IA emocional e design capcioso: a questão da soberania para a subjetividade

Emotional AI and Deceptive Design: The Issue of Sovereignty for Subjectivity

Fernanda Bruno [a, *]

Paulo Faltay [b]

Alice Lerner [a]

Helena Strecker [a]

RESUMO: Este artigo aborda a questão da soberania digital não somente como questão política e econômica, mas no âmbito da relação humano-máquina e, especialmente, do sujeito e da subjetividade. Focalizaremos dois domínios em que a questão da soberania do sujeito se coloca de forma mais explícita: os sistemas de inteligência artificial (IA) voltados para a inferência automatizada de características emocionais e psicológicas e as práticas de design capcioso, que consistem na incorporação de padrões obscuros de design herdados da economia comportamental na construção das plataformas. Nos dois casos, analisaremos as implicações para a questão da soberania do sujeito e para o debate sobre regulação da IA. Na tentativa de impor barreiras à manipulação ou à influência insidiosa sobre indivíduos e populações, os marcos regulatórios correm o risco de reivindicar, de um lado, um sujeito plenamente autônomo e livre de influências ou, de outro, um sujeito definitivamente vulnerável que precisa ser protegido. Apontamos como nenhum desses extremos se sustenta quando se trata de compreender como as subjetividades, os processos psicológicos e emocionais são interpelados em plataformas de IA. Propomos, por fim, que nos ambientes digitais, plataformas e aplicações de IA visados nesse artigo, a soberania do sujeito ou da subjetividade não pode ser assegurada de modo definitivo no plano individual nem no plano jurídico. Ela é, antes, um problema sociotécnico e tecnopolítico coletivo a ser continuamente trabalhado e revisado.

Palavras-chave: Soberania do Sujeito; Regulação de IA; Inferência de Emoções; Design Comportamental.

ABSTRACT: This article addresses the issue of digital sovereignty not only as a political and economic matter but also within the realm of the human-machine relationship and, more specifically, the subject and subjectivity. We focus on two domains where the question of subject sovereignty is most explicitly raised: AI systems aimed at the automated inference of emotional and psychological characteristics, and deceptive design practices, which involves the incorporation of dark design patterns inherited from behavioral economics in platform construction. In both cases, we analyze the implications for the question of subject sovereignty and the debate on AI regulation, guided by the central question: what kind of subject sovereignty could be proposed in AI platforms? We discuss how, in the attempt to impose barriers against manipulation or insidious influence on individuals and populations, regulatory frameworks risk asserting either a fully autonomous subject, free from influences, or a definitively vulnerable subject that needs protection. We point out that neither of these extremes holds up when it comes to understanding how subjectivities and psychological and emotional processes are engaged by AI platforms. Finally, we propose that in the digital environments, platforms, and AI applications targeted in this article, subject or

[a] *Programa de Pós-Graduação em Comunicação e Cultura, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil*
[b] *Programa de Pós-Graduação em Comunicação, Universidade Federal de Pernambuco, Recife, PE, Brasil.*

[*] Correspondência para/*Correspondence to* Fernanda Bruno . E-mail: bruno.fernanda@gmail.com

Recebido em/*Received*: 11/10/2024; Aprovado em/*Approved*: 17/12/2024.

subjectivity sovereignty cannot be definitively ensured at the individual or legal level. Instead, it is a sociotechnical and technopolitical issue that must be continuously addressed and revised collectively.

Keywords: Subject Sovereignty; AI Regulation; Emotion Inference; Behavioral Design.

INTRODUÇÃO

Num conhecido texto, publicado em 1917, Freud anuncia a terceira grande ferida que golpearia o narcisismo da humanidade. A primeira ferida narcísica teria sido de ordem cosmológica e romperia a ilusão de que a morada da humanidade ocuparia o centro imóvel do universo. Copérnico, que desferiu o derradeiro golpe, na esteira de outros que o precederam, retirou a Terra do centro do universo e num mesmo movimento a ilusão de que a humanidade habitava um lugar privilegiado na ordem cósmica. A segunda ferida narcísica seria um golpe "biológico", tributário das pesquisas de Charles Darwin e colaboradores, atestando nosso parentesco com o mundo animal, contra a pretensão de que seríamos apartados e superiores a este. A terceira, de autoria da psicanálise e do próprio Freud, incidiria sobre a centralidade da consciência no seio do psiquismo. A existência de processos inconscientes retira do Eu e da consciência o domínio sobre a vida psíquica. A consciência encontra-se, a partir de então, determinada e afetada por processos que ela desconhece. Tal ferida narcísica, que segundo Freud seria a mais dolorosa, nos diz, irremediavelmente: "o Eu não é senhor em sua própria casa" (Freud, 2010, p. 186).

Pouco mais de trinta anos se passaram para que um novo golpe narcísico começasse a se desenhar. A publicação do trabalho *Computer Machinery and Intelligence*, de Alan Turing (1950), anuncia a questão que se desdobrará, a partir de 1957, num programa de pesquisa controverso e vigoroso: "*As máquinas podem pensar?*". A ferida narcísica aqui anunciada consiste na extensão do pensamento, da cognição ou da inteligência (a depender da abordagem em jogo) a entes artificiais. Não se trata mais ou apenas, como na ferida anterior, de afirmar que a consciência não é o centro do sujeito, mas de postular uma nova forma de descentramento ao estender a seres artificiais aquilo mesmo que acreditávamos nos diferenciar deles.

Como sabemos, a pergunta lançada por Turing permanece aberta e tem sido tão contestada quanto relançada ao longo da história da Inteligência Artificial (IA). Essa história, ao mesmo tempo, acabou por desenvolver modelos computacionais que não necessariamente simulam o pensamento humano, mas ainda assim executam tarefas complexas e exibem capacidade de aprendizado maquínico com impactos expressivos em diversos setores como a visão computacional, o processamento de linguagem natural, entre outros. Tais máquinas saíram dos laboratórios e ganharam o mundo, integrando-se às infraestruturas do capital financeiro, da comunicação, do governo, da saúde, da segurança, do trabalho etc. Além disso, o modelo hegemônico de Inteligência Artificial em curso alimenta-se de uma voraz extração de dados de nossa experiência em ambientes, aplicações e plataformas digitais, atuando como mediadores de nossas interações e ações em diversos domínios.

Em boa parte das plataformas e aplicações de inteligência artificial, os mecanismos de extração de dados, o conhecimento que se produz a nosso respeito, os processos de tomada de decisão e a modulação de nossas condutas se dão muitas vezes nas margens de nossa percepção, consciência e atenção. Ao mesmo tempo, uma diminuta e custosa margem de negociação e contestação desses processos nos é concedida. Ainda que sejamos constantemente incitadas a interagir em tais aplicações e plataformas, tal agência é limitada, conduzida e modulada por operações que nos escapam. Se por um lado a inteligência artificial não realizou o sonho de simular o pensamento humano, por outro, passa a integrar cada vez mais as engrenagens do modo como percebemos e agimos no mundo, tendo, entretanto, a nossa capacidade de agência nesse processo constantemente constrangida.

A questão da soberania se coloca claramente aqui. Na literatura acadêmica e nos embates tecnopolíticos, a soberania digital (Allen, 2021; D'Almonte, Santos, 2024) é mobilizada comumente como a recuperação por parte dos Estados da capacidade de controle da infraestrutura material e tecnológica, como também da definição de políticas, regras, atores e valores nacionais nos ambientes e economia digitais. Entretanto, como aponta Barbosa, (2022), o termo é polissêmico, abarcando perspectivas que vão além da volta dos estados nacionais como autoridade normativa central, podendo ser utilizado também no âmbito da autodeterminação informacional ou digital.

