OPEN ACCESS

ESCOLA DE HUMANIDADES
PUCRS

VERITAS (PORTO ALEGRE)
Revista de Filosofia da PUCRS

Veritas, Porto Alegre, v. 70, n. 1, p. 1-16, jan.-dez. 2025
e-ISSN: 1984-6746 | **ISSN-L:** 0042-3955

http://dx.doi.org/10.15448/1984-6746.2025.1.48601

SEÇÃO: ÉTICA E FILOSOFIA POLÍTICA

Revisitando o equilíbrio reflexivo amplo numa teoria crítica da IA: ecofeminismo, decolonialidade, sustentabilidade

Revisiting Wide Reflective Equilibrium in a Critical Theory of AI: Ecofeminism, Decoloniality, Sustainability

Revisando el equilibrio reflexivo amplio en una teoría crítica de la IA: ecofeminismo, decolonialidad, sostenibilidad

Nythamar de Oliveira[1]
orcid.org/0000-0001-9241-1031
nythamar.oliveira@pucrs.br

Recebido em: 28 jul. 2025.
Aprovado em: 28 set. 2025.
Publicado em: 10 dez. 2025.

Resumo: O artigo propõe uma reformulação do equilíbrio reflexivo amplo, originariamente concebido por John Rawls e posteriormente evocado por eticistas e teóricos da Bioética, Neuroética e, mais recentemente, da Ética da Inteligência Artificial, visando à promoção de um igualitarismo interseccional (socioeconômico, de gênero, racial-étnico, ambiental) por meio da inclusão digital. Argumenta-se que somente a partir de uma perspectiva decolonial da teoria crítica combinada com uma visão ecofeminista de sustentabilidade pode-se recorrer a esse dispositivo procedimental para saldar o chamado déficit fenomenológico de teorias normativas e naturalistas, evitando o normativismo de modelos principialistas e o reducionismo de modelos eliminacionistas e fisicistas.

Palavras-chave: decolonialidade; ecofeminismo; equilíbrio reflexivo; inteligência artificial; sustentabilidade.

Abstract: The article proposes a reformulation of the wide reflective equilibrium, originally conceived by John Rawls and later evoked by ethicists and theorists of Bioethics, Neuroethics and, more recently, the Ethics of Artificial Intelligence, aiming at the promotion of an intersectional egalitarianism (socioeconomic, gender, racial-ethnic, environmental) through digital inclusion. It is argued that only from a decolonial perspective of critical theory, combined with an ecofeminist vision of sustainability, can one resort to reflective equilibrium to remedy the so-called phenomenological deficit of normative and naturalist theories, thereby avoiding the normativism of principlist models and the reductionism of eliminativist and physicalist models.

Keywords: Decoloniality; Ecofeminism; Reflective Equilibrium; Artificial Intelligence; Sustainability.

Resumen: Este artículo propone una reformulación del equilibrio reflexivo amplio, concebido originalmente por John Rawls y posteriormente evocado por eticistas y teóricos de la bioética, la neuroética y, más recientemente, la ética de la inteligencia artificial, con el objetivo de promover un igualitarismo interseccional (socioeconómico, de género, racial-étnico, ambiental) a través de la inclusión digital. Se argumenta que solo desde una perspectiva decolonial de la teoría crítica, combinada con una visión ecofeminista de la sostenibilidad, se puede recurrir al equilibrio reflexivo para remediar el llamado déficit fenomenológico de las teorías normativas y naturalistas, evitando el normativismo de los modelos principistas y el reduccionismo de los modelos eliminativistas y fisicalistas.

Palabras clave: decolonialidad; ecofeminismo; equilibrio reflexivo; inteligencia artificial; sostenibilidad.

[1] Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, Rio Grande do Sul, Brasil.

1 Introdução

O celebrado "equilíbrio reflexivo amplo" (*wide reflective equilibrium*), originariamente concebido por John Rawls, tem sido evocado por eticistas e teóricos da Bioética, da Neuroética e, mais recentemente, da Ética da Inteligência Artificial (IA), para atender a diferentes reivindicações normativas da Ética Aplicada, de forma a evitar as limitações e aporias de modelos ético-normativos *top-down* e de situações particulares de generalização *bottom-up* (Frey; Wellman 2003)[2]. Seguindo a virada feminista, pragmatista e decolonial na teoria crítica (Allen, 2016), proponho recorrer a uma crítica genealógica do poder (Forst, 2013) e a uma reformulação naturalista da imaginação política inspirada em Spinoza (Saar, 2002), de forma a revisitar a crítica de Habermas a Marcuse e reaproximá-la de uma concepção socioambiental de tecnologias de poder como formas transformadoras de subjetivação em regimes neoliberais de biopolítica, que podem ser evocadas para atualizar uma reconstrução normativa decolonial da sustentabilidade. Mesmo sem recorrer a políticas identitárias, um certo *ethos* democrático radical se identifica com seu *oikos* como lugar de fala ecológico para retomar o ecofeminismo na condição de teoria crítica da justiça socioambiental de forma a não apenas lidar com a atual crise climática que acomete o nosso planeta em pleno Antropoceno, mas também articular em termos normativos IA e Sustentabilidade. Assim, pode-se partir de uma situação concreta de injustiça, como no contexto das enchentes que assolaram o Rio Grande do Sul em maio de 2024 e que perdura até hoje com milhares de pessoas deslocadas e refugiados ambientais, para refletir sobre as variáveis causadas por humanos, além de inúmeras variáveis climáticas, como efeito do Antropoceno. Como alguém que não pode reivindicar na primeira pessoa o lugar de fala das mulheres, tenho buscado entender as contribuições radicais dos feminismos para uma teoria crítica decolonial da IA, lembrando que através do feminismo os movimentos sociais se descobrem em sua interseccionalidade decolonial, ambiental, de raça e de gênero. Dada a minha socialização (incluindo formação étnico-religiosa híbrida – judia católica reformada, afro-brasileira, ameríndia ou de povos indígenas no Brasil), proponho-me a reavaliar as contribuições seminais dos feminismos plurais para a mudança de paradigma da libertação para a decolonialidade, especialmente como as ecofeministas podem nos ajudar a saldar o que podemos chamar de déficit fenomenológico da teoria crítica e do naturalismo, numa abordagem interseccional, para além do normativismo e das abordagens reducionistas da neurociência, da natureza humana e das ciências naturais. Em última análise, trata-se de elaborar o que seria uma teoria crítica decolonial da IA, visando a uma reconstrução normativa, desde uma perspectiva teórico-crítica no Brasil e na América Latina, de um *ethos* democrático que, apesar de suas fraquezas e seus déficits normativos, seja capaz de promover uma democracia cada vez mais deliberativa, participativa e igualitária, fazendo uso extensivo de novas tecnologias digitais (abrangendo sistemas de IA e governança digital) na atual conjuntura geopolítica monopolizada pelas *big techs* e teorias desenvolvidas em países do Atlântico Norte. A IA está em todas as áreas do saber e tem encontrado aplicabilidade em todos os segmentos e em todas as atividades da existência humana. Sabemos, por exemplo, que hoje o campo da Inteligência Artificial e Sustentabilidade se tornou uma área de investigação interdisciplinar emergente, dentro do contexto das Humanidades e Ciências Sociais Aplicadas, e ganhou popularidade em todo o mundo nos últimos anos. Organizações privadas, públicas e não governamentais têm publicado diretrizes propondo uma IA sustentável capaz de promover a sustentabilidade (ambiental, social e econômica), através de princípios éticos para a regulação e a governança de sistemas inteligentes autônomos,

[2] Este texto reflete os resultados parciais de três projetos de pesquisa em andamento: "Uma Teoria Crítica Decolonial da Tecnologia: Naturalismo, Normatividade, Inteligência Artificial" (PQ-CNPq Processo 310285/2023-2); Projeto RAIES (Rede de IA Ética e Segura, 2022-26), apoiado pela FAPERGS; Projeto RAIP (Responsible AI Platform, 2023-25), apoiado pelo CNPq-MCTI.

