T. On the conflict between biological and social evolution and between psychology and moral tradition. *American Psychologist*, v. 30, p. 1103-1126, 1975; CELA-CONDE, C. *On genes, Gods and tyrants*: the biological causation of morality. Dordrecht: Reidel, 1987.

se eles podem ser racionalizados e modelados, cremos que possam ser computáveis em algoritmos.

Compreender como emerge a ideia do certo e errado nos humanos[462] e nas máquinas é um problema mais complexo e muito além dos propósitos deste trabalho. Contudo, pode-se encontrar na literatura diversos estudos sobre a *emergência* da moralidade, na natureza, nos seres humanos e, por que não, nas máquinas.

Um dos caminhos está no uso de algoritmos de aprendizado associativo, tendo como referência o utilizado para o aprendizado de jovens e crianças. A estratégia consistiria em dotar os agentes morais artificiais de um modelo de treinamento[463], envolvendo o *feedback* da aceitabilidade moral de suas ações[464], ou seja, mediante mecanismos de aprovação ou desaprovação, punição ou premiação. Os estudos não diferenciaram o modo de implementação, se por via de redes neurais ou outra forma de aprendizado de máquina.

O propósito, ao final do experimento, está na capacidade de superar o *Teste Moral de Turing* comparativo (cMTT). Contudo, os estudos com algoritmos morais evolucionários demonstram a emergência de comportamentos morais que não possuem justificativas para as suas ações, exatamente o cerne do comportamento de um agente moral artificial completo (*full moral agent*) é a sua capacidade em justificar as suas ações. As regras surgidas se limitavam à preocupação com a sobrevivência, sejam por cooperação ou ação individual. A busca do interesse individual do agente apenas aparenta a existência de uma moralidade[465], não comprovando a emergência de um agente moral artificial completo (*full moral agent*).

Os estudos não provocam uma demonstração negativa da possibilidade de algoritmos evolucionários ou da emergência de um agente moral artificial completo (*full moral agent*). Apenas relatam as dificuldades atuais dos modelos computacionais existentes. Antes de desmotivar as pesquisas futuras, eles apresentam uma nova agenda, ousada e promissora, de pesquisas em direção a descobertas[466].

Talvez as máquinas possam nos surpreender (*take us by surprise*). Mesmo que os algoritmos sejam exaustivos ao infinito, capazes de cobrir todas as possibilidades, todas as consequências; talvez, alertava *Turing*, as máquinas possam, ainda assim, nos surpreender.

---

[462] DE WAAL, Frans. *Good Natured*: the origins of right and wrong in humans and other animals. Cambridge, MA: Harvard University Press, 1996.
[463] DENNETT, Daniel C. True Believers: The Intentional Strategy and Why it Works. *In*: HAUGELAND, John; *Mind design II*: philosophy, psychology, artificial intelligence. Massachusetts: Massachusetts Institute of Technology, 1997.
[464] ALLEN; VARNER; ZINSER, 2000,
[465] ALLEN; VARNER; ZINSER, 2000.
[466] ALLEN; VARNER; ZINSER, 2000.

Enfim, a partir do exame de diversas questões conexas (linguagem, imaginação e crenças), parece claro que há a possibilidade de superação da *Lei de Lovelace*. Não só é possível, como é plausível e, talvez, provável, conforme os rápidos avanços em ciência da computação e teoria dos algoritmos. Uma teoria ética responsável deve considerar a possibilidade concreta do surgimento de *agentes morais artificiais completos* (*full moral agent*) e todas as consequências[467] de fenômeno divisor na história da humanidade.

[467] O Parlamento Europeu aprovou uma resolução determinando que a IA deve ser antropocêntrica e sob controle humano, especialmente, no caso de aprendizado de máquinas evolutivos. (European Parliament. 2019-2024, 8 out. 2020. *Plenary sitting with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies* [2020/2012(INL)]. Committee on Legal Affairs Rapporteur: Ibán García del Blanco. Disponível em: https://www.europarl.europa.eu/doceo/document/A-9-2020-0186_EN.pdf. Acesso em: 02 nov. 2020 às 00:12).

3 CONCLUSÕES

1. A Inteligência Artificial (IA) não é apenas um desenvolvimento tecnológico revolucionário, mas demarcador na história da humanidade. Seu impacto será gigantesco na economia, na política e no direito. A preocupação com o desenvolvimento dessa nova ferramenta se expande em toda as áreas. Diversos mecanismos têm sido pensados para reduzir ou mesmo eliminar as consequências danosas do seu mau uso. A presente Tese não versa sobre os efeitos e consequências do uso ou mau uso da IA.

2. O objeto da presente tese é verificar e assumir a possibilidade filosófica do surgimento de um agente moral artificial.

2. O conceito filosófico de autômato surgiu inicialmente em Aristóteles, em sua obra "Metafísica", na expressão *ta automata tôn thaumatôn* (*como nos fantoches*). Ele diferencia o humano dos animais e, provavelmente, diferenciaria dos autômatos pelas seguintes razões: i) os humanos se movimentam por deliberação, em direção a um fim; ii) possuem o dom da fala; iii) são agentes morais e iv) possuem comunicação.

3. A teoria moderna dos autômatos principia com *René Descartes*. Ele responde que deveríamos utilizar testes para identificar a presença de um "indivíduo real". O primeiro deles seria o uso da linguagem, ou seja, a capacidade de resposta articulada a tudo o que seja dito na presença deste ser, como ele consegue declarar com competência seus pensamentos. O segundo teste seria a incapacidade dos autômatos de ter um conhecimento prático ou abrangente.

4. *Lady Ada Lovelace* irá, elegante e rigorosamente, explicar o funcionamento das máquinas virtuais e afirmar que a máquina somente automatiza procedimentos, tal como faz um tear. Ele não cria padronagens novas, nem desenha novas e sublimes formas. Enfim, automatiza um procedimento. Não existiria "inteligência artificial", apenas um uso inteligente das máquinas automáticas, denominado de *Argumento de Lady Lovelace*.

5. *Allan Turing* irá produzir a primeira resposta consistente às objeções cartesianas à possibilidade de as máquinas utilizarem competências linguísticas. A resposta de Turing aparecerá no revolucionário artigo publicado sob o título "*Computing Machinery and Intelligence*", publicado na *Revista Mind* em 1950. O texto principia com a ambiciosa pergunta: podem as máquinas pensar? (*Can machines thinking*?). *Turing* sugere um artifício denominado "jogo da imitação". Para ele este substituiria a tradicional pergunta "podem as máquinas pensar?", por "podem as máquinas imitar com sucesso o comportamento humano?", a ponto de se tornarem indistinguíveis. A superação do *Teste de Turing* é o segundo desafio importante para o surgimento de agentes morais autênticos.

6. Coube a *John Searle* (1932-) elaborar a mais importante e bem formulada objeção ao TT. O autor irá distinguir, corretamente, entre IA forte (*strong*) e IA fraca (*weak*), conforme as suas funções. A IA forte sugere a possibilidade de máquinas que performam competências próprias de um ser humano, ou seja, não apenas aparentam como possuem igualmente todas as competências humanas, inclusive a consciência. *Searle* apresenta duas proposições encadeadas: 1) a intencionalidade nos seres humanos (e animais) é produto da características causais do cérebro (*causal features of the brain*) e 2) instanciar um programa de computador não é por si só uma condição suficiente de intencionalidade (*intentionality*). A conclusão de *Searle* é a de que toda tentativa de criar intencionalidade artificialmente (*strong* AI) deve duplicar os poderes causais do cérebro humano e não simplesmente elaborar um programa computacional.

7. *Searle*, partindo desses pressupostos, irá afirmar a sua tese sobre a *conexão entre consciência e intencionalidade*: *somente seres conscientes possuem intencionalidade e qualquer ato inconsciente intencional é no mínimo potencialmente consciente*. O desafio, expresso pelo autor, está no famoso e criativo argumento da *Sala Chinesa.* Toda tentativa de afirmar a possibilidade de um agente artificial autêntico precisa demonstrar que esses podem deter consciência e intencionalidade.