Nesse artigo, nos interessa abordar a questão da soberania - digital, tecnológica ou de dados - não somente como questão política e econômica, mas sobretudo no âmbito da relação humano-máquina e, mais especificamente, do sujeito e da subjetividade. Ressoando a provocação de Hui (2020) e Silveira (2023), que apontam para a tecnodiversidade nos modos de ser e de se relacionar com a tecnologia, nos interessa colocar o problema da soberania focalizando o modo como o sujeito e a subjetividade são interpelados pelas grandes plataformas, algoritmos e sistemas de inteligência artificial. Analisaremos, especialmente, dois domínios em que a questão da soberania do sujeito se coloca de forma mais explícita: os sistemas de IA voltados para a inferência automatizada de características emocionais e psicológicas, e a incorporação de *padrões obscuros* de design herdados da economia comportamental na construção das plataformas. Apontaremos brevemente, ainda, algumas implicações dessa reflexão para o debate sobre regulação da IA.

Vale ressaltar, contudo, que o problema da soberania, pensado no âmbito da subjetividade ou do sujeito, é atravessado por tensões e limites que precisam ser considerados. Que modalidade de soberania do sujeito poderia ser proposta em plataformas de IA, considerando que o "eu não é senhor em sua própria casa"? Como as subjetividades podem ser mobilizadas em plataformas digitais de modo a evitar os fantasmas extremos da servidão e da completa autodeterminação? Certamente, não cabe nos limites deste artigo responder a essas questões. Nosso propósito é compreender como as plataformas de IA nos interpelam a considerar a soberania do sujeito e da subjetividade como problema. A intenção deste texto é, sobretudo, colaborar na identificação das questões às quais devemos estar atentos nesse debate.

Antes de explorarmos os problemas e domínios que serão analisados nesse artigo, cabe apontar as conexões entre a consolidação do modelo hegemônico da IA e o avanço das operações extrativas sobre as emoções, o psiquismo e o comportamento.

AS MÁQUINAS CONEXIONISTAS E O EXTRATIVISMO EMOCIONAL

O amplo campo da Inteligência Artificial tem sido historicamente marcado pela disputa entre dois paradigmas - o simbólico e o conexionista - que propõem modelos computacionais com visões distintas acerca da própria concepção de pensamento e de inteligência. Dominique Cardon, Jean-Philippe Cointet e Antoine Mazières (2018) mostram como, durante quase cinco décadas, o campo da Inteligência Artificial foi dominado por uma abordagem hipotético-dedutiva que apostava em um modelo lógico alinhado com o cognitivismo funcionalista. O paradigma simbólico, capitaneado por John McCarthy, Marvin Minsky, Herbert Simon e Allen Newell, pressupunha o pensamento como cálculo de símbolos que teriam tanto realidade material quanto representação semântica, de modo que o modelo de inteligência incorporado nas máquinas se aproximava ao de um expert que deveria conhecer, *a priori*, todos os resultados possíveis do cálculo. Em contraposição a esse paradigma, as técnicas de aprendizado de máquina e aprendizado profundo trabalham o pensamento como cálculo de funções elementares distribuídas em redes neurais que dispensam qualquer programa anterior (Cardon, Cointet, Mazières, 2018). O paradigma conexionista resgata princípios das primeiras tentativas de modelização matemática de redes neurais nos primórdios da cibernética, visando criar máquinas indutivas que aprendem *a partir* do mundo, sem a necessidade de hipóteses prévias incorporadas na infraestrutura de cálculo. Ganha corpo um sonho antigo de criar uma máquina inteligente e adaptativa, que se relaciona intimamente com o ambiente para dele extrair o seu conhecimento.

Esta disputa epistêmica ganhou recentemente contornos mais definidos na medida em que o modelo conexionista-indutivo se mostra mais adequado às dinâmicas e interesses da economia digital de plataformização da internet (Helmond, 2015). As redes neurais de aprendizado profundo que protagonizam o modelo conexionista se difundiram por todas as camadas do tecido social, da comunicação à saúde e à educação, passando pelo trabalho, segurança pública, planejamento urbano, lazer e entretenimento, alterando drasticamente a correlação de forças que vigorava no campo da Inteligência Artificial até então. Ao passo que a plataformização da internet canalizava a produção e coleta de dados e de informações sem precedentes para um conjunto restrito de atores, essa concentração permitiu ampliar radicalmente o volume e a variedade de dados que as máquinas conexionistas precisavam para dar seu salto "cognitivo", uma vez que os dados - e não mais o programa - são a base a partir da qual a máquina conexionista aprende e calcula. Dito de outra forma, para que o modelo conexionista possa operar a indução que lhe é característica, é necessário que a quantidade e variedade dos dados disponíveis seja abundante. Essa necessidade, por sua vez, alimenta e é alimentada por uma empreitada extrativista das grandes

empresas de tecnologia que buscam traduzir cada vez mais aspectos do mundo, dos indivíduos e das relações psicossociais às suas engrenagens de processamento maquínico (Gago, Mezzadra, 2017; Bruno, Bentes, Faltay, 2019; Zuboff, 2021; Ricaurte, 2022).

O interesse das plataformas e corporações de tecnologia por dados que permitam inferir características psicológicas e emocionais dos usuários vem ganhando visibilidade desde a repercussão de casos como o experimento sobre contágio emocional realizado pelo Facebook em 2014 e o escândalo da Cambridge Analytica em 2018, entre outros menos notórios. Mobilizado pela inquietação diante desse interesse crescente, o MediaLab.UFRJ investiga, desde 2018, a emergência de tecnologias que pretendem inferir emoções, traços psíquicos e de personalidade, vieses cognitivos e vulnerabilidades comportamentais por meio de mecanismos automatizados. A noção de *economia psíquica dos algoritmos*, desenvolvida em trabalhos anteriores (Bruno, 2018; Bruno, Bentes, Faltay, 2019), ressalta justamente o gradual investimento em processos algorítmicos de captura, análise e utilização de dados psíquicos e emocionais. A vitória do paradigma conexionista e indutivo da IA, mencionada acima, está intimamente relacionada a esse avanço da empreitada extrativa sobre o psiquismo e as emoções.

A recente consolidação do campo da IA emocional tende a ampliar e complexificar as demandas insaciáveis das máquinas inteligentes por mais inferências e incidências sobre o mundo e sobre nós mesmos. Na tentativa de reagir e regular a rápida penetração dessas tecnologias, recentes legislações e projetos regulatórios da inteligência artificial têm se esforçado para dar conta tanto das sutis influências do design das aplicações digitais quanto do processo de extração e utilização de dados que incidem sobre nossos corpos, emoções e subjetividades. Aprovada em março de 2024, a Lei da IA da União Europeia (*EU AI Act*) é uma das primeiras regulações abrangentes sobre inteligência artificial no mundo, e por isso vem servindo de inspiração para diversos países que procuram legislar sobre estas novas tecnologias, dentre eles o Brasil. O Artigo 5 da legislação tem sido foco de muita atenção, uma vez que define os sistemas de inteligência artificial classificados como inaceitáveis e que serão portanto proibidos pela União Europeia. O esforço regulatório é desafiador, dada a natureza insidiosa do design comportamental e dos métodos de inferência emocional, além da dificuldade em traçar fronteiras claras entre categorias como autonomia e manipulação ou agência e influência.