incluindo também projetos de educação ambiental e letramento digital, de forma a integrar regiões do Sul Global e promover uma maior interlocução com sociedades do Atlântico Norte. No entanto, regiões como a América Latina e o chamado Sul Global continuam excluídas desse debate, até o momento bastante monopolizado por Estados Unidos, União Europeia e países da Commonwealth (UK, Canadá, Austrália e Nova Zelândia). No Brasil, o Ministério da Ciência, Tecnologia e Inovações (MCTI) prevê investir R$ 500 milhões no Programa de Sustentabilidade e Energias Renováveis para IA, como parte do Plano Brasileiro de Inteligência Artificial (PBIA)[3]. As pesquisas no Instituto Nacional de IA e Sustentabilidade (INCT-IAS) se propõem a preencher essa lacuna e assumir o protagonismo através da proposição de diretivas para a implementação de políticas públicas digitais, na consultoria ético-jurídica-educacional junto a desenvolvedores e empresas que produzem aplicações por meio de sistemas de IA e fornecendo estruturas e programas educacionais para a sociedade brasileira, assim como em outras cidades, regiões e em outros países. Trata-se de implementar um projeto de pesquisa interdisciplinar de forma a educar para o uso crítico, ético e responsável da IA, considerando a sustentabilidade ambiental, social e econômica, para promover tanto uma IA sustentável quanto uma sustentabilidade fomentada por sistemas de IA. Os processos formativos disseminam, destarte, princípios e questões éticas essenciais, para qualificar tanto os produtores de sistemas de IA quanto os usuários desses sistemas, incluindo o apoio às *startups* e empresas consolidadas na área de tecnologias junto aos ecossistemas de inovação. Neste artigo, limito-me aos aspectos metaéticos e ético-normativos que subjazem aos desdobramentos prático-teóricos de pesquisas em ética da IA. Ora, a Inteligência Artificial (IA) e as novas tecnologias de informação, através de plataformas digitais globais, integram hoje um número exponencialmente crescente (com mais de 4 bilhões de usuários) de comunidades virtuais, aplicativos e redes sociais, utilizando também *big data*, a Internet das Coisas, IA generativa, *chatbots* e inovadores recursos interativos. Se postularmos que a inteligência humana é o que permite a execução de tarefas como a percepção visual, o reconhecimento de fala, a tomada de decisões e a tradução entre idiomas, a IA poderia ser entendida como um complexo técnico-teórico de sistemas computacionais e de novas tecnologias e algoritmos que otimizam a capacidade de aprender e executar tarefas cognitivas levando a resultados com grande precisão e celeridade em ambientes materiais e virtuais, incluindo o aprendizado de máquina (*machine learning*), o aprendizado profundo (*deep learning*) e o aprendizado por reforço (esses dois últimos são geralmente incluídos numa visão abrangente do primeiro), assim como diferentes aplicações da engenharia de LLM (*large language models*). O aprendizado profundo representa, de resto, uma das principais partes da subárea preocupada com aprendizado de máquina (*machine learning*): trata-se de treinar e ensinar o computador a realizar tarefas a partir de dados, de instruções e de textos, treinando computadores a realizar tarefas com base em grandes volumes de dados, instruções e textos, tornando-se uma das áreas centrais da IA. A fim de lidarmos com os inúmeros desafios normativos e problemas ético-morais neste cenário, enfocando a questão crucial do alinhamento de valores numa perspectiva teórico-crítica (em autores como Marcuse, Habermas e Feenberg), tenho postulado um programa de pesquisas que parte do déficit fenomenológico da teoria crítica para propor uma tentativa de solução numa abordagem procedimental do chamado equilíbrio reflexivo amplo, aproximando os modelos algorítmicos dos modelos normativos de inspiração rawlsiana, viabilizando, destarte, uma leitura naturalista da fenomenologia moral, contraposta a modelos normativistas e fisicistas.

---

[3] Conferir a notícia no Portal do MCTI: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2025/02/pbia-preve-r-500-milhoes-para-data-centers-verdes-que-aliam-tecnologia-e-sustentabilidade. O projeto de criação de um INCT-IAS na PUCRS, a partir da RAIES, foi aprovado em 20 de março de 2025 (CNPq Processo No. 408332/2024-7) para ser implementado em cinco anos.

2 O equilíbrio reflexivo amplo em Rawls

O núcleo argumentativo dessa reformulação do equilíbrio reflexivo rawlsiano se resume à promoção do igualitarismo interseccional (socioeconômico, de gênero, racial-étnico, ambiental) por meio da inclusão digital, o que só parece viável a partir de uma perspectiva capaz de acomodar as reivindicações normativas de uma teoria crítica decolonial combinada com uma visão naturalista da sustentabilidade, dentro de um programa de pesquisa que denominei "construcionismo social mitigado" em resposta ao déficit fenomenológico de teorias normativas e naturalistas (incluindo a teoria crítica, a neurofilosofia e as ciências cognitivas, amplamente concebidas). Se o que importa, afinal, é a normatividade, a fim de evitar uma divisão entre o naturalismo e a normatividade não naturalista, por um lado, e os modelos cognitivistas e não cognitivistas, por outro, uma alternativa não fundacionista consiste em recorrer a relatos fenomenológico-morais, hermenêuticos e procedimentais da normatividade como pistas úteis para dar sentido ao "problema naturalismo-normatividade", evitando interpretações reducionistas tanto do naturalismo (Churchland, 2011) quanto do normativismo (Parfit, 2011). Eu proponho retomar a proposta quase realista de Simon Blackburn (1993, 1998) de forma a responder ao dogmatismo dos modelos normativistas, religiosos e moralistas (especialmente o realismo moral) e abordar os desafios normativos do expressivismo e do relativismo moral. A minha hipótese de trabalho é a de que o equilíbrio reflexivo amplo de Rawls pode ser, portanto, reformulado na ética e governança da IA sem os compromissos ontológicos (nos termos de Quine, algo existe em uma teoria científica se e somente se esse algo puder ser formalizado pela lógica de predicados) da divisão realista-antirrealista antes da concepção de uma "epistemologia naturalizada" e sem as aporias epistemológicas antes da virada metodológica do conhecimento situado (*situated knowledge*) e das epistemologias do ponto de vista (*standpoint epistemology*), especialmente após as críticas sistemáticas dos feminismos. Mesmo antes de reivindicar um lugar normativo de fala de uma teoria queer, o ecofeminismo desde sempre se uniu a militantes LGBTQIAPN+ para retomar contribuições seminais dos feminismos plurais (recorrendo ao termo empregado por Djamila Ribeiro e independentemente reivindicado por feministas tão diversas quanto originais, tais como Angela Davis, Seyla Benhabib, Nancy Fraser, Judith Butler, Drucilla Cornell, Amy Allen, Lélia Gonzalez, Conceição Evaristo, Rahel Jaeggi, Wendy Brown *et al.*) para propor o que seria uma teoria crítica decolonial da IA. Dada a minha socialização (incluindo uma formação étnica-religiosa híbrida — judeus católicos reformados, afro-brasileiros, ameríndios ou povos originários no Brasil), ao fazer uma reconstrução normativa decolonial como crítica imanente genealógica, proponho evocar tais contribuições dos feminismos plurais para a gramática moral e as mudanças de paradigma (guinadas, reviravoltas, viradas, conversões) da libertação para a decolonialidade, destacando como a crítica feminista pode nos ajudar a saldar o déficit fenomenológico da teoria crítica e do naturalismo através de uma abordagem interseccional, para além do normativismo e das abordagens reducionistas na neurociência, neuroética e ética da IA. As contribuições feministas e do pensamento decolonial se inserem, portanto, no âmbito de um equilíbrio reflexivo amplo. Françoise d'Eaubonne (1974) já antecipava a ampliação da interseccionalidade dos anos 1960 para entender o ecofeminismo como uma plataforma integrada para combater a opressão das mulheres e a destruição ambiental: a libertação das mulheres (*women's liberation*) era então reconhecida como a chave para a transformação social e ecológica. Afinal, os dois flagelos que ameaçavam então a humanidade eram a superpopulação e a destruição de recursos. Um novo humanismo podia ser pensado e defendido, emergindo com o fim da sociedade clássica. Afinal, foram os homens, os "falocratas", que levaram a Terra a esta situação em que a ameaça de morte é a mais iminente pela hecatombe nuclear e pela ameaça de aniquilação ambiental. Pode-se, ademais, partir do chamado "problema de pano de fundo" (*background pro-*