8. Os argumentos de *Searle* foram tão desconcertantes que implicaram sucessivas respostas e tentativas de superação da tese da Sala Chinesa. Diversas sugestões foram apresentadas para superar a objeção à IA forte. Todas muito criativas e sucessivamente descartadas. Algumas reformulações, contudo, começaram a atrair a atenção de filósofos e cientistas, por exemplo, a tese de um *autômato* (*robot*) com sensores para interagir com o meio ambiente, tal como ver, ouvir e mesmo sentir, superaria o obstáculo da conexão entre mente e ambiente por meio de experiências sensoriais únicas por parte do autômato[468]. A resposta propõe uma mudança de uma tese computacional da mente para uma tese robótica da mente. A inteligência artificial deixaria de ser um programa instalado no cérebro e passaria a ser entendida como um sistema incorporado no cérebro (*embodied* AI). Após sucessivas investidas, as objeções contra a tese da Sala Chinesa parecem ter tomado corpo e vislumbrado a possibilidade de que talvez as máquinas pudessem pensar e, incrivelmente, adquirir consciência.

9. A ideia de possibilidade filosófica cinge-se à noção de uma ordem de coisas consistente, ou seja, que não viole as regras lógicas decorrentes da aplicação do princípio da contradição. É possível refutar algumas das principais objeções apresentadas por *Turing* contra

[468] PIEK, [*s.d.*].

o argumento de que as máquinas podem pensar, mais propriamente: a objeção da consciência, das imperfeições, da intencionalidade, da limitação algorítmica (Argumento de Ada Lovelace), biológica e teológica. As máquinas podem adquirir, teoricamente, consciência e intencionalidade. Suas imperfeições não se constituem em uma limitação absoluta nem mesmo os limites biológicos se colocam como barreiras definitivas.

10. A presente tese parte da afirmativa da possibilidade de superação do Teste de Turing, da Sala Chinesa e do Teste de Ada Lovelace, sendo possível a emergência de um autêntico agente artificial moral, com deliberações intencionais em perspectiva de primeira pessoa. Parte-se da aceitação da possibilidade de um código computacional capaz de dar origem à emergência.

11. Existem três sentidos possíveis para falarmos de uma ética artificial, e esta pode ser a ética *aplicada* à inteligência artificial (IA), *decorrente* de sua aplicação ou da *própria* da IA. No primeiro caso, tratamos dos limites e das diretrizes éticas para pesquisa e desenvolvimento da IA. Poder-se-ia questionar, nesse campo, quais são os princípios que devem nortear as pesquisas sobre autômatos, robôs e algoritmos. No segundo caso, trata-se dos desafios éticos decorrentes da *aplicação* da IA. A presente Tese verifica a possibilidade de uma ética artificial, e a aceita como possível.

12. O objeto da presente tese é investigar a possibilidade filosófica de uma ética artificial, como decorrente da vontade e racionalidade *própria* de um sujeito artificial. *A inteligência artificial como sujeito moral*. Uma tarefa é analisar a inteligência artificial como *objeto* ou artefato humano que pode ou não ser bem utilizado no *agir humano*; outro problema muito distinto é tratar da possibilidade de um agir por parte de um sujeito artificial.

13. A realidade tem apresentado, contudo, novos artefatos que reproduzem mecanismos de escolhas éticas, deliberadas, arquitetadas, desenhadas e implementadas *ex ante* para uso e produção de consequências, conforme o modelo pensado por humanos. Essas máquinas seriam "agentes éticos implícitos", nos quais as máquinas são programadas para suportar comportamentos éticos ou evitar os comportamentos antiéticos. Contudo, a arquitetura, o desenho, o algoritmo, as funções, a implementação e o uso são fruto de uma mente humana, que projeta a sua racionalidade por extensão em uma máquina.

14. Um agente ético, seja humano ou artificial, deve sê-lo por características próprias e não agir conforme uma programação externa predeterminada. A ética deve ser interna e não externa ao autômato. Ela deve ser fruto de um agir em primeira pessoa e nunca como instrumento de um terceiro, ou seja, deve garantir a sua natureza subjetiva e não ser objeto da

ação de outrem. Essa é a diferença entre uma máquina lógica e uma máquina moral. De um lado, uma possui autonomia mecânica e a outra, autonomia moral.

15. Uma máquina, para pleitear o posto de agente artificial deveria ser racional, possuir uma existência incomunicável (*incommunicabilis existentia*) e liberdade. Ou seja, deve agente artificial possuir experiência singular, em primeira pessoa e dotada de liberdade. A moderna teoria computacional, apesar de não vislumbrar em um horizonte próximo tal possibilidade, não afirma existir uma impossibilidade, técnica ou abstrata, *ab initio* para o surgimento de um agente artificial autêntico.

16. Um sujeito artificial poderá ser um sujeito moral, dado que dotado de liberdade, racionalidade e autorregulado. Poderá elencar fins para conduzir a sua ação. Não se trata apenas de uma máquina possuidora de um algoritmo moral implementado originariamente de modo externo. Ele poderá possuir as características decorrentes da liberdade.

17. O interesse pela *ética das virtudes* decorre do acelerado e acentuado grau de autonomia dos agentes artificiais. Não somente modelos de robôs cada vez mais sofisticados se sucedem, com novas e mais surpreendentes capacidades reais e possíveis. Novos dispositivos dotados de autonomia e sistemas inteligentes embarcados se multiplicam em formas, tamanho e funcionalidades.

18. Um modelo proposto e com crescente aceitação, e que demonstra essa possibilidade computacional, é o de uma moralidade que se constrói *de baixo para cima* (*bottom-up*), e nesse caso o sistema pode passar a adquirir capacidades morais de modo independente. Esse modelo se aproxima da *ética aristotélica das virtudes*. Outra forma possível é a união de um modelo computacional de piso, com modelos fundados na deontologia, com a formulação mais geral de deveres e máximas. De uma outra forma, demonstra-se que pelo menos em um caso é possível a construção de um modelo de moralidade artificial viável e autônomo.

19. Parte-se do entendimento de que estatuto moral da humanidade pode não ser uma condição excepcional da espécie humana. Não há uma "agência moral excepcionalmente humana" (*essentially human agency*), estabelecida em bases ontológicas ou *a priori.* O surgimento, de modo controlado ou espontâneo de uma linguagem artificial compreensível ou não à racionalidade humana, permite aceitar que poderiam surgir, igualmente, regras morais próprias desses agentes artificiais, compreensíveis ou não, para os programadores humanos.

20. Admite-se a tese de que há similaridade entre a *moralidade humana e a artificial*, de tal modo que há teoricamente a possibilidade de condutas morais comparáveis, entre humanos e agentes artificiais em situações relevantes. O Teste Moral comparativo de *Turing* (cMTT), entre os agentes, esbarra ainda hoje em dificuldades tecnológicas, intransponíveis,

que, talvez, sejam no futuro superadas pelo desenvolvimento exponencial dos sistemas autônomos.

21. Pode-se afirmar que o desenvolvimento da modelagem matemática, da compreensão neuromatemática do cérebro, dos avanços em instrumentos de análise por imagem demonstram a possibilidade de que os *modelos computacionais de consciência podem ser suficientes para se compreender o mecanismo da consciência*. Os estudos de *Chalmers* servirão para defender que a tese de que a "explanação computacional" (*computational explanation*) nos permite uma linguagem adequada (*perfect language*) para a compreensão da organização causal dos processos cognitivos. Por sua vez, a tese da "suficiência computacional" (*computational sufficiency*) se sustenta, dado que todas as implementações computacionais conseguem replicar adequadamente a estrutura da mente.

22. Não há objeção instransponível à possibilidade de os agentes morais artificiais possuírem emoções artificiais. Os *sentimentos artificiais*, para a modelagem computacional, são considerados como *motivadores primitivos para a seleção de ações* (*primitive motivators*). Eles são representados como nós (*nodes*) de memória perceptual, em que cada nó representa sua própria valência, positiva ou negativa, e segundo uma determinada graduação. Considerando que existe similaridade entre compreender emoções e poder senti-las, é possível afirmar que elas possam se desenvolver. Afinal, seres menos complexos podem igualmente sentir o sofrimento. A igualdade em sofrer permitiria a possibilidade do surgimento de autômatos racionais, morais e emocionais

23. Os estudos sobre IA mudaram o foco da modelagem de regras morais gerais e abstratas, aplicáveis a todas as situações, para o aprendizado pela experiência. A imitação a partir de exemplos morais deve ser o conceito central em ética da virtude artificial. A ideia de *self-improvement* é primordial nesse conceito, superando a noção de agente dotado de corpo de regras pronto, acabado e aplicável a qualquer situação.