Retorna aqui o problema da soberania, já mencionado. Ainda que este texto não pretenda se deter nos pormenores dos marcos regulatórios, nos interessa apontar alguns limites da regulação da Inteligência Artificial no que tange à inferência automatizada de características psicológicas, emocionais e comportamentais voltadas para previsão e intervenção sobre condutas e oportunidades dos sujeitos. Na louvável tentativa de impor barreiras à manipulação ou à influência insidiosa sobre indivíduos e populações, os marcos regulatórios correm o risco de reivindicar, de um lado, um sujeito plenamente autônomo e livre de influências ou, de outro, um sujeito definitivamente vulnerável que precisa ser protegido. Nenhum desses extremos se

sustenta quando se trata de compreender como as subjetividades, os processos psicológicos e emocionais são interpelados em plataformas de IA. Somos simultaneamente produtores e produzidos pelas relações que estabelecemos nos diversos ambientes, sejam estes contextos tecnologicamente mediados ou não. Retomamos assim a pergunta sobre como construir um debate em torno da soberania do sujeito que não perca de vista a dimensão profundamente relacional e compósita das subjetividades. Um debate que reconheça que não somos senhores em nossa própria casa, mas que não se furte a contestar a empreitada extrativa da IA sobre as emoções e o psiquismo.

Nos próximos tópicos, abordaremos os dois domínios focalizados nesse artigo: os sistemas de inferência automatizada de estados emocionais e as estratégias utilizadas pelo design comportamental, agora atualizadas pela adoção de Inteligências Artificiais, para atuar nas brechas da consciência das usuárias a fim de conduzir suas condutas dentro e fora do ambiente digital. Nos dois casos, analisaremos as implicações para a questão da soberania do sujeito.

A INFERÊNCIA ARTIFICIAL DE EMOÇÕES E SUAS CONTROVÉRSIAS

A definição do que seria um 'sistema de reconhecimento emocional' no *AI ACT* consta no Artigo 3, parágrafo 34. Diz o texto: "'Sistema de reconhecimento de emoções' significa um sistema de IA destinado a identificar ou inferir emoções ou intenções de pessoas singulares com base nos seus dados biométricos"[1] (European Parliament, 2024, p. 173). A definição, entretanto, como aponta uma nota redigida pela *Access Now* e outras organizações da sociedade civil, é extremamente vaga e possui imprecisões, além de ser tecnicamente falha. As organizações argumentam que a definição é limitada a sistemas que utilizam dados biométricos relacionados "às características físicas, fisiológicas ou comportamentais de uma pessoa física, que permitam ou confirmem a identificação única dessa pessoa física"[2]. Tomando como exemplo a resposta galvânica da pele, o texto aponta uma limitação chave dessa definição: a utilização de informações de dados fisiológicos não necessariamente permite a identificação de um indivíduo.

O exemplo é ilustrativo. A resposta galvânica da pele mede a atividade elétrica das glândulas que produzem suor nas palmas das mãos. A criação de um método para observar e quantificar a atividade elétrica da pele é considerada o quarto e último elemento que permitiu a criação do polígrafo, conhecido popularmente como o aparelho *detector de mentiras*[3]. A nota da *Access Now* argumenta que desenvolvedores

[1] Livre tradução de: "emotion recognition system' means an AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data".

[2] Ver: https://www.accessnow.org/wp-content/uploads/2022/05/Prohibit-emotion-recognition-in-the-Artificial-Intelligence-Act.pdf

[3] Os outros elementos que compõem as métricas de um aparelho de polígrafo são a análise da pressão arterial, a frequência cardíaca e a amplitude e ritmo respiratórios.

e proprietários poderiam utilizar um dispositivo similar e alegar, embasado na lei, que este não estaria sujeito às obrigações decorrentes da legislação. É uma provocação interessante que nos convoca a pensar além do caso utilizado como exemplo.

O argumento toca em um ponto essencial do debate sobre regulação de IA. A detecção - ou inferência, como preferimos nomear - de emoções não é problemática apenas por ameaçar a privacidade dos indivíduos e coletar seus dados - ainda que estes aspectos sejam sensíveis e de extrema importância para a garantia de direitos. Antes, é preciso perguntar como as inferências emocionais dos sistemas de IA produzem pretensas verdades sobre indivíduos e populações, e de que maneira isso afeta a vida das pessoas.

Como apontamos anteriormente, essa inquietação move a pesquisa que se desenvolve no MediaLab.UFRJ desde 2018. No âmbito das plataformas e aplicações digitais, observamos que a conversão de aspectos psicológicos e emocionais em dados tem se transformado nos últimos anos: uma primeira geração de tecnologias que alimentam a economia psíquica dos algoritmos focalizou os dados comportamentais, as ações e interações *online* (textos, cliques, curtidas, compartilhamentos, postagens, compras, padrões e tempo de navegação, movimentos do mouse, velocidade de digitação) como fontes para a inferência de estados emocionais e psicológicos. Também observamos como foram progressivamente incorporados às interfaces de plataformas e aplicativos uma série de elementos e funcionalidades que buscavam tornar as emoções mais facilmente legíveis para o cálculo computacional, como os botões de curtir, os ícones de reação e os ícones de emoção (*emoticons*) hoje onipresentes (Bruno, Bentes, Faltay, 2019).

Numa plataforma como o *Facebook*, podemos ver claramente a ampliação dessas funcionalidades: o primeiro passo explícito nessa direção se dá em 2009, com o lançamento do botão de "Curtir" (*Like*); em 2013, o espectro de expressão de emoções e estados psíquicos amplia-se com a opção de "Atualização de Status" (*Status Update fields*), permitindo que o usuário utilize uma grande diversidade de ícones gráficos para indicar a tonalidade emocional e psíquica de sua postagem. Na categoria "Sentimento/Atividade" o usuário pode escolher o sentimento que lhe é mais pertinente entre um leque de mais de 200 expressões que correspondem a confiante, inspirado, esperançoso, frustrado, exausto, nostálgico, sexy etc. Em 2016, uma nova funcionalidade emocional passa a acompanhar o já banal botão "*Like*": os "Ícones de Reação" (*Reaction Icons*) permitem que qualifiquemos as postagens dos outros segundo um espectro de seis emoções básicas (Curtir, Amei, Haha, Uau, Triste e Grr) (Cf. Stark, 2018).

No último levantamento de casos, lançado em 2024[4], observamos que além da extração de dados comportamentais, as tecnologias de inferência sobre estados emocionais e características psíquicas passam a se ancorar em outros territórios

[4] Ver: MEDIALAB.UFRJ. Economia Psíquica dos Algoritmos em Linha do Tempo. *Blog do MediaLab.UFRJ*, 2024. Disponível em: https://medialabufrj.net/projetos/economia-psiquica-dos-algoritmos-em-linha-do-tempo/

extrativos, em especial na análise do rosto (microexpressões faciais, movimentos oculares etc.) e no reconhecimento de voz (aspectos e mudanças no timbre e velocidade vocal, por exemplo). Num âmbito mais geral, destacamos na pesquisa uma inflexão nas tecnologias de reconhecimento emocional, que cada vez mais privilegiam a leitura automatizada de emoções corporificadas. Ou seja, tais tecnologias visam inferir aspectos de difícil observação e mensuração (emoções, afetos, caráter, vícios, virtudes, sentimentos, estados psíquicos) a partir de um conjunto de indícios corporais, como expressões faciais, movimentos oculares, postura e gestos corporais, tom e cadência da voz, sinais fisiológicos, batimentos cardíacos, resposta galvânica da pele etc.