blem) nas formulações originais de uma teoria da justiça como equidade, em que temos, antes de mais nada, os problemas de conflitos de interesse e do senso de justiça nas circunstâncias da justiça (como Rawls os formulou, por exemplo, no capítulo 10 de *A Theory of Justice*, de 1999), incluindo o conhecimento limitado das pessoas, seus vieses e raciocínios tendenciosos, as tendências egoístas, as teorias de fundo relevantes e as diferenças doutrinárias que levam a conflitos de interesse – todas essas diferenças podem ser filosóficas, religiosas, sociais ou políticas. Como exemplo de uma divergência entre teorias de fundo relevantes e de uma diferença ético-moral, acredito que a crítica rawlsiana ao intuicionismo moral e ao problema da relação entre ética e matemática permite-nos partir das observações de Rawls sobre o construtivismo em ética e filosofia política e na matemática, recorrendo aos trabalhos de Charles Parsons (*mathematical intuitionism as constructivism*) em contraposição aos modelos intuicionistas de Cambridge (Moore) e de Oxford (Parfit). É nesse sentido preciso que tenho buscado desenvolver reflexões filosóficas sobre o construtivismo rawlsiano, em que se insere o problema do arcabouço teórico-conceitual de seu equilíbrio reflexivo amplo: para além da revisão de juízos morais e da articulação entre princípios e normas (que muitos ainda entendem, erroneamente, como um procedimento jurídico-legal, num certo tipo de principialismo aplicado), tenho procurado articular essa visão estreita (*narrow*) com uma revisão de teorias normativas decorrentes de uma concepção ampla (*wide*), como propôs o próprio Rawls em sua esquiva de concepções metafísicas, epistemológicas e teórico-filosóficas que ainda permitiriam uma leitura fundacionista de sua proposta original de uma teoria da justiça como equidade (*justice as fairness*), concebida como doutrina abrangente (*comprehensive doctrine*) – de resto, foi por isso mesmo que Rawls a reformulou em seus escritos culminando com o *Political Liberalism*, após as celebradas Dewey Lectures de 1990, assumindo explicitamente uma postura pragmatista[4].

Posso dividir essa breve comunicação em três momentos, a saber:

(i) porque Rawls se situa dentro da chamada teoria crítica da sociedade e porque sua contribuição pode responder ao problema do déficit fenomenológico em teorias da normatividade e teorias políticas como um todo;

(ii) porque podemos reformular o equilíbrio reflexivo amplo em Rawls, revisitando as ideias de dispositivo e de algoritmo como formas de modelagem, em consonância com uma "gramática moral" de inspiração wittgensteiniana em Rawls, Honneth e Forst, com uma teoria normativa da pessoa (de forma alternativa a modelos transcendentais em Kant, Strawson e Parfit) e com um programa de reconstrução normativa que encontra sua inspiração teórico-conceitual em Dewey, Chomsky e Kohlberg e que se mostra como forma alternativa de deliberação pública a modelos de democracia reflexiva (Giddens), de liberdade reflexiva (Habermas) e de liberdade social (Honneth);

porque os desafios normativos de uma teoria crítica da tecnologia podem ser satisfatoriamente respondidos por uma versão decolonial de uma teoria crítica da IA, incluindo os problemas fundamentais do alinhamento de valores, da governança de IA e da articulação orgânica entre sociedade e indivíduo. O programa de pesquisa perpassa esses três momentos como partes integrantes de uma teoria da justiça como libertação, sendo esta metáfora da "libertação" entendida de forma concreta, a partir de práticas populares e movimentos sociais "pé no chão" (*down-to-earth liberation, based on grassroots practices*), em contraposição a relatos teórico-estecistas como a recente proposta de Christoph Menke (*Theorie der Befreiung*) e de autores eurocêntricos que partem de concepções hegelianas ou de teorias visando a uma certa aplicabilidade. A palavra "libertação" é retomada em continuidade com todos os movi-

[4] *Cf.* "Revisiting Political Pragmatism and Education: Rawls, Dewey, Bernstein". *Cognitio*, São Paulo, v. 25, n. 1, 2024. Disponível em: https://revistas.pucsp.br/index.php/cognitiofilosofia/article/view/64973/44931.

mentos de resistência ao autoritarismo e opressão sistêmicos desde que o termo foi utilizado em oposição a "escravidão" (por exemplo, entre os hebreus antigos em seus relatos mítico-religiosos do êxodo) ou a dominação e "ocupação colonial" (por exemplo, na ocupação nazista da França e de territórios europeus e nas guerras anticoloniais e lutas pela independência política em ex-colônias francesas, portuguesas e europeias na África, Ásia e América Latina). Ademais, o termo "libertação" foi autoproclamado e estendido nos anos 1950, 1970 e 1970 a movimentos feministas (*women's liberation*), movimentos negros (*black liberation*), movimentos LGBTQIAPN+ (*gay liberation*) e movimentos de libertação animal (*animal liberation*). No caso da libertação latino-americana, desde a Revolução Cubana de 1959 e através de todos os movimentos liberacionistas que resistiram às ditaduras militares e aos regimes autoritários que foram impostos nesse subcontinente, a crítica descolonizante à modernidade e a suas patologias sociais do racismo e sexismo estruturais permitem uma clara antecipação do que seria agora denominada uma virada decolonial.

Em um programa de pesquisa sobre uma teoria crítica da libertação, tenho procurado articular o discurso liberacionista – mais conhecido pelos escritos teológicos da libertação a partir dos anos 1960, mas que também conheceu alguns escritos seminais em "filosofia da libertação" – a partir da recepção brasileira da chamada primeira geração da Escola de Frankfurt e do pensamento crítico-identitário que vai da Semana de Arte Moderna de 1922 até a ditadura militar de 1964-1985. Como recordamos meio século depois, *Uma Teoria da Justiça* foi publicada no mesmo ano em que Gustavo Gutiérrez tornou públicas suas reflexões sobre *Uma Teologia da Libertação*, consagrando o mais importante movimento intelectual daquela época, ao mesmo tempo que cristalizava os movimentos sociais populares e os movimentos culturais como o chamado grupo cepalino (da Comissão Econômica para América Latina, Cepal) em torno de Celso Furtado e seu concorrente opositor, a "teoria da dependência", que ao suposto subdesenvolvimento e à suposta falta de modernização das estruturas nacionais opunham o diagnóstico do atraso periférico à inserção dependente de ex-colônias dentro do sistema capitalista mundial, com suas novas colônias, semicolônias culturais e socioeconômicas. A minha contribuição tem sido elaborada à luz da recepção brasileira da teoria crítica, que é anterior à emergência da chamada filosofia da libertação nos anos 1960 e com ela se integra de forma simbiótica em sua práxis emancipadora e suas abordagens de uma hermenêutica descolonizante. Segundo a corrente definição de "filosofia da libertação" postulada por um de seus mais reputados estudiosos, Eduardo Mendieta (2020), na *Stanford Encyclopedia of Philosophy*,

> Filosofia da Libertação é o nome coletivo de um movimento filosófico e método de fazer filosofia que surgiu inicialmente na Argentina no final dos anos 1960, mas que se espalhou pela América Latina no início dos anos 1970. É por essa razão que, por vezes, alguns críticos e historiógrafos da filosofia da libertação fazem referência a uma concepção estrita e a uma concepção ampla da filosofia da libertação, a fim de se referir ao contexto imediato de suas primeiras articulações e à sua posterior disseminação e desenvolvimento geral.