24. A conclusão a que diversos cientistas de programação chegaram é que um modelo de agência artificial fundado em *machine learning*, combinado com a *ética da virtude*, é um caminho natural, coeso, coerente, integrado e "bem costurado" (*seamless*). Assim, existe uma resposta coerente, consistente e bem fundamentada que *indica que não é provada a impossibilidade de um agente moral artificial autêntico.*

25. É afastada a objeção teológica apresentada por *Turing* para refutar a possibilidade da existência de *agentes morais artificiais completos* (*full* AMAs), bem como a objeção, denominada de *Tese de Lovelace*, da impossibilidade de autênticos agentes artificiais emergirem de algoritmos, dada a possibilidade de algoritmos evolucionários.

26. O emergentismo surge hodiernamente com a teoria dos sistemas complexos, a neurociência e a filosofia da mente. *David Chalmers* (2006) afirmava que o *emergentismo fraco* (*weak emergence*) é comum e compatível com as noções de auto-organização, complexidade e não linearidade. Trata-se de uma noção epistemológica e não metafísica, sendo definida em termos de *imprevisibilidade* (*unpredictability or unexpectedness*). Assim, dadas as características e propriedades das partes de baixo nível ou fundamentais, podem emergir propriedades imprevisíveis.

27. O surgimento emergente de comportamentos multiagentes foi comprovado em diversos estudos. Foi constatada a possibilidade de emergência de linguagem composicional fundamentada (*grounded compositional language*) para atingir finalidades e objetivos em populações de multiagentes. Não se pode concluir pela impossibilidade de emergência de comportamentos comunicacionais em algoritmos, que utilizem o *Aprendizado de Máquina por Reforço em Multiagentes* (*multi-agent reinforcement learning* – MARL). Pelo contrário, os estudos tendem a avançar em complexidade, profundidade e ousadia nas possibilidades computacionais, expandindo o caminho para a superação do *Teste de Lovelace*.

28. A ideia de que os sistemas morais devem ser considerados como sistemas evolucionários é partilhada por outros autores. E, assim, se eles podem ser racionalizados e modelados, cremos que possam ser computáveis em algoritmos.

29. Os estudos não demonstram a impossibilidade de algoritmos evolucionários ou da emergência de um *agente moral artificial completo* (*full moral agent*). Apenas relatam as dificuldades atuais dos modelos computacionais existentes. Enfim, a partir do exame de diversas questões conexas (linguagem, imaginação e crenças), parece claro que há a possibilidade de superação da *Lei de Lovelace*. Não só possível como é plausível e, talvez, provável, conforme os rápidos avanços em ciência da computação e teoria dos algoritmos.

30. Por fim, uma teoria ética responsável deve considerar a possibilidade concreta do surgimento de *agentes morais artificiais completos* (*full moral agent*) e todas as consequências desse fenômeno divisor na história da humanidade.

REFERÊNCIAS

AESOP'S FABLES. A new translation by Laura Gibbs. New York: Oxford University Press, 2002. (World's Classics). Disponível em: http://mythfolklore.net/aesopica/oxford/514.htm. Acesso em: 23 maio 2020 às 23:56.

AGAR, N. How to treat machines that might have minds. *Philos Technol.*, 2019. Disponível em: https://doi.org/10.1007/s13347-019-00357-8. Acesso em: 08 jun. 2020.

ALEXANDER, R. D. *The biology of moral systems*. New York: Aldine de Gruyter, 1987.

ALLEN, C.; SMIT, I.; WALLACH, W. Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, v. 7, n. 3, p. 149-155, 2005.

ALLEN, C.; VARNER, G.; ZINSER, J. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 12, n. 3, p. 251-261, 2000.

ALMEIDA, J. R. Pessoa e relação em João Duns Scotus. *Enrahonar, Supplement Issue*, p. 79-87, p. 84, 2018. Disponível em: https://ddd.uab.cat/pub/enrahonar/enrahonar_a2018nsupissue/enrahonar_a2018nSupplp79.pdf. Acesso em: 11 jun. 2020 às 11:04.

ALMEIDA, R. T. de. Evolução histórica do conceito de pessoa – enquanto categoria Ontológica. *Revista Interdisciplinar de Direito*, [*S.l.*], v. 10, n. 1, p. 229, out. 2017. ISSN 2447-4290. Disponível em: http://revistas.faa.edu.br/index.php/FDV/article/view/202. Acesso em: 08 jun. 2020.

ANDERSON, M.; ANDERSON, S. L. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, v. 28, n. 4, p. 15-27, 2007. p. 17.

ANDERSON, M.; ANDERSON, S. L. The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines*, v. 17, p. 1-10, 2007. Disponível em: https://link.springer.com/article/10.1007%2Fs11023-007-9053-7. Acesso em: 05 jun. 2020 às 23:07.

ANDERSON, S. L. Asimov's "three laws of robotics" and machine metaethics. *AI & Soc*, v. 22, p. 477-493, 2008.

ANGIONI, L. Aristotle's Modal Syllogistic. *Ancient Philosophy*, v. 38, p. 211-216, 2018.

ANGIONI, L. As relações entre fins e meios e a relevância moral da *Phronesis* na ética de Aristóteles. *Revista Filosófica de Coimbra*, v. 18, p. 185-203, 2009.

ANGIONI, L. Em que sentido a virtude é mais exata que a técnica? Notas sobre Ethica Nicomachea 1106b 14-16. *Revista Dissertatio de Filosofia*, v. 29, p. 43, 2009.

ANGIONI, L. Explanation and method in Eudemian ethics I.6. *Revista Archai: Revista de Estudos sobre as Origens do Pensamento Ocidental*, p. 191-229, 2017.

ANGIONI, L. Notas sobre a definição de virtude moral em Aristóteles (EN 1106b 36- 1107a 2). *Revista de Filosofia Antiga*, v. 3, p. 1-17, 2009. (USP. Ed. português)

ANGIONI, L. Phronesis e virtude do caráter em Aristóteles: Comentários à Ética a Nicômaco VI. *Dissertatio*, v. 34, 2011, p. 303-345.

ANNAS, J. Learning virtue rules: the issue of thick concepts. *In:* ANNAS, J.; NARVAEZ, D.; SNOW, N. E. (ed.). *Developing the virtues*: integrating perspectives. New York: Oxford University Press, 2016.

ANSCOMBE. Modern Moral Philosophy. *Philosophy*, v. 33, n. 124, p. 1-19, January 1958.

AQUINO, T. de. *Suma Teológica*, 1, 29, 1. Disponível em: https://sumateologica.files.wordpress.com/2017/04/suma-teolc3b3gica.pdf. Acesso em: 08 jun. 2020 às 01h02.

ARAÚJO, L. B. de. A filosofia natural de Thomas Hobbes: a composição das paixões humanas. *Controvérsia*, São Leopoldo, v. 14, n. 3, p. 75-96, set.-dez. 2018.

ARISTÓTELES. De generatione animalium I, 22, 730, b 11-23. Trad. Arthur Platt. *In*: ROSS, W. D. (ed.). *The works of Aristotle translated into English*. Oxford: Clarendon, 1963.

ARISTÓTELES. *Ética a Nicômaco*. Trad. Leonel Vallandro e Gerd Bornheim da versão inglesa de W. D. Ross. São Paulo: Nova Cultural, 1991. Livro VI, 8.

ARISTÓTELES. *Sobre a Alma*. I, 3, 407b 20-24. Lisboa: Imprensa Nacional Casa da Moeda, 2020.

ARKIN, R. C. *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press, 2009.

ARNOLD, T.; SCHEUTZ, M. Against the Moral Turing Test: accountable design and the moral reasoning of autonomous systems. *In*: SCHEUTZ, Matthias (dir.). *Hrilab*. Medford, 2016. Disponível em: https://hrilab.tufts.edu/publications/arnoldscheutz16mtt.pdf. Acesso em: 02 jul. 2020 às 22:48.

ARTOSI, A.; PIERI, B.; SARTOR, G. *Leibniz*: LogicoPhilosophical Puzzles in the Law. Philosophical Questions and Perplexing Cases in the Law. Heildelberg: Springer, 2013.

AUDI, R. Intuitions, intuitionism, and moral judgment. *In:* AUDI, R. *Reasons, Rights, and Values*. Cambridge: Cambridge University Press, 2015.

AVRAMIDES, A. Descartes and Other Minds. *Teorema*, v. XVI/1, p. 27-46, 1996. Disponível em: file:///Users/caliendo/Downloads/Dialnet-DescartesAndOtherMinds-4244359.pdf. Acesso em: 28 maio 2020 às 12:05.

AYALA, F. J. The biological roots of morality. *Biology & Philosophy*, v. 2, p. 235-252, 1987.