Um primeiro problema que atravessa a discussão sobre os instrumentos de mensuração, detecção ou reconhecimento de emoções e de aspectos psicológicos diz respeito ao estatuto desses "objetos". Diferentemente do que certos tecnoentusiastas nos fazem pensar, aspectos psíquicos e emocionais não são naturais e disponíveis de antemão. Tanto o psiquismo quanto as emoções são fortemente atravessados por processos culturais, históricos, sociais e contextuais que tornam inconsistentes qualquer suposição universalizante ou essencializante. Trata-se de processos extremamente relacionais e de difícil apreensão em termos quantificáveis e calculáveis. A perspectiva corrente em torno da detecção automatizada de emoções naturaliza nossa relação com os dados em geral e com os dados psicológicos e emocionais em particular, como se eles fossem uma fonte de informação natural que antecede as operações de extração (Cf. Boehner et al 2007; Ven, 2017; Rhue, 2018; Barrett et al., 2019; Crawford, 2021; Stark e Hoey, 2021). Em outras palavras, os dados psíquicos e emocionais são, como qualquer outro dado, produzidos. Entender a sua cadeia de produção e o modo como aspectos psicológicos e emocionais são convertidos em dados que sejam legíveis para o processamento computacional (Cf. Bruno, 2022) é essencial para avaliar os sistemas automatizados de reconhecimento de emoções e suas implicações individuais e coletivas.

Vejamos o caso do rosto, que vem se convertendo no território extrativo mais conhecido e visado das ferramentas de inferência emocional (McStay, 2018; Crawford, 2021). As tecnologias de reconhecimento facial de emoções alegam ter a capacidade de, ao detectar um rosto na imagem, traduzir automaticamente as expressões em parâmetros numéricos análogos às expressões de emoção supostamente reconhecidas como universais. A base desses sistemas é a "teoria da universalidade das emoções", proposta pelo psicólogo americano Paul Ekman a partir de pesquisas desenvolvidas com o povo Fore, na Papua Nova Guiné. Como os Fores eram considerados isolados e tinham pouco ou quase nenhum contato com o mundo ocidental, Ekman acreditava que a pesquisa poderia provar se as emoções são expressadas e compreendidas da mesma forma em diferentes contextos culturais. Apesar das grandes dificuldades na condução geral do experimento, uma vez que o psicólogo pouco conhecia o idioma e a cultura Fore, os resultados foram posteriormente interpretados como bem-sucedidos, sustentando o argumento de que as emoções são inatas e universais (Ekman, Friesen, 1971).

As fragilidades do trabalho de Ekman e as contestações que recebeu à época, sobretudo pela pretensão à universalidade, não impediram que o modelo *FACS*[5] e o das seis emoções básicas (posteriormente alterado para sete)[6] tenha se tornado o paradigma dominante nas tecnologias de inferência emocional, respaldando a concepção de que as emoções são dados comportamentais passíveis de serem observados e medidos de forma objetiva (Crawford, 2021). Ou seja, a escolha pelo modelo de Ekman não se deve ao fato de ela ser considerada a melhor, mais correta ou mais completa teoria sobre as emoções humanas. Como argumenta McStay (2018), o modelo de Ekman funciona muito bem para os tecnólogos porque a suposição de que as emoções são inatas e mensuráveis universalmente por expressões faciais comuns permite que as máquinas possam ser treinadas para identificá-las. Mesmo o setor estando ciente de que essas abordagens são problemáticas, o fato de funcionarem bem com a tecnologia de visão computacional fez com que os limites metodológicos fossem deliberadamente ignorados, ou seja, "há uma simplicidade atraente nas emoções básicas que os tecnólogos têm se agarrado"[7] (McStay, 2018, p. 19). Kate Crawford faz uma análise semelhante quando afirma que "as teorias de Ekman pareciam ideais para o campo emergente da visão computacional porque podiam ser automatizadas em grande escala"[8] (Crawford, 2021, p. 175).

Para além do debate científico em torno da associação emoção-expressão facial, diferentes autores apontam seus problemas políticos e tendências discriminatórias. Ao longo da história, a associação entre determinados traços faciais e estados mentais serviu como base para teorias que contribuíram para a legitimação científica de preconceitos e hierarquias de raça e gênero. Esse ponto é levantado em críticas recentes às tecnologias de biometria e reconhecimento facial, apontando que tais dispositivos têm raízes históricas em práticas como a frenologia, a fisiognomia, a antropometria criminal e a eugenia (Lyon, 2001; Browne, 2010; Birhane, 2021).

Associando características físicas e raciais a características mentais ou morais – como tendência à criminalidade e nível de inteligência –, essas teorias serviram para justificar a defesa de diferenças hierárquicas entre as raças, legitimando uma série de opressões e violências coloniais. O trabalho de Rhue (2018) é especialmente importante ao chamar atenção para a disparidade racial na análise de emoções. Ao analisar os sistemas de reconhecimento facial de emoções da *Microsoft* e da *Face++,* a pesquisadora demonstrou que alguns softwares tendem a atribuir a pessoas negras emoções consideradas mais negativas, como raiva e desdém.

[5] Facial Action Coding System (FACS) é um sistema desenvolvido por Paul Ekman e Wallace Friesen em 1976 para descrever todos os movimentos faciais visualmente discerníveis.

[6] As seis emoções básicas e universais seriam felicidade, tristeza, raiva, medo, surpresa e nojo. Posteriormente, uma sétima emoção foi incluída no modelo, o desprezo.

[7] Livre tradução de: "there is an attractive simplicity to ‘basic emotions’ (Ekman and Friesen, 1971) that technologists have latched onto".

[8] Livre tradução de: "Ekman’s theories seemed ideal for the emerging field of computer vision because they could be automated at scale".

Diversos autores questionam, portanto, não somente os pressupostos de que as emoções seriam inatas e universais, mas a própria ideia de que estados emocionais podem ser mensuráveis e quantificáveis, apontando os problemas e riscos desse tipo de projeto[9] (Boehner et al 2007; Barrett, 2017; Barrett et al, 2019). Nos últimos anos, o debate em torno dessas bases pseudocientíficas e discriminatórias levou ativistas a defenderem o banimento de tecnologias de visão computacional que utilizam reconhecimento facial ou de afetos e emoções (*AI Now Institute*, 2019; *Ada Lovelace Institute*, 2019). Houve também importantes iniciativas em apontar os dilemas éticos das IAs emocionais, como a cartilha de diretrizes para uso ético da Inteligência Artificial Emocional desenvolvida por McStay e Pavliscak (2019).

Em resposta a essas críticas, o texto final do *AI Act* da União Europeia decidiu proibir o uso de sistemas de reconhecimento de emoções em locais de trabalho e instituições de ensino, além de classificar o uso dessas tecnologias em outros ambientes como de alto risco. O texto, entretanto, abre uma exceção ao uso de sistemas de inferência emocional para razões médicas e de segurança, o que significa que eles ainda poderão ser utilizados em práticas bastante contestáveis, como o policiamento preditivo.

No Brasil, a Coalizão Direitos na Rede também sugere a proibição dos sistemas de reconhecimento de emoções. A organização argumenta que a proposta de permitir o uso de sistemas de reconhecimento de emoções ou categorização biométrica, desde que sejam informadas sobre essa utilização, acaba por admitir o uso dessas tecnologias apesar dos riscos inerentes ao uso da IA para esse fim, tanto pela falta de "fundamento científico de que seja possível identificar emoções somente com base em expressões faciais", como pela violação à intimidade do indivíduo (Coalizão Direitos na Rede, 2023).