3 Rawls e a teoria crítica

Se quisermos recapitular por que Rawls pode ser situado dentro do amplo espectro de uma teoria crítica da sociedade, podemos recorrer a vários escritos comparativos que já deram conta dessa filiação ou *rapprochement*, ainda nos anos 1990, em expoentes da teoria crítica como Ken Baynes (1992) e Rainer Forst (2020). Como estou assumindo essa postura, limitar-me-ei apenas a esboçar em que sentido essa identificação teórico-política pode nos ajudar a responder ao problema do déficit fenomenológico em teorias da normatividade e, particularmente, em teoria crítica. Políticas públicas que visam promover o igualitarismo social favorecem a leitura de Rawls, para além dos debates opondo liberais e comunitaristas, como um igualitarista relacional, em contraposição a uma errônea interpretação da justiça como equidade como se fosse um tipo de *luck egalitarianism* (igualitarismo de sorte), tal

qual defendem libertários e meritocratas. Contra tal leitura, Baynes (2014) e Forst (2020) enfatizam a má interpretação da justiça distributiva em Rawls quando insistem que a filosofia política rawlsiana, mesmo no nível da teoria ideal, pressupõe uma descrição socialmente situada e sensível aos fatos de sua tarefa normativa, embora isso não signifique abandonar qualquer reivindicação de objetividade ou de verdade adequadamente interpretada. Afinal, como argumentava Rawls (1999, p. 398), as concepções de justiça devem ser justificadas pelas condições de nossa vida tal como a conhecemos[5]. Portanto, não podemos reduzir o problema do igualitarismo a uma funcionalidade distributiva, mas temos de levar em consideração os projetos de vida pessoais e como bens primários (que incluem as ideias de autoestima e autorrespeito) podem ser equitativamente acessíveis a todas as pessoas que subscrevem ao sistema equitativo de cooperação social.

Embora o termo "equilíbrio reflexivo" só tenha sido popularizado (ao menos, em círculos acadêmicos) em 1971, com a publicação da obra-prima de Rawls, sabemos que o processo de equilíbrio reflexivo foi proposto pela primeira vez por Nelson Goodman (1965) como um método para justificar nossa teoria dedutiva (e, subsequentemente, indutiva) como lógica. Em termos gerais, a proposta de Goodman é que venhamos a ser justificados em acreditar em certos princípios lógicos (dedutivos) na busca de um estado de equilíbrio entre nossos juízos iniciais sobre a validade de argumentos particulares (linguagem natural) e os princípios lógicos que constituem nossas teorias lógicas, assim como no equilíbrio que se obtém ao calibrarmos diferentes intuições e crenças mesmo que embasadas em teorias científicas nas quais há desacordo epistêmico entre pares – como na meteorologia, em sua época, e como acontece ainda hoje com relação ao aquecimento global e à mudança climática (mesmo descartando os negacionistas). Desde a apresentação inicial do método por Goodman, o equilíbrio reflexivo foi apropriado por outros subcampos da filosofia e proposto como aplicável a uma gama de áreas do conhecimento, incluindo a ética, as ciências naturais e a filosofia da mente e da linguagem.

Assim, de acordo com a ampliação da proposta, à medida que teorias lógicas podem ser justificadas pelo estabelecimento de algum estado adequado de equilíbrio entre nossos julgamentos iniciais sobre casos específicos e a teoria, as teorias em outras áreas do conhecimento também poderiam ser justificadas pelo estabelecimento de um estado de equilíbrio reflexivo entre a teoria e os dados relevantes. Rawls implicitamente assumiu que a deliberação entre cidadãos sobre as questões dos fundamentos constitucionais tende a girar em torno da justificação das normas orientadoras e que a razão pública é inevitável e necessariamente evocada para assegurar legitimidade e, portanto, estabilidade para a concepção de justiça. Quando Rawls introduziu as diretrizes para uma ideia democrática de deliberação pública, essas diretrizes se referem àquelas que podem ser introduzidas para justificar qualquer norma específica em equilíbrio reflexivo. O algoritmo do equilíbrio reflexivo que Rawls postula se traduz pelo processo através do qual a razão pública ou a esfera pública (assumindo a aproximação familiar entre Rawls e Habermas nessa reformulação deliberativa da ideia kantiana de publicidade, Öffentlichkeit) exige da parte daqueles que apresentam a norma ao propor a sua justificação. No entanto, também podemos pensar na deliberação pública como um processo através do qual "os públicos e contrapúblicos" – usando aqui os termos de Fraser (1992) – indagam sobre as relações de poder na sociedade. A razão pública especifica os fundamentos normativos a partir dos quais a "dimensão de profundidade" das relações de poder na sociedade é examinada e avaliada. Podemos reformular a justiça como equidade em uma preparação do cenário para tal avaliação crítica, calibrando uma teoria ideal

[5] Sobre este e outros problemas afins na mais recente recepção de Rawls, *cf.* OLIVEIRA, Nythamar de; MOURA, Julia S.; CONSANI, Cristina F. (org.). *Justiça e Libertação*: A Tribute to John Rawls. Porto Alegre: Fundação Editora Fênix, 2021. Disponivel em: https://doi.org/10.36592/9786581110482.

da justiça com uma teoria não ideal, nos termos consagrados pela reformulação rawlsiana da justiça como equidade em seu *Liberalismo Político* de 1993.

Interessantemente, foi nesse contexto de calibragem reflexiva entre teoria ideal (por exemplo, os princípios de justiça e a sociedade bem-ordenada) e teoria não ideal (em que pessoas de carne e osso escolhem seus representantes, legisladores e governantes em processos eleitorais que pressupõem algum tipo de deliberação e iteratividade sistêmicas) que a contribuição rawlsiana mais se aproximara de uma hermenêutica e de uma fenomenologia moral em reconstrução normativa. Chomsky chegou, inclusive, a sugerir com propriedade que a teoria rawlsiana é, neste sentido, muito devedora da linguística – como Rawls, de resto, o reconhece ao referir-se à gramaticalidade em sua *Teoria da Justiça* (Mikhail, 2011). Como foi recapitulado, Forst (2014, p. 175) pôde refutar, na esteira de Habermas, a filosofia da história hegeliana e seu eurocentrismo metodológico ao reconstruir normativamente ideias universalizáveis familiares sem, no entanto, confundi-las com as concepções familiares particulares de um movimento ou grupo social determinado. Forst optou por manter a intuição rawlsiana de que a força normativa dos ordenamentos justificatórios reside no argumento de raciocínio moral (*moral reasoning*), inerente ao senso de justiça e ao senso comum de intuições e crenças compartilhadas num *ethos* democrático em termos pragmáticos, favorecendo a universalizabilidade de direitos humanos que têm uma base moral intransponível, mesmo quando variamos os contextos de justificação de uma cultura a outra. Observo, *en passant*, que tal articulação entre crenças compartilhas e o senso de justiça *vis-à-vis* princípios regulativos de um *ethos* social (por exemplo, em termos constitucionais e de eticidade, como diriam os hegelianos) também pode ser evocada como exemplar de um equilíbrio reflexivo amplo (Andreazza, 2015) e restrito (Silveira, 2009).

Destarte, pode-se reexaminar em que sentido a articulação entre uma teoria ideal e uma teoria não ideal na trilogia rawlsiana logra reabilitar um modelo deontológico procedimental de inspiração kantiana de forma a responder aos desafios de um igualitarismo social num modelo cognitivista pragmatista e pluralista, capaz de abrigar diferenças identitárias nas suas aspirações, pautas e reivindicações normativas. Neste sentido, pode-se mostrar que o conceito jurídico-formal de *igualdade* em Rawls, de matriz kantiana, torna sua utopia política realista, não apenas no sentido de mostrar-se exequível, mas, ainda, de ser defensável e capaz de responder às exigências da instável condição humana de insociável sociabilidade e de conflitos persistentes. À medida que Rawls rejeita a tese da meritocracia em sua defesa do igualitarismo, procura-se aproximar tal procedimento da igualdade jurídico-formal, em termos da "universalizabilidade" da máxima de que "somos todos iguais na medida apenas em que temos todos a mesma liberdade". Ou seja, não tanto que sejamos todos livres do mesmo modo *de facto*, mas que sejamos todos *de jure* igualmente livres. Esta situação hipotética é obviamente um construto da razão prática de que, embora todo mundo conheça a existência de algum artigo na Constituição de seu país que postule tal igualdade e disso se sirva para reivindicar direitos concretos particulares (aqui e alhures), o que em Kant seria uma proposição sintética *a priori* em Rawls não passa de um dispositivo procedimental de representação. O construtivismo rawlsiano e o equilíbrio reflexivo de seu correlato coerentismo epistêmico-moral servem para explicitar a correlação que se busca estabelecer entre igualdade e liberdade na própria formulação de princípios universalizáveis de justiça, segundo tal igualitarismo, reformulando a concepção kantiana de igualdade nos termos democráticos da justiça distributiva e a influência de fatores arbitrários – loterias naturais e sociais – nos desdobramentos de uma vida exitosa.