AZAMBUJA, C. C. *Prometeu*: a sabedoria pelo trabalho e pela dor. *Archai*, n. 10, jan.-jul 2013.

AZEEM M. M. *et al.* Emotions in Robots. *In*: CHOWDHRY B. S. *et al.* (ed.). *Emerging trends and applications in information communication technologies*. Communications in Computer and Information Science. Berlin: Springer, 2012. v. 281.

BABBAGE, C. *Sketch of the analytical engine invented*. [1942] Disponível em: http://www.fourmilab.ch/babbage/sketch.html. Acesso em: 29 maio 2020 às13:07.

BACON, F. New Atlantis. *Project Gutenberg*. 2020. Disponível em: http://www.gutenberg.org/ebooks/2434. Acesso em: 05 jun. 2020 às 01:24.

BATES, D. Cartesian Robotics. *Representations*, v. 124, n. 1, p. 43-68, Fall 2013.

BATTERMAN, R. W. Is Weak Emergence Just in the Mind? *Minds and Machines*, v. 18, p. 443-459, 2008.

BATTERMAN, R. W. *The devil in the details*: asymptotic reasoning in explanation, reduction, and emergence. Oxford Studies in Philosophy of Science. Oxford: Oxford University Press, 2001.

BEDAU, M. A. Weak Emergence. *In*: TOMBERLIN, J. (ed.). *Philosophical perspectives*: mind, causation and world. Malden, MA: Blackwell, 1997. v. 11.

BEDAU, M. A.; HUMPHREYS, Paul (ed.). *Emergence*: contemporary readings in philosophy and science. London, UK: MIT Press, 2007.

BENTHAM, J. *An Introduction to the Principles of Morals and Legislation (1781)*. Batoche Books. Disponível em: https://socialsciences.mcmaster.ca/econ/ugcm/3ll3/bentham/morals.pdf. Acesso em: 20 jun. 2020 às 15:21.

BERBERICH, N.; DIEPOLD, K. *The virtuous machine*: old ethics for new technology? Munich: Munich Center for Technology in Society, 2018. Disponível em: https://arxiv.org/pdf/1806.10322.pdf. Acesso em: 21 jun. 2020 às 21:50.

BERRYMAN, S. Ancient automata and mechanical explanation. *Phronesis-A Journal for Ancient Philosophy*, v. 48, n. 4, p. 344-369, 2003.

BICKHARD, M.; CAMPBELL, D.T. Emergence. *In*: ANDERSEN, P.B. *et al.* (ed.). *Downward causation*. Aarhus: Aarhus University Press, 2000.

BIGMAN, Y. E.; GRAY, K. People are averse to machines making moral decisions. *Cognition*, v. 181, p. 21-34, 2018 . Disponível em: https://doi.org/10.1016/j.cognition.2018.08.003.

BISCHOF, N. On the Phylogeny of Human Morality. *In*: STENT, G. S. (ed.). *Morality as a biological phenomenon*. Berlin: Springer, 1978.

BJORK, Russell C. Artificial Intelligence and the Soul. *Perspectives on Science and Christian Faith*, v. 60, n. 2, p. 95-102, June 2008.

BODEN, M. A. *AI*: its nature and future. New York: Oxford University Press, 2016.

BODEN, M. A. *Mind as machine*: a history of cognitive science. Clarendon Press, Oxford, 2006.

BOÉCIO. Escritos (OPUSCULA SACRA). Tradução, introdução, estudos introdutórios e notas Juvenal Savian Filho. Prefácio de Marilena Chauí. São Paulo: Martins Fontes, 2005.

BORDINI, R. H., HÜBNER, J. F.; WOOLDRIDGE, M. *Programming Multi-Agent Systems in Agent Speak Using Jason.* John Wiley & Sons, 2007.

BORDINI, R. H.; MOREIRA A. F. Proving BDI Properties of agent-oriented programming languages. *Ann. Math. Artif. Intell*., v. 42, n. 1-3, p. 197-226, 2004.

BOSTROM, N.; YUDKOWSKY, E. The ethics of artificial intelligence. *In*: FRANKISH, K.; RAMSEY, W. M. (ed.). *The Cambridge Handbook of artificial intelligence*. Cambridge: Cambridge University Press, 2014. p. 316-334. Disponível em: https://intelligence.org/files/EthicsofAI.pdf. Acesso em: 06 jun. 2020.

BOWE, G. Alexander's Metaphysics commentary and some scholastic understandings of automata. *Schole. Ancient Philosophy and the Classical Tradition*, v. XIV, 2020, Issue 1. Disponível em: https://nsu.ru/classics/schole/14/schole-14-1.pdf. Acesso às 00:05.

BRAITHWAITE, R. B. *Theory of games as a tool for the moral philosopher*. Cambridge: Cambridge University Press, 1955.

BRATMAN, M. E. Intentions, plans, and practical reason. CSLI, 1987.

BRINGSJ, S. God, souls, and Turing: in defense of the theological objection to the Turing test. *Kybernetes*, v. 39, n. 03, p. 414-422, 2010, p. 417. Disponível em: http://kryten.mm.rpi.edu/SB_theo_obj_tt_offprint.pdf. Acesso em: 18 jul. 2020 às 20:41

BRITO, A. R. T. *As abelhas egoístas*: vício e virtude na obra de Bernard Mandeville. 2006. Tese (doutorado em Filosofia) – Programa de Pós-Graduação em Filosofia. Universidade de São Paulo, São Paulo, 2006. p. 128. Disponível em: http://livros01.livrosgratis.com.br/cp077816.pdf. Acesso em: 20 jun. 2020 às 18:03.

BROAD, C.D. *The mind and its place in nature*. London: Routledge and Kegan Paul, 1925. p. 71-72.

BRONI, L. de. A filosofia natural de Thomas Hobbes: a composição das paixões humanas. *Controvérsia*, São Leopoldo, v. 14, n. 3, p. 75-96, set.-dez. 2018.

BROWN, W. S.; MURPHY, N.; MALONY, H. N. *Whatever Happened to the Soul?* Minneapolis: Fortress, 1998.

BUNGE, M. Emergence and the Mind. *Neuroscience*, v. 2, p. 501-509, 1977.

CAMPBELL, D. T. On the conflict between biological and social evolution and between psychology and moral tradition. *American Psychologist*, v. 30, p. 1103-1126, 1975.

CARDOZO CIACCO, F. Sobre o monstro, a natureza e a origem: uma releitura de Frankenstein ou o Prometeu moderno. *Outra Travessia*, Florianópolis, n. 22, p. 161-174, ago. 2016. ISSN 2176-8552. Disponível em: https://periodicos.ufsc.br/index.php/Outra/article/view/2176-8552.2016n22p161/34652. Acesso em: 02 jul. 2020.

CELA-CONDE, C. *On genes, Gods and tyrants*: the biological causation of morality. Dordrecht: Reidel, 1987.

CHALMERS, D. A Computational Foundation for the Study of Cognition. *Journal of Cognitive Science*, Seoul Republic of Korea, p. 323-357, 2011,

CHALMERS, D. Strong and Weak Emergence. *In*: DAVIES, P.; CLAYTON (ed.). *The re-emergence of emergence*. New York: Oxford University Press, 2006.

CHIAROTTINOI, Z. R.; FREIRE, J.-J. O dualismo de Descartes como princípio de sua Filosofia Natural. *Estudos Avançados*, v. 27, n. 79, p. 158, 2013.

CHRISTIANSEN, M. H.; KIRBY, S. Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, v. 7, n. 7, p. 300-307, 2003.

CHURCHLAND, P. S.; WINKIELMAN, P. Modulating social behavior with oxytocin: How does it work? What does it mean? *Hormones and Behavior*, v. 61, n. 3, p. 392-399. March 2012. Disponível em: https://www-sciencedirect.ez94.periodicos.capes.gov.br/science/article/pii/S0018506X11002807?via%3Dihub. Acesso em: 30 jun. 2020 às 00:51.

COLE, D. The Chinese Room Argument. *In*: ZALTA, E. N. (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). Disponível em: https://plato.stanford.edu/archives/spr2020/entries/chinese-room/. Acesso em: 31 jul. 2020 às 00:04.

CONEE, E. A. The Preface Paradox. *In*: DANCY, J.; SOSA, E.; STEUP, M. (ed.). *Companion to Epistemology*. 2. ed. Malden, MA: Wiley-Blackwell, 2009. p. 604-605.