Para além da discussão sobre a fragilidade científica, técnica e operacional dos atuais sistemas de inferência emocional, consideramos importante chamar atenção para um outro ponto: a dimensão performativa tanto desses sistemas como dessas teorias. Esta preocupação fica especialmente evidente quando observamos o potencial uso da classificação de emoções para a tomada de decisões que afetarão o futuro das pessoas – como embasar decisões de contratação e demissão de trabalhadores em alguma empresa ou a avaliação de desempenho acadêmico em instituições de ensino (Crawford, 2021; Israel e Firmino, 2023).

A preocupação aumenta se considerarmos que essas tecnologias incidem diretamente sobre a vida coletiva e individual. Os resultados de sistemas de IA, usados em larga escala e atendendo a interesses corporativos específicos, tendem a ser categóricos.

[9] A neurocientista Lisa Feldman Barrett conduziu uma extensa pesquisa em torno das evidências científicas da associação entre expressões faciais e expressões emocionais. Barrett argumenta que as emoções são expressões aprendidas em determinadas culturas e contextos, e seus estudos concluem que não há evidências de que as emoções podem ser detectadas – seja por humanos, seja por máquinas –, a partir de padrões de movimentos faciais (Barrett, 2017; Barrett et al., 2019). Ao comentar as tentativas de detecção automatizada dos afetos, a autora pontua que as empresas de tecnologia podem ser capazes de detectar movimentos do rosto (como um sorriso, uma testa franzida), mas isso não é a mesma coisa que detectar emoções.

Ou seja, o fato de os conhecimentos produzidos por estas tecnologias de inferência emocional serem cientificamente contestáveis e reducionistas não significa que suas consequências sejam menos nocivas (Stark e Hoey, 2021). Ao contrário. Considerar aspectos involuntários - os micromovimentos da face, por exemplo - como evidências de estados emocionais para a produção de decisões envolvendo indivíduos e populações irá, inevitavelmente, retirar dos sujeitos a possibilidade de negociação e contestação dos critérios segundo os quais suas características e potencialidades são definidas. O risco de se aprofundar assimetrias e injustiças no acesso a serviços, direitos e oportunidades é patente.

Nota-se como a questão da soberania do sujeito está em jogo. Há, nos ambientes mediados por sistemas de inferência artificial de emoções, uma clara redução da margem que os sujeitos têm de conhecer, deliberar, negociar e contestar os processos de tomada de decisão a seu respeito. Como encaminhar o problema da soberania nesses contextos? Que forma de soberania poderia ser aí reivindicada? Como compreender as relações entre soberania, autonomia e determinação do sujeito em ambientes tecnologicamente mediados por inteligência artificial? Fiquemos com essas perguntas, que serão retomadas na última seção deste texto. Vejamos, agora, como esse golpe na soberania do sujeito não é privilégio dos sistemas automatizados de análise facial das emoções. Como notaremos no próximo tópico, o design comportamental há muito explora as brechas da consciência dos usuários.

DESIGN CAPCIOSO[10] E O DRIBLE DA CONSCIÊNCIA

Legislações recentes vêm questionando um aspecto da construção dos produtos digitais há muito invisibilizado pelos interesses comerciais das grandes corporações e agora exacerbado pela potência das Inteligências Artificiais: a utilização de práticas de design capcioso, técnicas herdadas da economia comportamental para driblar a consciência dos usuários e garantir maior engajamento. Como sabemos, este último permite tanto escalar a coleta de dados quanto orientar desejos e monetizá-los com propagandas direcionadas. Notaremos a seguir como, nos últimos 20 anos, essa matriz epistemológica comportamental (Faltay, 2020) penetrou nos mais diversos campos, carregando consigo a premissa de que as atitudes são definidas, majoritariamente, não por escolhas racionais e autorreflexivas, mas como fruto de vieses cognitivos (Kahneman, 2012), de modo que os sujeitos seriam "previsivelmente irracionais" (Ariely, 2008).

Tomemos como exemplo os sistemas de recomendação. Em um primeiro momento, o principal alvo de interesse eram as declarações explícitas dos usuários a respeito de suas preferências, na forma de avaliações, comentários e likes. Mas essas logo passam a ser consideradas insuficientes (McNee, Riedl, Konstan, 2006), por três motivos primordiais. O primeiro deles é temporal: as avaliações são bons indicadores da relação pregressa das pessoas com um conteúdo, mas dizem pouco sobre o porvir. Esse

[10] Design capcioso é a tradução que propomos para o termo em inglês *deceptive design,* também traduzido para o português como "design malicioso" (Calonga et al, 2022).

"aprisionamento" ao passado, que limita a capacidade de incidir sobre as possibilidades futuras, deixa de ser suficiente em um mercado cujo modelo de negócio depende cada vez mais da assertividade da previsão comportamental. Em busca de sucesso no *mercado de comportamentos futuros* (Zuboff, 2021), interessa às empresas a adoção de um modelo muito mais performativo, que intervenha sobre a conduta dos usuários em tempo real, enquanto elas acontecem.

O segundo ponto fraco das declarações explícitas que vigoravam nos primórdios da coleta de dados para recomendações algorítmicas está diretamente ligado à consolidação do paradigma conexionista do qual vínhamos falando: trata-se de sua parca quantidade. De modo geral, poucas pessoas dão opinião sobre aquilo que consumiram, e nesse modelo de coleta é preciso contar com a boa vontade dos usuários na concessão das avaliações para que se possa extrair o conhecimento necessário para oferecer novas recomendações (Oard; Kim, 1998). Além disso, direcionar a recomendação pela expectativa da acuidade depende de um corolário: para que se possa avaliar a performance da máquina, as pessoas precisam manifestar também as contraprovas das recomendações de maneira explícita (Faltay, 2020). Por si só, a predição de como as pessoas irão classificar um item contribui para a escassez de dados das populações existentes e dificulta a montagem de perfis de grandes populações (Oard; Kim, 1998). Não se mostrando interessante, portanto, para um modelo de negócios cujas virtudes são a rapidez e a flexibilidade no processamento de grande volume de dados pessoais.

O terceiro inconveniente das declarações explícitas diz respeito a sua suposta falibilidade em comparação com aquilo que pode ser observado e concluído de forma automatizada pelo processamento maquínico. Este traço faz parte de um deslocamento epistemológico mais amplo no qual a reflexão crítica característica do ideal iluminista de sujeito começa a ser sobrepujada por uma nova forma de racionalidade algorítmica (Bruno, 2022) cujo conhecimento, calcado em cálculos probabilísticos, promete mais objetividade, velocidade e eficácia. Em meio a esta mudança na forma de saber e em consonância com as teorias behavioristas, ganha corpo a crença de que as declarações individuais seriam incontornavelmente contaminadas de imprecisões. A fim de prever comportamentos futuros, o ideal, portanto, seria evitá-las.