Seguindo uma intuição de Brian Barry (1989), a ideia filosófica de justiça tem oscilado através dos séculos entre duas tradições que remontam ao argumento de Glaucon na *República* de Platão e ao Iluminismo moderno, remetendo-nos ora ao

regramento de vantagens e de interesses mútuos (reformulado por Hobbes, Hume e Gauthier), ora à noção reguladora de imparcialidade (Kant e utilitaristas). Essa tensão parece ainda persistir na própria concepção rawlsiana da "posição original" (Rawls, 1999, § 4), precisamente quando buscava explorar e esgotar os argumentos então disponíveis para resolvê-la nos termos de uma teoria da escolha racional. Com efeito, é nesse mesmo contexto conceitual que devemos entender que se postulou sua apropriação crítica do modelo procedimental de "equilíbrio reflexivo", quando Rawls o aproxima da justificação de princípios de inferência em Goodman e o afasta da neutralidade imparcial defendida por Thomas Nagel: a objetividade em questão, segundo Rawls, serve apenas para descartar as aporias opondo posicionamentos extremos de relativismos e objetivismos. E foi nesse sentido preciso que Rawls encontrou no construtivismo kantiano uma terceira via entre concepções teleológicas (éticas das virtudes e utilitarismos) e intuicionistas da moral. Como observa Baynes (1992, p. 8) no seu estudo seminal sobre Kant, Rawls e Habermas, a formulação construtivista da filosofia prática sustentada por estes pensadores visa a "um procedimento capaz de avaliar criticamente a legitimidade de normas e instituições sociais pelo crivo de uma concepção normativa de razão prática". Outrossim, ao explorar os argumentos centrais de tais versões de construtivismo, este se mostra uma defensável elucidação dos elementos normativos da crítica social, cuja justificação é em última análise reflexiva ou recorrente, no sentido não fundacionista de não se poder mais apelar para alguma instância além da ideia do que pode ser racional e consensualmente aceito por pessoas livres e iguais. Se o construtivismo rawlsiano é mais defensável e viável do que o de seus precursores e interlocutores, permanece questão em aberto; mas o seu programa pragmatista de justificação pública da justiça social em nossas democracias constitucionais mantém-se fiel ao princípio socrático (Rawls, 1999, p. 579) à medida que uma teoria moral sempre nos conduz a rever nossos princípios e juízos, concedendo que "a justificação reside na concepção [da justiça] como um todo e como esta se encaixa e organiza nossos juízos em equilíbrio reflexivo". Somente assim poderíamos passar a uma "teoria substantiva da justiça". Pela sua implícita reformulação de uma teoria procedimental da sociedade e de uma teoria normativa da pessoa moral, uma teoria da justiça como equidade deve nos parecer mais defensável e mais viável do que outras versões do contratualismo (Rawls, 1999, p. 584).

4 Reformulando o equilíbrio reflexivo

O procedimentalismo rawlsiano coincide precisamente com a sua apropriação do construtivismo kantiano, na autorregulação recorrente de uma cooperação social entre pessoas livres e iguais. Portanto, à medida que direitos, valores e normas politicamente objetivados numa Constituição são reivindicados através de práticas cotidianas intersubjetivas (pelo voto, por reformas constitucionais, por atos de desobediência civil, pelo exercício pleno da cidadania, através de reivindicações normativas e movimentos sociais e culturais, desde a reforma agrária até os movimentos negros, feministas e LGBTQIAPN+), as aparentes defasagens entre os ideais reguladores de uma situação hipotética (situação original, sociedade bem-ordenada, os dois princípios da justiça) e nossas experiências concretas de existência social são gradativamente corrigidas de forma a "consolidar" (*to entrench*) o processo democrático-constitucional. O equilíbrio reflexivo (tanto no sentido restrito dos princípios morais e juízos ponderados particulares quanto no sentido amplo da natureza humana, teorias econômicas, teorias identitárias e suas formas de vida sociais) sempre nos remete ao processo de construção de uma sociedade bem-ordenada, de forma a nos integrar com a interminável tarefa de recorrer à posição original como dispositivo procedimental de representação. Embora somente em suas reformulações tardias do construtivismo político encontremos essa explícita articulação entre normatividade e facticidade sociais, já no seu texto seminal de 1971 (*Uma Teoria da Justiça*) identificamos 23 ocorrências do equilíbrio reflexi-

vo. Pode-se destacar algumas dessas passagens tão perspicazes quanto instrutivas, servindo não apenas para desvelar a dimensão social de uma teoria crítica da justiça como equidade mas, ainda, a sua sensibilidade ao contexto de formação do eu (*self*) como agente moral e político, que não se dá *in abstracto* mas em todas as dimensões que poderiam ser reconstruídas normativamente em termos fenomenológico-hermenêuticos, por exemplo, numa psicologia social, numa ideia coconstitutiva de significado e razão pública e numa justificação pública dessa mesma ideia:

> O papel do Princípio Aristotélico na teoria do bem é que ele enuncia um fato psicológico profundo que, em conjunto com outros fatos gerais e a concepção de um plano racional, explica nossos juízos de valor ponderados. As coisas comumente consideradas bens humanos devem se tornar os fins e atividades que ocupam um lugar importante nos planos racionais. O princípio faz parte do contexto que regula esses juízos. Desde que seja verdadeiro e leve a conclusões que correspondam às nossas convicções sobre o que é bom e mau (em equilíbrio reflexivo), ele tem um lugar apropriado na teoria moral. Mesmo que essa concepção não seja verdadeira para algumas pessoas, a ideia de um plano racional de longo prazo ainda se aplica. Podemos descobrir o que é bom para elas da mesma forma que antes (Rawls, 1999, p. 370).

Segundo Kaufman (2017), os algoritmos de IA são agentes, mas não são agentes morais, porque carecem de consciência, intencionalidade, emoções e sentimento. Os algoritmos de IA podem agir de forma mais ou menos autônoma, mas não têm consciência ou emoções. Por isso, segundo Kaufman, são apenas agentes técnicos, mas não chegam a ser agentes morais, à medida que não podem ser responsabilizados eticamente. Destarte, o que Rawls insere em sua concepção de psicologia social está em consonância com a ideia de intencionalidade que, para John Searle (1983), é a capacidade de a mente se referir a algo no mundo. Todo desejo, crença ou pensamento está sempre voltado para algo, como já havia sido tematizado por Edmund Husserl: "consciência de algo". Isso distingue os seres vivos das máquinas, que não têm tal direcionamento mental. As intuições e os juízos morais em seres humanos, que agem segundo intenções, metas e planos de vida, podem ser sistematicamente revisados em equilíbrio reflexivo, como postulara Rawls (1999, p. 392), que recorre a esse dispositivo procedimental para justificar de forma coerente nossos processos decisórios e suas representações racionais:

> Não procedi, então, como se os primeiros princípios, ou as condições a eles subjacentes, ou mesmo as definições, tivessem características especiais que lhes permitissem ocupar um lugar peculiar na justificação de uma doutrina moral. Eles são elementos e dispositivos centrais da teoria, mas a justificação repousa sobre toda a concepção e como ela se encaixa e organiza nossos juízos considerados em equilíbrio reflexivo. Como observamos anteriormente, a justificação é uma questão de apoio mútuo de muitas considerações, de tudo se encaixar em uma visão coerente (§ 4). Aceitar essa ideia nos permite deixar de lado as questões de significado e definição e prosseguir com a tarefa de desenvolver uma teoria substantiva da justiça (Rawls, 1999, p. 507).

O método de equilíbrio reflexivo amplo se define, portanto, como uma tentativa de produzir coerência em um triplo ordenado de conjuntos de crenças mantidos por uma pessoa em particular, a saber: (a) um conjunto de juízos morais ponderados, (b) um conjunto de princípios morais e (c) um conjunto de teorias de fundo relevantes (Daniels 1996, p. 22). Como foi inicialmente apropriado em Bioética e Ética Aplicada, seguindo esse entendimento, o equilíbrio reflexivo amplo inclui e pressupõe um equilíbrio reflexivo restrito e adiciona ainda um terceiro nível: Daniels considera essa extensão de teorias de fundo (*background theories*) relevantes não apenas inevitável, mas ainda necessária. Isso ocorre porque as teorias de fundo podem apoiar princípios éticos independentemente de julgamentos morais existentes e justificar sua aceitação diante de alternativas (Daniels, 1996, p. 49).