CROCKETT, L. AI Ethics: the thin line between computer simulation and deception. *In:*

CULLETON, A. S. O conceito de pessoa em Ricardo de São Vitor. *Problemata – Rev. Int. de Filosofia*, v. 2, n. 1, p. 11-26, 2011.

DE WAAL, F. *Good natured*: the origins of right and wrong in humans and other animals. Cambridge: Harvard University Press, 1996.

DEJEANNE, S. de M. *A fundamentação da moral no limite da razão em Kant*. 2020. Tese (Doutorado em Filosofia) – Programa de Pós-Graduação em Filosofia, PUCRS, Porto Alegre, 2020. Disponível em: https://doi.org/http://tede2.pucrs.br/tede2/handle/tede/2783. Acesso em: 15 dez. 20 às 00:42.

DENNETT, D. C. True believers: the intentional strategy and why it works. *In*: HAUGELAND, J.; *Mind Design II*: philosophy, psychology, artificial intelligence. Massachusetts: Massachusetts Institute of Technology, 1997.

DESCARTES, R. *Discurso do Método*. São Paulo: Martins Fontes, 2009.

DESCARTES, R. *Meditações sobre a Filosofia Primeira*. Campinas: Unicamp, 2004. (Coleção Multilíngues de Filosofia Tradução Fausto Castillo.)

DI MARZO, G.; GLEIZES, M.-P.; KARAGEORGOS, A. Self-Organisation and Emergence. *MAS: An Overview Informatica*, v. 30, p. 45-54, 2006.

DI NAPOLI, R. B. Conflitos de deveres e a casuística na filosofia moral de Kant. *Studia Kantiana*, 11, p. 178-200.

DI NAPOLI, R. B. O intuicionismo moral e os dilemas morais. *Dissertatio*, UFPel, v. 35, p. 79-98, 2012.

DIGNUM, V. *Artificial Intelligence*: foundations, theory, and algorithms. Cham: Springer, 2019.

DIGNUM, V. *Responsible Artificial Intelligence*: How to develop and use ai in a responsible way. Cham: Springer, 2019.

DONAGAN, A. Moral dilemmas, genuine and spurious: a comparative anatomy. *In:* MASON, H. E. (org.). *Moral dilemmas and moral theory*. Oxford: Oxford University Press, 1996.

DOUGHERTY, C. *Prometheus*. London/New York: Taylor & Francis, 2006.

DOUGHERTY, M. V. *Moral dilemmas in medieval thought*: from Gratian to Aquinas. Cambridge: Cambridge University Press, 2011.

EINEVOL, T. *et al.* The scientific case for brain simulations. *Neuron.*, v 102, Issue 4, pp. 735-744, 22 May 2019.

EMMECHE, C.; KØPPE, S.; STJERNFELT, F. *Levels, emergence, and three versions of downward causation*. Disponível em: http://www.nbi.dk/~emmeche/coPubl/2000d.le3DC.v4b.html. Acesso em: 24 ago. 2020 às 15:04.

ENGBERG-PEDERSEN, T. Aristotle's Theory of Moral Insight. Review by Alfred R. Mele. *The Philosophical Review*, v. 94, n. 2, Apr. 1985, pp. 273-275.

ETZIONI, A.; ETZIONI, O. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, v. 21, n. 4, p. 403-418, 2017.

FAUSTO, J. A cadela sem nome de Descartes: notas sobre vivissecção e mecanomorfose no século XVII. 44. *Doispontos*, Curitiba, v. 15, n. 1, p. 43-59, abril de 2018.

FELDHAUS, C. De Schopenhauer a ética de virtudes contemporânea. *Revista Guairacá*, v. 29, n. 2, p. 46, 2013.

FERRER, V. Stéphan Geonget. La notion de perplexité à la Renaissance. *Revue de l'histoire des religions*, v. 3, 2008. Disponível em: http://journals.openedition.org/rhr/6763. Acesso em: 29 jun. 2020 às 23:36.

FILINGERI, D. *et al.* Why wet feels wet? A neurophysiological model of human cutaneous wetness sensitivity. *J Neurophysiol.*, v. 112, Issue 6, p. 1457-1469, September 2014.

FLORIDI, L. *The ethics of information*. New York: Oxford University Press, 2013.

FLORIDI, L.; SANDERS, J. On the morality of artificial agents. *Minds and Machines*, v. 14, p. 349-379, 2004.

FODOR, J. A. Searle on what only brains can do. *The Behavioral and Brain Sciences*, p. 431-432, 2010.

FONSECA, E. G. da. *A Fábula das Abelhas*. Braudel Papers. The Tinker Foundation & Champion Papel e Celulose, 1994.

FOOT, P. Moral dilemmas revisited. *In:* FOOT, P. *Moral Dilemmas*. Oxford: Oxford University Press, 2002.

FOOT, P. Morality as a system of hypothetical imperatives. *The Philosophical Review*, v. 81, n. 3, p. 305-316, jul. 1972.

FOOT, P. The problem of abortion and the doctrine of double effect. *Oxford Review*, v. 5, p. 5-15, 1967.

FREEMAN, W. J. *Nonlinear Brain Dynamics and Intention According to Aquinas,* Seattle, AI2, p. 232. Disponível em: https://www.semanticscholar.org/paper/Nonlinear-Brain-Dynamics-and-Intention-According-to-Freeman/1058e99a0036f6f9b9a76a9b7dc59e6b16cf736a. Acesso em: 31 jul. 2020 às 00:51.

FREY, J. A. Against autonomy: why practical reason cannot be pure. *Manuscrito*, v. 41, n. 4, p. 159-93, Dec. 2018. Disponível em: https://doi.org/10.1590/0100-6045.2018.v41n4.jf. Acesso em: 08 nov. 2020 às 01:21.

FURROW, D. *Ética*: conceitos-chave em filosofia. Porto Alegre: Artmed, 2007.

GAMEZ, P. *et al.* Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY*, Springer, 2020.

GERDES, A.; ØHRSTRØM, P. Issues in robot ethics seen through the lens of a moral Turing Test. *Journal of Information, Communication and Ethics in Society*, v. 13, n. 2, p. 98-109, 2015. Disponível em: https://portal.findresearcher.sdu.dk/da/publications/issues-in-robot-ethics-seen-through-the-lens-of-a-moral-turing-te. Acesso em: 04 jul. 2020 às 00:59.

GILLON, Raanan. Autonomy and the principle of respect for autonomy. *British Medical Journal. Clinical Research Edition*, v. 290, n. 6.484, p. 1.806-1.808, jun. 1985.

GOMES, N. Um panorama da lógica deôntica. *Kriterion*, 2008, p. 9-38.

GOVINDARAJULU, N. S. *et al.* Toward the engineering of virtuous machines. *In:* PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY, 2009, p. 29-35.

GOWANS, C. W. *Innocence lost*: An examination of inescapable moral wrongdoing. Oxford: Oxford University Press, 1994.

GRAY, A. C. M. K.; WEGNER, D. Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition*, v. 125, n. 1, p. 125-130, 2012.

GRAY, A. C. M. K.; YOUNG, L.; WAYTZ, A. Mind perception is the essence of morality. *Psychol Inq*, v. 23, n. 2, p. 101-124, 2012. Disponível em: https://doi.org/10.1080/1047840x.2012.651387.

GRIFFITHS, P.; NOWSHADE, M. K. *Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics*. Oxford: ACPI, 2019.

HÅKANSSON, S. *The Chinese Room and Turing's Wager*: moral status in the age of artificial intelligence. 2016. Disponível em: https://www.researchgate.net/profile/Simon_Hakansson/publication/309634694_The_Chinese_Room_and_Turing%27s_Wager_Moral_Status_in_the_Age_of_Artificial_Intelligence/links/581aecf308ae30a2c01d53b5/The-Chinese-Room-and-Turings-Wager-Moral-Status-in-the-Age-of-Artificial-Intelligence.pdf?origin=publication_detail. Acesso em: 05 jul. 2020 às 15:16.

HAMMING, R. W. The Theory of Automata. Reviewed work: computation: finite and infinite machines by Marvin L. Minsky. *Science*, New Series, v. 159, n. 3818, p. 966-967, 1968.

HARARI, Y. N. *Homo Deus*: a brief history of tomorrow. Random House, 2016.

HARARI, Y. N. *Sapiens*: uma breve História da humanidade. São Paulo: L&PM Editores, 2015.

HARE, R. M. Moral Conflicts (Ed.). *Moral thinking*: its levels, method and point. Oxford: Oxford University Press, 1981. p. 25-35.