A resposta do mercado aos inconvenientes das declarações explícitas veio na forma de uma nova estratégia na qual a coleta de dados, as práticas de marketing, o design das plataformas e os sistemas de recomendação algorítmica passam a ser guiados por uma matriz epistemológica comportamental com o objetivo de influenciar e persuadir os usuários a realizar determinadas ações e não outras (Bentes, 2022). Nick Seaver (2018) chama de virada captológica o momento em que os sistemas de recomendação, retratado por ele como armadilhas, abandonam métricas baseadas na assertividade das previsões das avaliações em favor de métricas implícitas extraídas tacitamente do comportamento. As plataformas passam a privilegiar a produção de métricas a partir da identificação de padrões observáveis da interação das pessoas com suas interfaces (Oard; Kim, 1998). São percebidos como evidências dos juízos e preferências das

pessoas, por exemplo, o tempo médio gasto em posts ou vídeos, a pausa em um vídeo, pular uma música recomendada ou um story etc.

Zuboff (2021) nomeia essa camada de dados abandonados não intencionalmente de *shadow text*, cujo papel é central na geração do *superávit comportamental* necessário para alimentar a inteligência das máquinas. Segundo a autora, por trás de um "primeiro texto" legível formado por *posts*, músicas, fotos, mensagens, *likes* e toda forma de declaração, haveria uma camada opaca de rastros involuntários inapreensível ao olhar humano. Justamente pelo caráter virtualmente infinito e pretensamente verdadeiro desse segundo texto, ele ganha cada vez mais relevância em relação ao primeiro. A condução das condutas e as estratégias para influenciar os indivíduos e populações voltam-se, então, para a criação de ambientes de captura que buscam driblar, sempre que possível, a consciência dos usuários. Empresas investem no design de interfaces e na organização da arquitetura da informação para a elaboração de um sistema adaptativo e flexível cujos algoritmos procuram capturar e analisar os dados na maior velocidade possível. A eficiência de um mecanismo de direcionamento de conteúdo passa, então, a ser medida pela capacidade em capturar a atenção e reter pelo máximo de tempo a pessoa conectada (Bentes, 2022).

Essa mesma lógica de captura que passa a orientar os sistemas de recomendação se manifesta de forma similar também em outras áreas. A *virada comportamental* destacada por Nadler e McGuigan (2017) diz respeito ao impacto da economia comportamental nas práticas de marketing digital. Os autores evidenciam o duplo discurso feito pela indústria: para evitar a regulação da coleta de dados, as agências de marketing argumentam que os usuários são indivíduos racionais e soberanos que cedem conscientemente seus dados em troca de ofertas "relevantes"; por outro lado, essas mesmas agências vendem para os seus clientes técnicas herdadas da economia comportamental que pressupõem os tais indivíduos "previsivelmente irracionais" (Ariely, 2008).

Os exemplos supracitados evidenciam uma empreitada na qual o sujeito reflexivo característico da racionalidade moderna passa a ser colocado sob suspeita. Suas decisões conscientes, lentas e eventualmente contraditórias que garantiriam sua "soberania" não convêm ao ritmo acelerado do processamento maquínico e devem ser dribladas ou superadas. Essa superação se dá, como vimos, principalmente por meio da incorporação de uma matriz epistemológica comportamental em cada uma das camadas de construção dos produtos digitais, com o objetivo de conduzir a conduta dos usuários em tempo real. É no design comportamental das plataformas, no entanto, que esse drible da consciência se torna mais evidente, justamente porque tangibiliza ideias abstratas (modelos de sujeito, consciência, autonomia, vieses cognitivos etc.) na forma de padrões que podem ser nomeados e reconhecidos na materialidade das interfaces.

Em 2010, à luz das estratégias cada vez mais explícitas das empresas para induzir comportamentos nos produtos digitais, Harry Brignull (2010) cunha o termo *dark*

patterns[11] para se referir aos truques usados por *designers* de sites e aplicativos para compelir os usuários a determinadas ações irrefletidas, como realizar uma compra ou assinar um serviço. O caráter obscuro do nome não vinha por acaso; embora a dissimulação já fizesse parte do manual da economia digital há bastante tempo, as técnicas através das quais ela se dava ainda eram desconhecidas do público em geral e evidentemente não regulamentadas. Faltavam ainda seis anos para Tristan Harris ganhar notoriedade denunciando as 10 formas através das quais as tecnologias estariam "sequestrando a mente das pessoas"[12] e outros quatro para que o debate acerca da manipulação ganhasse as mesas de bar na esteira do lançamento de *O dilema das redes* (2020). Em suma, em 2010 os vieses cognitivos ainda não eram um território de batalha moral entre benfeitores paternalistas e manipuladores malvados, ambos tendo em comum a crença em duas premissas epistemológicas básicas: 1) o ser humano tem atributos dados que devem ser desvelados (e não produzidos na relação com o mundo, com os artefatos e demais viventes); e 2) esse ser humano é também uma criatura de hábitos, preguiçosa e esquecida (Hall, 2013) cujas vulnerabilidades são claramente definidas, podendo então ser a) protegidas ou b) exploradas (Fogg, 2002; Eyal, 2014; Harris, 2016).

Em grande medida graças ao trabalho de Brignull (2010), que na última década e meia se dedicou incansavelmente a revelar e nomear essas práticas de design até então invisibilizadas, hoje os *dark patterns* estão não só nos fóruns de *UX design*, mas também no centro das discussões sobre regulação das plataformas digitais. Isso porque ilustram de forma didática o drible na consciência e a decorrente perda da agência que conferiria ao sujeito a soberania sobre suas condutas e escolhas em ambientes digitais.

Divididos em 16 categorias que incluem falsa urgência, obstrução, prevenção de comparação, falsa escassez e confirmação da vergonha (*confirmshaming*), os padrões obscuros - que agora se chamam padrões enganosos (*deceptive patterns*) em resposta a uma evolução da sua natureza cuja dissimulação já prescinde da obscuridade original - estão espalhados pelos mais diferentes tipos de produtos digitais camuflando, dificultando ou seduzindo as pessoas para atender às prioridades de negócios das empresas às quais os produtos pertencem. Encontramos um exemplo bastante elucidativo em uma atualização recente do Uber de maio de 2023 que esconde o botão de cancelamento, antes visível de antemão desde o momento da solicitação, atrás de duas outras ações: primeiro, é preciso clicar no ícone de "Informações da viagem"; em seguida, ele demanda que se informe o motivo do cancelamento. Só então, após o pedágio de dois cliques e uma justificativa, nos é oferecida a possibilidade da ação inicialmente desejada. Note que esse fluxo independe de algum motorista aceitar a corrida, fato este que poderia explicar o inconveniente "humano" do cancelamento que a plataforma busca evitar. Aqui temos um caso ilustrativo das barreiras ao

[11] Ver: https://90percentofeverything.com/2010/07/08/dark-patterns-dirty-tricks-designers-use-to-make-people-do-stuff/index.html

[12] Ver: medium.com/how-technology-hijacks-peoples-minds-from-a-magician-and-google-design-ethicist

cancelamento, que são ubíquas o suficiente para terem sido agraciadas com uma categoria particular de padrão enganoso (*deceptive pattern*) chamada "*hard to cancel*".

Esses exemplos ilustram o "drible da consciência" performado pela indústria digital, que agora ganha novas camadas em meio à proliferação de tecnologias de Inteligência Artificial que prometem uma adaptação das interfaces em tempo real de acordo com o contexto e a situação psíquica e emocional de cada um. A mudança do termo "*dark pattern*" para "*deceptive design*", observada pelo pesquisador Mark Leiser (2024), ilustra bem o enraizamento das práticas enganosas, que deixam a superfície para se embrenhar nas camadas mais profundas das máquinas inteligentes. Antes, os truques de manipulação se davam de forma observável nas interfaces dos sites e aplicativos; por mais que fossem dissimulados, como sugeriria o nome, uma vez apontados e "desmascarados" era possível ao olhar humano reconhecê-los. Agora, temos um segundo estágio no qual as práticas evoluem para estratégias mais sofisticadas de modulação constante na arquitetura dos sistemas, que acontecem mais rápido do que somos capazes de perceber e afetam a experiência do usuário a longo prazo. Em suma, trata-se de um *design capcioso* que limita a autonomia do sujeito uma vez que direciona as escolhas e comportamentos de forma sutil, discreta e atraente.