5 Uma teoria crítica da IA

Aqui essas reflexões já podem ser encaminhadas em direção ao uso bastante problemático que tem sido feito do equilíbrio reflexivo amplo em modelos computacionais de linguagem, LLM (*large language models*), a partir de redes neurais artificiais e com grande aplicabilidade

em IA generativa, como tem sido popularizado pela disseminação de diferentes versões comercializadas do ChatGPT (OpenAI) e similares. As redes neurais artificiais procuram reproduzir computacionalmente alguns aspectos do sistema nervoso humano, ou seja, combinam unidades de processamento simples (os "neurônios artificiais") em camadas que se ligam de forma inspirada nas sinapses do cérebro humano. As redes neurais artificiais simulam o funcionamento do cérebro humano por meio de neurônios artificiais, isto é, neurônios digitais interligados, organizados em camadas, imitando as conexões sinápticas para processar informações (Kaufman, 2017).

Assim como ocorreu com as primeiras aplicações do modelo rawlsiano em bioética, houve sempre uma desconfiança sobre a indeterminação e a falta de confiabilidade em tal metodologia que parece promover um certo relativismo moral, à medida que intuições, juízos e crenças morais passam por revisão *ad infinitum*, sem nenhuma garantia de ancoragem em alguma crença básica (ou num determinado conjunto de crenças básicas, como ocorre com modelos religiosos, teológicos, metafísicos, intuicionistas, teleológicos, utilitaristas, deontológicos e fundacionistas em geral). Por analogia com a lógica da pesquisa científica, no sentido popperiano de falsificacionismo, creio ser possível postular teorias morais que podem ser aperfeiçoadas ou até mesmo superadas por outras mais defensáveis e críveis, à medida que podem ser refutadas por meio da experiência e das evidências disponíveis, podendo ser consideradas mitos ou incompatíveis com a ciência e nosso conhecimento atualizado da natureza como um todo. A ideia que defendo é a de levar a sério o naturalismo e o realismo científico num modelo heurístico que, dentro da perspectiva que se põe como tarefa incessante de preencher o déficit fenomenológico da teoria crítica, de teorias normativas e do próprio naturalismo científico, tenho caracterizado como um construtivismo social mitigado: nem tudo é construção social (ao contrário de pós-modernos e pós-estruturalistas que tendem ao subjetivismo e ao relativismo) mas várias instituições e concepções o são, por exemplo, o dinheiro, os fatos sociais e os chamados "fatos normativos" ou objetos da moral, do direito e da política. No caso da IA e dos modelos algorítmicos, o problema ético-normativo se configura em torno do chamado *alinhamento de valores*, sendo necessário especificar o que está em jogo ao traçar a linha divisória entre valores morais e valores não morais, uma distinção que, em última análise, conduz ao problema do naturalismo-normatividade. Afinal, o que são valores? E o que faz dos valores que sejam morais? Já em obra originalmente de 2012, Nick Bostrom (2018) vaticinava sobre "valor final" e carga de valor na IA, mesmo antes do alinhamento de valores se tornar um problema para a investigação em IA. A ética da IA emergiu recentemente como um campo caracterizado por questões normativas sobre o potencial aparentemente infinito e imprevisível de uma IA forte ou AGI (*Artificial General Intelligence*, Inteligência Artificial Geral), seguindo a afirmação de Bostrom (2018, p. 14) de que a

> [...] tese da ortogonalidade sugere que não podemos ingenuamente supor que uma superinteligência necessariamente compartilhará qualquer um dos valores finais estereotipados associados à sabedoria e ao desenvolvimento intelectual dos humanos – curiosidade científica, preocupação benevolente pelos outros, iluminação e contemplação espiritual, renúncia à ganância material, gosto pela cultura refinada ou pelos prazeres simples na vida, humildade e altruísmo, e assim por diante.

Com efeito, como observou acertadamente Paula Boddington (2023, p. 43), o termo "alinhamento de valores" é por vezes utilizado quase como sinônimo de "ética da IA". Ora, para tecer uma reflexão normativa que justifique em termos públicos a passagem de um problema ético aplicado (no caso, ética da IA) a uma teoria crítica, temos de evitar modelos de aplicabilidade (do tipo kantiano, que por extensão aplica à política e ao direito o que está justificado em filosofia moral) ou de justificação normativista (como acaba ocorrendo com modelos rawlsianos, à medida que simplesmente assumem uma teoria da equidade como uma doutrina abrangente – algo rechaçado pelo próprio Rawls em sua reformulação de um

liberalismo político).

Portanto, ao seguir uma aproximação como a de Waelen (2022) entre a ética da IA e a teoria crítica, tenho defendido uma reformulação decolonial e emancipatória do problema ético-normativo do alinhamento de valores, conforme corretamente identificado por Russell e Norvig (2022): desafio normativo para alcançar um acordo entre valores e objetivos humanos em sistemas de IA e aprendizado de máquina. No entanto, recuso a ir tão longe a ponto de equiparar a ética da IA à teoria crítica por vários motivos que podemos desvendar aqui. Pois, embora eu concorde que os princípios éticos mais comuns da IA estão primariamente preocupados com o empoderamento individual (poder disposicional) ou com a proteção daqueles que estão sujeitos a relações de poder (poder relacional), não creio que nem a Ética da IA nem a teoria crítica tenham abordado com êxito questões de poder e de emancipação para os impotentes e oprimidos do mundo, à medida que ambas precisam levar em conta uma viragem pragmatista e decolonial em suas premissas e seus objetivos programáticos. Ademais, a ética da IA, como ética aplicada, não pode ser equiparada a uma teoria crítica ou a qualquer teoria política da sociedade democrática. Nesse sentido, podemos retomar o principialismo bioético, nossos juízos e nossas intuições morais em pesquisas em ética da IA, para concluirmos provisoriamente que podemos manter formulações de modelos ético-normativos híbridos, em equilíbrio reflexivo, de forma a conjugar premissas tão distintas quanto as que encontramos nos três mais conhecidos modelos ocidentais, a saber:

(i) modelo deontológico: a humanidade deve ser preservada e tomada sempre também como um fim em si (em termos kantianos);

(i) ética das virtudes: há valores humanos que devem ser cultivados como virtudes, de uma geração a outra, constituindo nossas intuições morais e crenças compartilhadas em um *ethos* democrático, republicano, deliberativo e inclusivo;

(i) há princípios e normativas utilitaristas que podem beneficiar o progresso moral, na medida em promovem o florescimento humano e o bem-estar de todos, maximizando a felicidade e o prazer de grupos sociais e mitigando a dor e o sofrimento, em concepções capazes de promover também o altruísmo.

Todos esses princípios se mostram compatíveis com a beneficência, a não maleficência, a autonomia moral-política e a justiça social, assim como podemos reivindicar de pesquisas em IA transparência, equidade, confiabilidade, interpretabilidade e explicabilidade (Crisp, 2006, 2015; Floridi, 1999).