HARNAD, S. Minds, machines, and Searle. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 1, n. 1, 1989.

HARNAD, S. *Minds, Machines, and Searle*: what's right and wrong about the chinese room argument. Preston and Bishop, 2002.

HARRIS, P. *Wet mind, a new cognitive neuroscience and its implications for behavioral optometry*. 2020. Disponível em: https://www.oepf.org/sites/default/files/referencearticles/WET_MIND_A_NEW_COGNITIVE_N.pdf. Acesso em: 05 jul. 2020 às 00:07.

HASKER, W. *The emergent self*. Ithaca, NY: Cornell University Press, 1999.

HEATH, G. Origins of the binary code. *Scientific American*, v. 227, n. 2, p. 76-83, August 1972.

HEGEL, G. W. F. *Princípios da Filosofia do Direito*. São Paulo: Martins Fontes, 1997.

HEIDEGGER, M. *O que é isto* — A filosofia? São Paulo: Abril Cultural, 1973. p. 219. (Col. Os Pensadores).

HERRERO, F. Javier. A ética filosófica de Henrique Cláudio de Lima Vaz. *Síntese*, Belo Horizonte, v. 39, n. 125, p. 393-432, 2012.

HESÍODO. *Teogonia*: a origem dos deuses. Trad. Jaa Torrano. São Paulo: Iluminuras, 2003.

HILGENDORF, E. *Robotik im Kontext von Recht und Moral*. Bd. 3 Robotik und Recht: Nomos Verlag, Baden-Baden, 2014.

HILL, T, E. Moral dilemmas, gaps and residues: a Kantian perspective. *In:* MASON, H. E. (org.). *Moral dilemmas and moral theory*. Oxford: Oxford University Press, 1996. p. 167-198.

HILL, T, E. *The Blackwell guide to Kant's ethics*. Chichester: Wiley-Blackwell, 2009.

HITLIN, S. *Moral selves, evil selves*: the social psychology of conscience. New York: Palgrave Macmillan, 2008.

HOLBO, J. Moral dilemmas and the logic of obligation. *American Philosophical Quarterly*, n. 3, p. 259-274, 2002.

HOLLINGS, C.; MARTIN, U.; RICE, A. The early mathematical education of Ada Lovelace. *BSHM Bulletin, Journal of the British Society for the History of Mathematics*, 2017. Disponível em: https://www.claymath.org/sites/default/files/the_early_mathematical_education_of_ada_lovelace.pdf. Acesso em: 02 ago. 2020 às 21:36.

HOMERO. *Ilíada*. Trad. Manoel Odorico Mendes. eBooksBrasil, 2009. Livro XVIII. Disponível em: http://www.ebooksbrasil.org/adobeebook/iliadap.pdf. Acesso em 23.12.2020 às 09:47.

HOWARD, D.; MUNTEAN, I. A minimalist model of the artificial autonomous moral agent (AAMA). *In:* AAAI SPRING SYMPOSIUM SERIES, Palo Alto, California, March 2016. Disponível em: https://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12760/11954. Acesso em: 02 jul. 2020 às 05:11.

HOWARD, D.; MUNTEAN, I. Artificial moral cognition: moral functionalism and autonomous moral agency. *Philos Comput*, pp. 121-159, Springer, 2017.

ILLIAMS, B. A. O.; ATKINSON W. F. Ethical Consistency. Proceedings of the Aristotelian Society. *Supplementary Volumes*, v. 39, p. 103-138, 1965.

JADERBERG, M. *et al*. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science31*, p. 859-886, May 2019.

JAEGER, W. *Paideia*: The ideals of greek culture. Trans. Gilbert Highet. Oxford University Press, 1945. v. I-III.

JOHNSON, A. M.; AXINN, S. The morality of autonomous robots. *Journal of Military Ethics*, v. 12, n. 2, p. 129-141, 2013.

JOHNSON, D.G. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, v. 8, n. 4, p. 195-204, 2006.

JOHNSTON, V. S. *Why we feel*: The science of human emotions. Reading, MA: Perseus Books, 1999.

KANG, M. The mechanical daughter of Rene Descartes: the origin and history of an intellectual fable. *Modern Intellectual History*, v. 14, n. 3, p. 633-660.

KANT, I. *A metafísica dos costumes*. Petrópolis: Vozes, 2013.

KANT, I. *A metafísica dos costumes*. Trad. de Edson Bini. Bauru: Edipro, 2003.

KANT, I. *Die Metaphysik der Sitten*. Frankfurt am Main: Suhrkamp, 1982.

KEOWN, D. *Buddhist ethics*: a very short introduction. New York: Oxford University Press, 2005.

KIM, E. E.; TOOLE, B. A. Ada and the First Computer. *Scientific American*, v. 280, n. 5, p. 76-81, May 1999.

KIRBY, S. Natural language from artificial life. *Artificial Life*, 2002.

KIRBY, S. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. IEEE Trans. *Evolutionary Computation*, v. 5, p. 102-110, 2001.

KNOBE, J. Intentional action and side effects in ordinary language. *Analysis*, v. 63, v. 3, p. 190-194, 2003b.

KNOBE, J. Intentional action in folk psychology: an experimental investigation. *Philos Psychol*, v. 16, n. 2, p. 309-324, 2003a.

KOBIS, N.; MOSSINK, L. Creative artificial intelligence – Algorithms vs. humans in an incentivized writing competition. *ResearchGate.* 2020. Disponível em: https://www.researchgate.net/publication/338689473_Creative_Artificial_IntelligenceAlgorithms_vs_humans_in_an_incentivized_writing_competition. Acesso em: 25 jul. 2020 às 10:11.

KOTTUR, S. *et al.* Natural language does not emerge 'naturally' in multi-agent dialog. *In*: EMNLP, 2017. Disponível em: https://arxiv.org/pdf/1706.08502.pdf. Acesso em: 26 jul. 2020 às 00:37.

KRÜGER, L. Ethics according to nature in the age of evolutionary thinking. *Grazer Philosophische Studien*, v. 30, p. 25-42, 1987.

KUHN, S. T. Reflections on Ethics and Game Theory. *Synthese*, v. 141, n. 1, p. 1-44, 2004.

KURZWEIL, R. *The age of spiritual machines*. New York: Viking, 1999.

KURZWEIL, R. *The singularity is near*. New York: Viking, 2005.

LAKATOS, I. History of science and its rational reconstructions. *In*: HACKING, I. (org.). *Scientific revolutions*. Hong-Kong: Oxford University, 1983.

LAKATOS, I. O falseamento e a metodologia dos programas de pesquisa científica. *In*: LAKATOS, I.; MUSGRAVE, A. (org.) *A crítica e o desenvolvimento do conhecimento*. São Paulo: Cultrix, 1979.

LARGE, W. *Levinas "Totality and Infinity"*: a reader's guide. New York: Bloomsbury Publishing.

LAVELLE, J. S. What is it to have a mind? *In*: CHRISMAN, M.; PRITCHARD, D. *Philosophy for Everyone*. London/New York: Routledge, 2014.

LAZARIDOU, A.; HERMANN, K. M.; TUYLS, K.; CLARK, S. Emergence of linguistic communication from referential games with symbolic and pixel input. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), Vancouver, 2018.

LAZARIDOU, A.; PEYSAKHOVICH, A.; BARONI, M. Multi-agent cooperation and the emergence of (natural) language. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), Vancouver, 2017.

LEBEN, D. *Ethics for robots*: how to design a moral algorithm. Oxon/NewYork: Routledge, 2018.

LEE, J. D. *et al.* Emergent translation in multiagent communication. *CoRR*, abs/1710.06922, 2018.

LEIBNIZ, G. W. *Disputatio de casibus perplexis in jure, in Samtliche Schriften und Briefe*. Berlin, Darmstadt: Otto Reichl Verlag, 1930. p. 231-256. vi.1.

LEIVAS, C. R. C. *Representação e Vontade em Hobbes*. 2005. Tese (Doutorado em Filosofia) – Instituto de Filosofia e Ciências Humanas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

LEMMON, E. J. Moral dilemmas. *The Philosophical Review*, n. 2, p. 139-158, 1962.

LEYHAUSEN. The biological basis of ethics and morality. *Science, Medicine & Man*, v. 1, p. 215-235, 1974.

MACDONALD, C.; MACDONALD, G. *Emergence in mind.* New York: Oxford University Press, 2010.

MACINTYRE, A. *Dependent Rational Animals*. Illinois: Carus Publishing, 1999. p. 9.