Essa rapidez cada vez maior de adaptação das máquinas inteligentes, cujo efeito ainda não conhecemos completamente, alija o sujeito do conhecimento que é produzido sobre ele próprio, reduzindo sua autonomia na medida em que o usuário digital é visto como um objeto (Fisher, 2022), uma presa (Seaver, 2018) ou um ambiente para a agência de sistemas não humanos (Cesarino, 2022) que, como tal, pode ser capturado pelo ímpeto extrativista de empresas do norte global. Assim como as nações reivindicam sua autodeterminação informacional, seria bem-vindo um movimento coletivo de emancipação similar em relação às engrenagens maquínicas que tentam driblar nossas consciências. Essa emancipação, no entanto, reconhece a impossibilidade de se estar plenamente no controle de si, já apontada por Freud em 1917, ao mesmo tempo em que aposta na construção de um ecossistema sociotécnico rico em conexões capazes de fortalecer coletivamente a autonomia e soberania dos sujeitos e subjetividades, num trabalho contínuo de co-produção, conforme retomaremos na próxima seção.

CONSIDERAÇÕES FINAIS

Nos tópicos anteriores, vimos como aspectos psicológicos e emocionais tornam-se alvos privilegiados para extração de dados que alimentam a racionalidade algorítmica e os sistemas de inteligência artificial das grandes corporações de tecnologia. Destacamos como os sistemas automatizados de inferência emocional baseiam-se em modelos epistemológicos reducionistas, frágeis e controversos sobre as emoções humanas, tornando seus efeitos não apenas potencialmente imprecisos e falhos, como assimétricos e injustos. Também ressaltamos, a esse respeito, duas inquietações que merecem ser retomadas. A primeira é a impossibilidade de se estabelecer uma definição universal das emoções humanas que possa estar na base seja de sistemas de

detecção emocional de alcance massivo, seja de leis que limitem ou regulem seus usos. A segunda é a performatividade dos sistemas de categorização das emoções humanas, que historicamente produziram efeitos de verdade sobre indivíduos e populações, incidindo tanto sobre o modo como as pessoas entendem a si mesmas, quanto sobre as formas como são classificadas socialmente, com impactos discriminatórios sobre suas vidas, corpos e oportunidades. Para dar um exemplo amplamente conhecido, mulheres e populações racializadas foram historicamente discriminadas e sujeitas a violências físicas, políticas e sociais com base em teorias sobre o caráter excessivo, indomesticável ou agressivo de suas emoções (Cf. Shields, 2020).

À performatividade dos sistemas de categorização soma-se a performatividade dos sistemas algorítmicos e de inteligência artificial na modulação e orientação de condutas e decisões, na incitação de hábitos, desejos e interesses, já fartamente documentada (Cf. Bruno, 2013; Gillespie, 2014; Grosman, Reigeluth, 2019; Bucher, 2018). Incorporar sistemas de inferência emocional em processos maquínicos de corporações que operam em larga escala envolve, no mínimo, um duplo cinismo que se retroalimenta: utilizar, de um lado, métodos cientificamente contestáveis para escalar a extração de dados e o engajamento de usuários; e lucrar, de outro lado, com a performatividade desses sistemas alegando estar personalizando e otimizando a qualidade dos seus serviços em favor de uma melhor experiência do usuário.

Diante dessas inquietações e outras apontadas ao longo do texto, não há margem para a construção de condições ou dispositivos legais que garantam aos sujeitos implicados a suficiente proteção ou a efetiva agência sobre tais mecanismos de inferência acerca de seus estados psicológicos e emocionais. Em outros termos, não há margem para qualquer nível de autonomia ou soberania do sujeito e das subjetividades nesse domínio. Tal governo algorítmico, em última instância, retira as margens de negociação e contestação, ou a própria "capacidade do sujeito de lutar contra os dispositivos estabelecidos para conduzi-lo" (Alves; Andrade, 2022, p. 1019). Desse modo, ressoamos as proposições pela moratória e/ou banimento de sistemas automatizados de detecção emocional não apenas nos campos do trabalho e da educação, como também da segurança e de todo o espectro de atuação da inteligência artificial por corporações de tecnologia.

Outro problema explorado ao longo do texto com implicações para o debate sobre regulação de IA diz respeito ao estatuto do sujeito reflexivo. Vimos como a empreitada extrativa de dados psicológicos e emocionais liderada pelas grandes corporações de tecnologia digital e de inteligência artificial colocam sob suspeita o sujeito reflexivo característico da racionalidade moderna. Tanto no caso da leitura automatizada de emoções em marcadores biométricos, manifestações corporais e fisiológicas quanto no design de plataformas e aplicações digitais, notamos como decisões conscientes, intrinsecamente mais lentas e eventualmente contraditórias se tornam uma barreira ao ritmo frenético do processamento maquínico que deve ser constantemente superada. Essa superação se dá, como vimos, incorporando estratégias comportamentais em diversas camadas dos produtos digitais, buscando orientar a conduta dos usuários em tempo real.

Leis recentes como o já mencionado Artigo 5 do *AI Act* Europeu e a própria proposta brasileira em avaliação citam profusamente termos como "autonomia individual", "persuasão", "manipulação psicológica", "vulnerabilidades", "padrões enganosos" e "dano", em um duplo movimento que morde, mas também assopra: reconhece os riscos associados aos "padrões obscuros" de design e desenvolvimento mas, ao fazê-lo, também referenda uma perspectiva em que as pessoas possuem atributos (comportamentos, pensamentos, desejos, vulnerabilidades, emoções) relativamente pré-definidos que estão à disposição; sempre prontos para ser desvelados e manejados por ferramentas e práticas mais ou menos eficazes, com as melhores ou as piores intenções (Faltay, 2020).

A constatação de que essas práticas algorítmicas capciosas operam nas margens da consciência humana tem animado o debate sobre os perigos da manipulação e da perda da autonomia individual (Susser; Roessler; Nissenbaum, 2019; Alves; Andrade, 2022). Ainda que as denúncias nessa direção sejam necessárias, elas não são suficientes. Não é raro encontrarmos, no fundo dessas denúncias, um apelo a ideais de autonomia ou de transparência bastante questionáveis, que ressoam a racionalidade neoliberal. O foco na manipulação supõe que em algum lugar de nós reside um sujeito autônomo e capaz de escolher sem nenhum tipo de influência. Ou que poderíamos criar um mundo em que relações de poder magicamente desapareceriam com regras e normas de conduta, boas práticas e transparência (Bruno, 2022). Na tentativa de preservar o que resta do sujeito reflexivo em tempos de ameaça sem precedentes ao excepcionalismo humano, há o risco de se perder de vista algo que consideramos fundamental no entendimento da relação que nós, humanos, mantemos com os arranjos sociotécnicos que produzimos.