Procedemos agora às questões pragmáticas da democracia, da justiça social e da sustentabilidade socioambiental para revisitar os desafios normativos da IA e das novas tecnologias para a consolidação de uma esfera pública democrática cada vez mais dominada por plataformas digitais, mídias sociais e o uso desenfreado de algoritmos para a formação da opinião, com o risco de disseminar a desinformação, as *fake news* e a manipulação de grupos sociais que se comportam como rebanhos humanos. Em seu último livro, Habermas (2022) argumenta que, se não lograrmos uma regulamentação adequada dos meios de comunicação digitais, essa nova transformação estrutural corre o risco de esvaziar as instituições através das quais as democracias podem moldar os processos sociais e econômicos para resolver problemas coletivos urgentes, que vão desde a crescente desigualdade social até a crise climática. Habermas defende, destarte, um conceito amplo de razão humana, um processo de aprendizagem colaborativa que opera através de discussões nas quais os participantes devem recorrer apenas à força do melhor argumento. Diferentes tipos de discussão – sobre fatos científicos, normas morais ou juízos estéticos – empregam diferentes padrões de justificação e o que conta como uma razão válida depende do contexto, mas todo progresso, independentemente do campo, depende de

seguirmos o caminho ao longo do qual a razão nos conduz. A principal afirmação de Habermas é a de que a razão humana, adequadamente utilizada, mantém o seu potencial emancipador para a nossa espécie. Assim como o seu primeiro livro, *A transformação estrutural da esfera pública* (Habermas, 1987) traçou o surgimento da esfera pública no século XVIII como espaço social funcionalmente distinto, localizado entre a privacidade da sociedade civil e os gabinetes formais do Estado moderno, em que os cidadãos podiam participar de processos de deliberação democrática, Habermas chama agora a atenção para uma série de fenômenos contemporâneos, incluindo a organização da opinião pelos partidos políticos e o desenvolvimento dos meios de comunicação de massa financiados pela publicidade, que perturbaram a possibilidade de um debate político generalizado e bem informado. A democracia moderna, argumenta Habermas, está cada vez mais caracterizada pela organização tecnocrática de interesses, e não pela discussão aberta de princípios e valores. Com o avanço das tecnologias de informação e comunicação, especialmente da Internet, novas mudanças surgiram na esfera pública, à medida que a imprensa escrita começou a perder espaço e importância para o jornalismo digital. Essas mudanças, consideradas por vários autores parte dos sintomas de uma condição pós-moderna, teriam dado origem a novas formas de interação social e a novos espaços públicos. No seu último livro, portanto, Habermas começa por abordar a relação entre a teoria normativa e a teoria empírica, antes de explicar por que e como devemos compreender o processo democrático, uma vez institucionalizado em condições sociais marcadas pelo individualismo e pelo pluralismo, à luz da política deliberativa, concluindo essas reflexões teóricas preliminares com uma recapitulação das condições improváveis que devem ser preenchidas para que uma democracia capitalista propensa a crises permaneça estável. Dentro desse quadro teórico, para o qual *A transformação estrutural* de 1962 forneceu uma análise histórico-social preliminar, Habermas se propõe a descrever como a digitalização está transformando hoje, cada vez mais, a estrutura dos meios de comunicação social, o impacto que essa transformação tem no processo político e suas polarizações hodiernas. O avanço tecnológico marcado pela comunicação digitalizada fomenta inicialmente tendências para a dissolução de fronteiras, mas também para a fragmentação da esfera pública. O caráter de plataforma dos novos meios de comunicação termina por criar, com a esfera pública editorial, um espaço de comunicação em que leitores, ouvintes e telespectadores podem assumir espontaneamente o papel de autores. O alcance dos novos meios de comunicação é demonstrado pelos resultados de um inquérito longitudinal sobre a utilização da oferta alargada de meios de comunicação social na Alemanha e alhures. Embora a utilização da Internet tenha aumentado exponencialmente nas últimas duas décadas e tanto a televisão como o rádio tenham conseguido se manter, em grande medida, o consumo de jornais e revistas impressos despencou, a ascensão dos novos meios de comunicação social está ocorrendo hoje à sombra da exploração comercial da comunicação virtual não regulamentada pela Internet. Por um lado, isso ameaça minar a base econômica dos editores de jornais tradicionais e dos jornalistas como grupo ocupacional responsável. Por outro lado, um modo de comunicação semipública, fragmentada e fechada em si mesmo parece se espalhar entre os utilizadores exclusivos dos meios de comunicação social, o que distorce a sua percepção da esfera pública política como tal. Se essa conjectura estiver correta, um importante pré-requisito subjetivo para um modo mais ou menos deliberativo de opinião e formação de vontade está comprometido entre uma parcela crescente da cidadania.

6 Conclusão

Com efeito, a tese programática habermasiana da colonização do mundo da vida (especialmente nos dois volumes da *Teoria do agir comunicativo*) reflete vários estudos seminais e reflexões anteriores sobre a alienação, o fetichismo do mercado

e a reificação num sentido que já antecipa a sua proposta normativa de resgate de um sistema comunicativo capaz de evitar a mera instrumentalização e tecnificação do mundo social e de suas relações de produção, reduzindo-os a algo independente e totalmente indiferente à vontade e às reivindicações normativas dos atores sociais. O prognóstico habermasiano é consistente com a denúncia atual do paradoxo crucial que acomete o desenvolvimento dos sistemas de IA, ou seja, quanto menor a participação de uma parte interessada no ciclo de vida do sistema de IA, mais influência terá na forma como o sistema funcionará. Isto implica que o impacto social na justiça do sistema está nas mãos daqueles que são menos impactados por ele, refletindo outros paradoxos da modernidade já apontados pela chamada "primeira geração da Escola de Frankfurt". Nas palavras de Habermas (1997, p. 310), "Um sistema democrático como um todo fica prejudicado quando a infraestrutura da esfera pública não é mais capaz de direcionar a atenção dos cidadãos para as questões relevantes que precisam ser decididas e, ainda, garantir a formação de opiniões públicas concorrentes. – e isso significa opiniões filtradas qualitativamente".

A crítica de Habermas à filosofia da tecnologia de Marcuse (*Técnica e ciência como ideologia*, 1968) já refletia a sua visão perspicaz de uma sociedade mais democrática e justa, caracterizada pela comunicação aberta e pelo discurso racional. Habermas mostrava-nos, então, que a tecnologia poderia desempenhar um papel importante na concretização dessa visão, mas apenas se fosse utilizada de forma consistente com os valores democráticos e o respeito pela dignidade humana. Segundo Habermas, Marcuse vinculou a racionalização progressiva da sociedade (seguindo a crítica marxista do capitalismo e a interpretação weberiana da secularização) à institucionalização do desenvolvimento tecnocientífico, à medida que a tecnociência permeia as instituições sociais e as transforma radicalmente, em detrimento de antigas legitimações e códigos tradicionais de normatividade social. A filosofia social de Marcuse denuncia, portanto, a peculiar fusão da tecnologia com a dominação e da racionalidade com a opressão, numa abordagem unidimensional da racionalidade instrumental que provoca alienação, reificação e colonização. Como Feenberg (2017) e Habermas (1987) observaram corretamente, Marcuse segue Heidegger na aparente demonização da tecnologia moderna, mas, em vez de procurar refúgio ontológico numa nova linguagem histórica do Ser, Marcuse defende a libertação humana e a utopia social através de movimentos sociais (especialmente estudantes, trabalhadores, mulheres e grupos discriminados). Habermas alerta-nos agora para as ameaças do controle algorítmico da comunicação que flui dos mercados hegemônicos e desregulamentados, bem como do poder concentrado nas grandes corporações da Internet (*big techs*). Assim, fica posta a tarefa inacabada de levar a cabo uma teoria crítica decolonial da IA e das novas tecnologias, que hoje parecem aumentar o abismo geopolítico que separa as nações do Hemisfério Norte do Sul Global.

Referências

ALLEN, Amy. *The End of Progress: Decolonizing the Normative Foundations of Critical Theory*. New York: Columbia University Press, 2016.

ANDREAZZA, Tiaraju. Equilíbrio Reflexivo Amplo e a Revisibilidade das Crenças Morais. *Ethic@*, Florianópolis, v. 14, n. 3, p. 473-489, 2015.

BARRY, Brian. *Theories of Justice*: A Treatise on Social Justice. v. I. Berkeley: University of California Press, 1989.

BAYNES, Kenneth. Critical Theory and Habermas. *In*: MANDLE, Jon; REIDY, David (ed.). *A Companion to Rawls*. Malden: Blackwell, 2014. p. 487-503.

BAYNES, Kenneth. *The Normative Grounds of Social Criticism: Kant, Rawls, Habermas.* Albany: SUNY Press, 1992.

BENHABIB, Seyla. The methodological illusions of modern political theory: The case of Rawls and Habermas. *Neue Hefte für Philosophie*, [*s. l.*], p. 47-74, 1982.

BENHABIB, Seyla. *Situating the Self.* Gender, Community and Postmodernism in Contemporary Ethics. [*S. l.*]: Polity Press, 1992.

BENHABIB, Seyla; CORNELL, Drucilla (ed.). *Feminism as Critique*: On the Politics of Gender. Minneapolis: University of Minnesota Press, 1987.

BLACKBURN, Simon. *Essays in Quasi-realism*: A defence of quasi-realism as applied to ethics. Oxford University Press, 1993.