MACINTYRE, A. *Depois da virtude*: um estudo em teoria moral. Trad. Jussara Simões. Bauru: Edusc, 2001. p. 115.

MANDEVILLE, B. M. The fable of the bees or private vices. *Public Benefits*, v. 1 [1732]. The Online Library of Liberty. Disponível em: http://oll-resources.s3.amazonaws.com/titles/846/Mandeville_0014-01_EBk_v6.0.pdf. Acesso em: 20 jun. 2020 às 10:05.

MARCUS, R. B. Moral dilemmas and consistency. *The Journal of Philosophy*, n. 3, p. 121-136, 1980.

MARIZ, D. A especificidade da natureza humana em relação aos demais animais no pensamento aristotélico. *Argumentos*, Fortaleza, ano 6, n. 12, p. 157-168, jul./dez. 2014.

MARKS, R.; DEMBSKI, W.; EWERT, W. *Introduction to evolutionary informatics*. Hackensack: World Scientific, 2017.

MATIAS, D. W. de S. *et al.* Mentira: aspectos sociais e neurobiológicos. *Psicologia: Teoria e Pesquisa*, v. 31, n. 3, p. 397-401, jul.-set. 2015.

MAYOR, A. *Gods and robots*: myths, machines, and ancient dreams of technology. Princeton: Princeton University Press, 2018. p. 19.

McCONNELL, T. C. Moral dilemmas and consistency in ethics. *Canadian Journal of Philosophy*, n. 2, p. 269-287, 1978.

MCCULLOCH, W.; STURGIS; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, v. 52, p. 99-115, 1990.

MCLAUGHLIN, B. P. Emergence and supervenience. *Intellectica*, v. 2, p. 25-43, 1997.

MILL, J. S. *O utilitarismo*. São Paulo: Iluminuras, 2000.

MINSKY, M. L. *Computation*: finite and infinite machines. Austin: Englewood Cliffs, 1967.

MORAVEC, H. *Robots*: mere machine to transcendent mind. New York: Oxford University Press, 1999.

MORDATCH, I.; ABBEEL, P. Emergence of grounded compositional language in multi-agent populations. *In*: THE THIRTY-SECOND AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-18), 32., 2018, New Orleans, p. 1495-1502, p. 1502. Disponível em: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17007/15846. Acesso em: 02 jul. 2020 às 14:54.

MÜLLER, V. C. Ethics of Artificial Intelligence. *In*: ELLIOTT, A. (ed.). *The Routledge social science handbook of AI*. London: Routledge, 2014.

NECHES, R. *et al.* Enabling technology for knowledge sharing. *AI Magazine*, v. 12, n. 3, p. 36-56, 1991.

NEGROTTI, Massimo. Naturoids — on the nature of the artificial. London/Singapore: World Scientific Publishing, 2002, p. 48.

NUNES, L. de L.; TRINDADE, G. G. da. Conflitos morais insolúveis e sistemas racionalistas: uma abordagem sobre consistência moral. *Princípios*, Natal, v. 18, n. 30, p. 85-100, jul./dez. 2011.

O'CONNOR, T. Emergent individuals. *The Philosophical Quarterly*, v. 53, n. 213, p. 540-555, 2003.

O'CONNOR, T. Emergent properties. *American Philosophical Quarterly*, v. 31, 1994, p. 91-104.

O'CONNOR, T.; HONG YU WONG. The metaphysics of emergence. *Noûs*, v. 39, n. 4, p. 58-678, 2005.

O'SHEA, T. *The essex autonomy project*. Critics of Autonomy. University of Essex. 2012. Disponível em: https://autonomy.essex.ac.uk/wp-content/uploads/2016/11/CriticsofAutonomyGPRJune2012.pdf. Acesso em: 08 nov. 2020 às 01:21.

OVÍDIO. *Metamorfoses*. Trad. Bocage e comentários de Rafael Falcón. Porto Alegre: Concreta, 2016. p. 49. (Coleção Clássica.). Disponível em: https://issuu.com/editoraconcreta/docs/metamorfoses_-_teste. Acesso em: 23 maio 2020 às 23:34.

OWEN, A. M. The Search for Consciousness. *NeuroView. Neuron.*, v. 102, Issue 3, p. 526-528, 8 May 2019.

PANISSON, A. R. *et al.* An approach for argumentation-based reasoning using defeasible logic in multi-agent programming languages. *In*: 11th INTERNATIONAL WORKSHOP ON ARGUMENTATION IN MULTIAGENT SYSTEMS, Paris, 2014.

PANISSON, A. R. *Et al.* Towards practical argumentation-based dialogues in multi-agent systems. *In*: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON INTELLIGENT AGENT TECHNOLOGY, Singapura, 2015.

PAPINEAU, D. Why supervenience? *Analysis*, v. 50, n. 2, p. 66-71, 1990.

PARK, K.; DASTON, L. J. Unnatural conceptions: the study of monsters in sixteenth-and seventeenth-century France and England. *Past & Present*, n. 92, p. 20-54, 1981.

PATON, H. J. *The categorical imperative*: a study in Kant's Moral Philosophy. Philadelphia: University of Pennsylvania Press, 1971.

PEPPER, S. C. Emergence. *Journal of Philosophy*, v. 23, p. 241- 245, 1926.

PEREZ, A.G.; RODRIGUEZ, F.O.; TERRAZAS, B.V. Legal ontologies for the Spanish e-government. *In*: CAEPIA. *Researchgate.net/*, 2006. p. 301-310. Disponível em: https://www.researchgate.net/profile/Asuncion_Gomez-Perez/publication/221275037_Legal_Ontologies_for_the_Spanish_e-Government/links/0fcfd50b23ad68a223000000/Legal-Ontologies-for-the-Spanish-e-Government.pdf. Acesso em: 17 dez. 2020 às 00:58.

PICARD, R. *Affective computing*. Cambridge, MA: MIT Press, 1997.

PICARD, R.W.; VYZAS, E.; HEALEY, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, n. 10, p. 1175-1191, 2001.

PICOLI, R. A. Utilitarismos, Bentham e a história da tradição. *Existência e Arte*, v. 2, p. 1-20, 2010.

PIEK, M. *The Chinese Room and the Robot Reply*. Tese. (Doutorado) – Tilburg University. Philosophy of Science and Society, Holanda, [*s.d.*]. Disponível em: http://arno.uvt.nl/show.cgi?fid=146015. Acesso em: 02 jun. 2020 às 06:44.

PLATÃO. *A República*. São Paulo: Edipro, 1994.

PLATÃO. *Protágoras*. Trad. Carlos Alberto Nunes. Belém: Editora da Universidade Federal do Pará, 2002. p. 155. Disponível em: https://edisciplinas.usp.br/pluginfile.php/270800/mod_resource/content/1/platao%20protagoras.pdf. Acesso em: 20 jun. 2020 às 14:44.

PORTO, L. S. Uma investigação filosófica sobre a inteligência artificial. *Informática na Educação: Teoria & Prática*, Porto Alegre, v. 9, n. 1, jan./jun. 2006.

RAWLS, J. *Theory of Justice*. Cambridge: Harvard University Press, 1999.

RICHARDS, R. A defense of evolutionary ethics. *Biology & Philosophy*, v. 1, p. 165-293, 1986.

RODRIGUES, L. R. O paradoxo do prefácio generalizado. *Intuitio*, Porto Alegre, v. 11, n. 1, jun. 2018, p. 7-18.

RODRIGUES, R. A. Severino Boécio e a invenção filosófica da dignidade Humana. *Seara Filosófica*, n. 5, p. 3-20, 2012.

ROSENBERG, Y. *The golem and the wondrous deeds of the Maharal of Prague*. Trad. Curt Leviant. New Haven: Yale University Press, 2007.

ROSS, W. D. *The Right and the Good*. Oxford: Oxford University Press, 2002.

SAINT-VICTOR, R. de. *La Trinité*. Edição bilingue Latim-Francés. Introdução, tradução e notas de Gaston Salet SJ. Paris: Les Editions du CERF, 1959.

SARTRE, J. P. *O existencialismo é um humanismo*. Paris: Les Éditions Nagel, 1970.

SCHUURMAN, D. C. *Artificial Intelligence*: discerning a Christian response. Ontario/Canada: Canadian Scientific & Christian Affiliation, 2020. Disponível em: https://www.csca.ca/uploads/18Jan20SchuurmanDiscerningAI.pdf. Acesso em: 18 jul. 2020 às 21:12.