BLACKBURN, Simon. *Ruling Passions*. Oxford: Oxford University Press, 1998.

BODDINGTON, Paula. *AI Ethics*. A Textbook. Artificial Intelligence: Foundations, Theory and Algorithms. London: Springer, 2023.

BOSTROM, Nick. *Superinteligência*: Caminhos, perigos e estratégias para um novo mundo. Rio de Janeiro: Darkside Books, 2018.

BUTLER, Judith. What is Critique? An Essay on Foucault's Virtue. *In*: INGRAM, David (ed.). *The Political*: Readings in Continental Philosophy. [*S. l.*]: Basil Blackwell, 2002.

BUTLER, Judith. What is Critique? An Essay on Foucault's Virtue. In: INGRAM, David (ed.). *The Political*: Readings in Continental Philosophy. [*S. l.*]: Basil Blackwell, 2002.

CHALMERS, David. *Reality+*: Virtual Worlds and the Problems of Philosophy. [*S. l.*]: W. W. Norton, 2022.

CHURCHLAND, Patricia. *Braintrust*: What neuroscience tells us about morality. Princeton: Princeton University Press, 2011.

CRISP, Roger (ed.). *The Oxford Handbook of the History of Ethics.* Oxford: Oxford University Press, 2015.

CRISP, Roger. *Reasons and the Good.* Oxford: Clarendon Press, 2006.

DANIELS, Norman. *Justice and Justification*: Reflective Equilibrium in Theory and Practice. Cambridge: Cambridge University Press, 1996.

D'EAUBONNE, Françoise. *Le féminisme ou la mort*. Paris: Horay, 1974.

FEENBERG, Andrew. *Technosystem*: The social life of reason. Cambridge: Harvard University Press, 2017.

FLORIDI, Luciano. Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, [*s. l.*], v. 1, n. 1, p. 33-52, 1999.

FORST, Rainer. The Point of Justice. On the Paradigmatic Incompatibility between Rawlsian 'Justice as Fairness' and Luck Egalitarianism. *In*: MANDLE, Jon; ROBERTS-CADY, Sarah (ed.). *John Rawls*: Debating the Major Questions. Oxford: Oxford University Press, 2020. p. 148-160.

FORST, Rainer. *Justification and Critique*: Towards a Critical Theory of Politics. Trans. C. Cronin. Cambridge: Polity Press, 2014.

FORST, Rainer. *Kontexte der Gerechtigkeit*. Politische Philosophie jenseits von Liberalismus und Kommunitarismus. Frankfurt: Suhrkamp, 1994. (Em português: *Contextos da justiça*. Para além de liberalismo e comunitarismo. Trad. D.L. Werle. São Paulo: Boitempo, 2010).

FORST, Rainer. *Justification and Critique: Towards a Critical Theory of Politics.* New York: Wiley, 2013.

FRASER, Nancy. Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *In*: CALHOUN, Craig J. *Habermas and the Public Sphere*. Boston: MIT Press, 1992.

FREY, R. G.; WELLMAN, Christopher H. (ed.). *A Companion to Applied Ethics.* Blackwell Companions to Philosophy. Malden: Blackwell, 2003.

GOODMAN, Nelson. *Fact, Fiction, and Forecast*, second edition. Indianapolis: Bobbs-Merrill, 1965.

HABERMAS, Jürgen. *Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Frankfurt: Suhrkamp, 1962.

HABERMAS, Jürgen. Technology and Science as 'Ideology'. *In*: TOWARD a Rational Society. Oxford: Polity Press, 1987. [*Technik und Wissenschaft als 'Ideologie'*, Frankfurt: Suhrkamp, 1968].

HABERMAS, Jürgen. *Direito e Democracia*: Entre facticidade e validade. v. I e II. Tradução de Flávio Beno Sieveneichler. Rio de Janeiro: Tempo Brasileiro, 1997.

HABERMAS, Jürgen. *Ein neuer Strukturwandel der* Öffentlichkeit *und die deliberative Politik.* Frankfurt: Suhrkamp, 2022. (Em português: *Uma nova mudança estrutural da esfera pública e a política deliberativa*. Trad. Denilson Werle. São Paulo: Unesp, 2023).

HABERMAS, Jürgen. *The Theory of Communicative Action I*: Reason and the Rationalization of Society. Trans. T. McCarthy. Boston: Beacon Press, 1984.

HABERMAS, Jürgen. *The Theory of Communicative Action II*: Lifeworld and System. Trans. T. McCarthy. Boston: Beacon Press, 1989.

KAUFMAN, Dora. *Desmistificando a Inteligência Artificial*. Belo Horizonte: Autêntica, 2017.

MARCUSE, Herbert. *One-Dimensional Man*. London: Routledge & Kegan Paul, 1964.

MENDIETA, Eduardo. Philosophy of Liberation. *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Winter 2020. Disponível em: https://plato.stanford.edu/archives/win2020/entries/liberation/. Acesso em: 20 jan. 2025.

MENKE, Christoph. *Theorie der Befreiung*. Frankfurt: Suhrkamp, 2024.

MIKHAIL, John. *Elements of Moral Cognition*: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment. Cambridge: Cambridge University Press, 2011.

MÜLLER, V. C. Ethics of Artificial Intelligence and Robotics. *The Stanford Encyclopedia of Philosophy*. Edited by E. N. Zalta. 2021. Disponível em: https://plato.stanford.edu/ archives/sum2021/entries/ethics-ai/. Acesso em: 20 jan. 2025.

OLIVEIRA, Nythamar de. Revisiting Political Pragmatism and Education: Rawls, Dewey, Bernstein. *Cognitio*, São Paulo, v. 25, n. 1, 2024. Disponível em: https://revistas.pucsp.br/index.php/cognitiofilosofia/article/view/64973/44931. Acesso em: 20 jan. 2025.

OLIVEIRA, Nythamar de; MOURA, Julia S.; CONSANI, Cristina F. (org.). *Justiça e Libertação*: A Tribute to John Rawls. Porto Alegre: Fundação Editora Fênix, 2021. Disponível em: https://doi.org/10.36592/9786581110482. Acesso em: 20 jan. 2025.

PARFIT, Derek. *On What Matters*. 3 v. Oxford: Oxford University Press, 2011.

PARSONS, Charles. *Philosophy of Mathematics in the Twentieth Century*: Selected Essays. Cambridge: Harvard University Press, 2014.

RAWLS, John. *A Theory of Justice*. Revised Edition. Cambridge: Harvard University Press, 1999.

RAWLS, John. *Political Liberalism*. New York: Columbia University Press, 1993. (Paperback edition, 1996).

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence*: A Modern Approach. 4. ed. London: Pearson, 2022.

SAAR, Martin. Genealogy and Subjectivity. *European Journal of Philosophy*, [*s. l.*], v. 10, n. 2, p. 231-245, 2002.

SEARLE, John. Intentionality: An Essay in the Philosophy of Mind. Cambridge: Cambridge University Press, 1983.

SILVEIRA, Denis Coitinho. Posição Original e Equilíbrio Reflexivo em John Rawls: O Problema da Justificação. *Trans/Form/Ação*, [*s. l.*], v. 32, n. 1, p. 139-157, 2009.

TEIXEIRA, João Fernandes. *Filosofia da Mente e Inteligência Artificial*. Campinas: Unicamp, 2016.

WAELEN, R. Why AI Ethics is a Critical Theory. *Philosophy & Technology*, [*s. l.*], v. 35, n. 9, p. 1-16, 2022.

YOUNG, H. P. Social Norms. *In*: DURLAUF, Steven N.; BLUME, Lawrence E. (ed.). *New Palgrave Dictionary of Economics*. [*S. l.*]: Palgrave MacMillan, 2008.

Nythamar de Oliveira

Ph.D. in Philosophy, State University of New York, 1994. Professor Titular, PPG-Filosofia, Escola de Humanidades, PUCRS. Pesquisador do CNPq desde 1995.

Endereço para correspondência

NYTHAMAR DE OLIVEIRA

Rua Irmão José Otão nº 395, apt. 605, Bom Fim, 90035-060

Porto Alegre, Rio Grande do Sul, Brasil

Os textos deste artigo foram revisados por Araceli Pimentel Godinho e submetidos para validação dos autores antes da publicação.