SEARLE, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, v.3, n. 3, p. 417-457, 1980.

SEARLE, J. R. *The mystery of consciousness*: John R. Searle and exchanges with Daniel C. Dennett and David Chalmers. New York: The New York Review of Books, 1997.

SEARLE, J. R. *The rediscovery of the mind*. Cambridge: MIT, 1995.

SHELLEY, M. *Frankenstein ou o Prometeu moderno*. Rio de Janeiro: Nova Fronteira, 2011.

SHELLEY, M. *Frankenstein, or the Modern Prometheus*. Engage Books, 2008.

SIDGWICK, H. *The Methods of Ethics*. 2011, p. 3. Disponível em: https://www.earlymoderntexts.com/assets/pdfs/sidgwick1874.pdf, acesso 20 jun. 2020 às 16:09.

SIMON, Herbert. The Sciences of the Artificial. London/Cambridge: MIT Press, 1996, p. 05. Disponível em https://epdf.pub/the-sciences-of-the-artificial-3rd-edition-pdf-5eccdc2f3d0e8.html. Acesso dia 26.12.20.

SIMON, H. A. A Theory of Emotional Behavior. Carnegie Mellon University Complex Information Processing (CIP) Working Paper #55, June 1, 1963. Disponível em http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=346072. Acesso dia 27.12.2020.

SINGER, P. *Vida ética*: os melhores ensaios do mais polêmico filósofo da atualidade. Rio de Janeiro: Ediouro, 2002.

SKALFIST, P. *A revolução da robótica*. Cambridge: Cambridge.

SLOMAN, A. Why robots will have emotions. Proceedings IJCAI. *Cognitive Science Research Paper*, Sussex University Vancouver, v. 176, p. 1, 1981.

SMITH, J. M. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

SOFFNER, R. *Algoritmos e programação em linguagem C*. São Paulo: Saraiva, 2013.

STAHL, B. C. Information, ethics, and computers: the problem of autonomous moral agents. *Minds and Machines*, v. 14, p. 67-83, 2004.

STATMAN, D. The debate over the so-called reality of moral dilemmas. *Philosophical Papers*, v. XIX, n. 3, p. 191-211, 1990.

STREFLING, S. R. A realidade da pessoa humana em Tomás de Aquino. Porto Alegre: Edipucrs, 2016. Disponível em: https://editora.pucrs.br/anais/seminario-internacional-de-antropologia-teologica/assets/2016/20.pdf. Acesso em: 08 jun. 2020 às 01:21.

STYRON, W. *A escolha de Sofia*. Rio de Janeiro: Record, 1979.

SWINBURNE, R. *Are we bodies or souls?* New York: Oxford University Press, 2019.

SWINBURNE, R. *Simplicity as evidence of truth (Aquinas lecture).* Milwaukee: Marquette University Press, 1997.

TCHAPEK, K. *A fábrica de robôs*. São Paulo: Hedra, 2012.

TEIXEIRA, J. de F. *O cérebro e o robô*: inteligência artificial, biotecnologia e a nova ética. São Paulo: Paulus, 2015.

TEIXEIRA, J. de F. *O pesadelo de Descartes*: do mundo mecânico à inteligência artificial. Porto Alegre: Editora Fi, 2018.

THOMPSON, E. *Mind in life*: biology, phenomenology, and the sciences of mind. Cambridge and London: Harvard University Press, 2007.

TIMMERMANN, Jens. Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory. *AGPh*, v. 95, n. 1, p. 36-64, 2013.

TOOLE, B. Poetical Science. *The Byron Journal*, 15, p. 55-65, 1987.

TURING, A. M. Computing Machinery and Intelligence. *Mind*, v. 49, p. 433-460, 1950.

VALLE, J. I. del. Inteligencia artificial ética: Un enfoque metaético a la moralidad de sistemas autónomos (TFG). Bruxelas: Universidad Nacional de Educación a Distancia, 2019. Disponível em: https://www.researchgate.net/publication/337797495_Inteligencia_Artificial_Etica_-_Un_Enfoque_Metaetico_a_la_Moralidad_de_Sistemas_Autonomos_TFG. Acesso em: 14.06.2020 às 20:05.

VALLOR, S. *Technology and the virtues*: a philosophical guide to a future worth wanting. New York: Oxford University Press, 2016.

VAZ, H. C. de Lima. *Escritos de filosofia*: Introdução à ética filosófica. Porto Alegre: Edições Loyola, 1986.

VELASQUEZ, J. A computational framework for emotion-based control. *In*: PROCEEDINGS OF THE WORKSHOP ON GROUNDING EMOTIONS IN ADAPTIVE SYSTEMS, INTERNATIONAL CONFERENCE ON SAB, University of Zurich, Switzerland August 21, 1998.

VERBEEK, B.; MORRIS, C. *Game Theory and Ethics*. 2009. Disponível em: https://stanford.library.sydney.edu.au/archives/spr2009/entries/game-ethics/#7. Acesso em: 26 jul. 2020 às 21:38.

VIEIRA, R *et al.* On the Formal semantics of speech-act based communication in an agent-oriented programming language. *J. Artif. Intell. Res.* (JAIR), v. 29, p. 221-267, 2007.

VINTIADIS, E. *Emergence*. Internet Encyclopedia of Philosophy. Disponível em: https://www.iep.utm.edu/emergenc/. Acesso em: 24 ago. 2020 às 01:52.

VOLKER, C. B. *As palavras do Oráculo de Delfos*: um estudo sobre o *De Phytiae Oraculis* de Plutarco. Disponível em: https://repositorio.ufmg.br/bitstream/1843/ECAP-6ZFG54/1/microsoft_word___camila_bylaardt_volker.pdf. Acesso em: 15 jun. 2020 às 23:07.

WALLACH, W.; ALLEN, C.; FRANKLIN, S. Consciousness and ethics: artificially conscious moral agents. *International Journal of Machine Consciousness*, v. 03, n. 01, p. 177-192, 2011.

WALLACH, W.; FRANKLIN, S.; ALLEN, C. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, v. 2, p. 454-485, 2010.

WALSH, T. *Machines that think*: the future of artificial intelligence. Prometheus Books, 2018.

WEBER, T. Autonomia e dignidade da pessoa humana em Kant. *Revista Brasileira de Direitos Fundamentais & Justiça*, v. 3, n. 9, p. 232-259, 2009, p. 233.

WEBER, T. *Ética e Filosofia política*: Hegel e o formalismo kantiano. Porto Alegre: Edipucrs, 2009.

WEBER, T. *Hegel*: Liberdade, Estado e História. Porto Alegre: Vozes, 1993

WEBER, T. Direito e justiça em Kant. *Revista de Estudos Constitucionais, Hermenêutica e Teoria do Direito* (RECHTD), v. 5, n. 1, p. 38-47, p. 38, jan.-jun. 2013.

WIENER, N. *Cybernetics or control and communication in the animal and the machine*. Massachusetts: MIT Press, 1965.

WIENER, N. *God and Golem*, Inc. A comment on certain points where cybernetics impinges on religion. Cambridge: MIT Press, 1963.

WIENER, N. *Some moral and technical consequences of automation science*, v. 131, Issue 3410, p. 1355-1358.

WILLIAMS, B. Ethical Consistency. *Proceedings of the Aristotelian Society*, v. 39, p. 103-124, 1965.

WRIGHT, G. Deontic Logic. *Mind*, n. 237, 1951, p. 1-15.

YOUNG, S. I am not a brain: philosophy of mind for the 21st Century. *Library Journal*, v. 142, n. 19, p. 84-84, 2017.

ZAGZEBSKI, L. Intellectual Autonomy. *Philosophical Issues*, v. 23, p. 244-261, 2013. Disponível em: http://www.investigacoesfilosoficas.com/wp-content/uploads/04-Zagzebski-2013-Intellectual-Autonomy.pdf.

ZHU, Q.; WILLIAMS, T.; WEN, R. *Confucian Robot Ethics*, 2019. Disponível em: https://www.researchgate.net/publication/339815118_Confucian_Robot_Ethics. Acesso em: 09 ago. 2020 às 04:20.

ZIEMKE, T. *The body of knowledge*: On the role of the living body in grounding embodied cognition, 2016. Disponível em: https://www.researchgate.net/publication/306350827_The_body_of_knowledge_On_the_role_of_the_living_body_in_grounding_embodied_cognition. Acesso em: 31.07.2020, às 00:07.

ZILLES, U. *Teoria do conhecimento e teoria da ciência*. São. Paulo: Paulus, 2005.