A questão não é ser ou não influenciado, mas quais redes de influências queremos construir (Bruno, 2022). É nesse sentido que propomos encaminhar o debate sobre soberania nos âmbitos do sujeito e da subjetividade. Por isso, é fundamental criar as condições para o exercício de um cuidado coletivo com nosso ecossistema sociotécnico, com as relações e redes de interdependência que nos constituem. Nos constituímos e nos transformamos todo o tempo na relação que mantemos com nós mesmos e com tantos outros - humanos e não humanos - que nos circundam. Atualmente, o modelo hegemônico da IA e das grandes corporações de tecnologia cria assimetrias brutais em nossas redes de influências, concentrando conhecimento e poder em um número restrito de atores com uma agenda de negócios específica, homogênea e sem nenhum compromisso com o bem comum. Nos ambientes digitais, plataformas e aplicações de IA visados nesse artigo, a soberania do sujeito ou da subjetividade não pode ser assegurada de modo definitivo no plano individual nem no plano jurídico. Ela é, antes, um problema sociotécnico e tecnopolítico coletivo a ser continuamente trabalhado e revisado.

Ressoando mais uma vez algumas proposições do Hui (2024), a soberania é, em geral, uma noção ambivalente. Por um lado, reivindicar uma soberania tecnológica, digital e de dados é importante e necessário para fazer frente à hegemonia do poder econômico, político e social das corporações de tecnologia. Por outro lado, a soberania

é historicamente assombrada, especialmente no âmbito dos estados-nação, pela tendência ao exercício de um poder que se pretende absoluto, centralizado e acima de leis internacionais.

Hui sugere, retomando Derrida, uma farmacologia da soberania. Tal como a leitura que este último autor faz do pharmakon no Fedro de Platão (2016), a escrita – e a técnica em geral – é tanto remédio quanto veneno. A escrita, como exteriorização da memória, impede o esquecimento e também faz esquecer, uma vez que deixa à mão o que precisa ser lembrado. Ela também permite que os discursos viajem para além dos corpos dos seus enunciadores, ampliando seu alcance, mas ao mesmo tempo os discursos se tornam órfãos e mais sujeitos a equívocos, uma vez que seus "pais" ou autores nem sempre estão presentes para defendê-los. De modo similar, uma farmacologia da soberania ressalta que ela tanto pode "impor uma tendência totalizante" quanto pode "usar esse poder para se proteger de influências externas, como o imperialismo" (Hui, 2024, p. 247).

Tal perspectiva farmacológica pode ser interessante para pensar também a soberania do sujeito e da subjetividade em plataformas de IA. Por um lado, a reivindicação da soberania deve ser tomada com cautela, dada a sua tendência em reafirmar a perigosa ilusão de indivíduos capazes de se autodeterminar. Por outro, ela pode fazer frente ao empenho da plataformização e da racionalidade algorítmica em definir unilateralmente os termos de nossa experiência em ambientes digitais. Ela também pode fazer frente à "monocultura da mente" (Shiva apud Hui, 2024) onipresente na lógica capitalística global e na monotecnologia impulsionada pelas plataformas digitais globais (Hui, op.cit). Além de reduzir as margens de contestação e agência dos sujeitos, tais plataformas incitam uma forte homogeneização das condutas, hábitos e práticas online. A soberania dos sujeitos e das subjetividades tem menos a ver com a manutenção ou restituição de um suposto poder perdido do que com a capacidade de diferir, transformar e reimaginar modos de viver, pensar e se relacionar com as tecnologias. Nos termos de Hui, "a tecnodiversidade se opõe ao apocalipse tecnológico e, em vez de convergência, propõe uma divergência ou bifurcação" (Hui, 2024, p. 225).

Juntamente com a perspectiva farmacológica, que não busca esconder nem resolver a ambivalência implicada em qualquer forma de soberania, vale também chamar a atenção para o desafio de reimaginar uma soberania da subjetividade e do sujeito sem recair no antropocentrismo. Recolocar o problema da soberania do sujeito e das subjetividades não pode se confundir, portanto, com a restituição de poder ao humano. Como já apontamos, trata-se, antes, de reconstruir um ecossistema sociotécnico, subjetivo e ambiental que leve a sério o emaranhado de relações e co-dependência entre humanos, viventes e entes técnicos.

Não se trata, portanto, de reafirmar ou defender um sujeito soberano nos moldes da racionalidade moderna, mas de tomar a sério a questão da soberania do sujeito e da subjetividade como problema sociotécnico e coletivo e, assim, como tarefa continuada. Reivindicar maior participação e agência na produção de conhecimento,

nas relações e nos processos de tomada de decisão nos ambientes sociotécnicos que habitamos não significa retornar à crença em um Eu capaz de orquestrar unilateralmente suas próprias condições de possibilidade. No sentido aqui proposto, tanto a soberania quanto a autonomia não implicam ser o senhor exclusivo de escolhas, ações ou decisões, mas sim a possibilidade de construir coletivamente as condições onde essas escolhas, ações e decisões – eventualmente enviesadas ou equivocadas – se dão. Por isso, também é fundamental garantir condições de dissenso e reconsideração de práticas, regras, parâmetros, protocolos e normas etc. A construção coletiva da soberania digital, tecnológica, de dados ou das subjetividades tem menos a ver com a defesa de atributos pré-definidos do que com a possibilidade de tecer sistemas sociotécnicos permeáveis à negociação, contestação, diferenciação e revisão.

FINANCIAMENTO

Este trabalho foi realizado no âmbito do projeto "Economia Psíquica dos Algoritmos: racionalidade, subjetividade e conduta em plataformas digitais", coordenado pela Profa Fernanda Bruno no MediaLab.UFRJ e financiado pela Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). As autoras Alice Lerner e Helena Strecker agradecem, respectivamente, à CAPES e à Faperj pelas bolsas de mestrado. Paulo Faltay agradece à Fundação de Amparo à Ciência e Tecnologia de Pernambuco (Facepe) e ao CNPq pela bolsa de pós-doutorado do Programa de Desenvolvimento Científico e Tecnológico Regional (PDCTR).

REFERÊNCIAS

ADA LOVELACE INSTITUTE, 2019. Beyond face value: public attitudes to facial recognition technology. *Ada Lovelace Institute*, 02 setembro 2019. [Acesso em 16 março 2024]. Disponível em: https://www.adalovelaceinstitute.org/report/beyond-face-value-public-attitudes-to-facial-recognition-technology/

AI NOW INSTITUTE, 2019. AI in 2019: A Year in Review. *AI Now Institute*, 9 outubro 2019. [Acesso em: 16 mar. 2024]. Disponível em: https://ainowinstitute.org/news/ai-in-2019-a-year-in-review

ALLEN, Seamus, 2021. European Sovereignty In: *The Digital Age. Europe's Digital Future*. [Acesso em 29 setembro 2024]. Disponível em: https://www.iiea.com/publications/european-sovereignty-in-the-digital-age

ALVES, Marco Antonio Souza ; ANDRADE, Otávio Morato de, 2022. Autonomia individual em risco? Governamentalidade algorítmica e a constituição do sujeito. Cadernos Metrópole, v. 24, n. 55, p. 1007–1024, 2022. [Acesso em: 10 dez. 2024]. Disponível em: https://www.scielo.br/j/cm/a/MhymSLPFzLcpSbWFcYBdpqy/abstract/?lang=pt

ARIELY, Dan, 2006. *Predictably irrational: The hidden forces that shape our decisions.* New York, NY: HarperCollins Publishers.

BARRETT, Lisa Feldman, 2017. *How Emotions Are Made: The Secret Life of the Brain*. Boston, MA: Houghton Mifflin Harcourt.

BARRETT, Lisa Feldman; ADOLPHS, Ralph; MARSELLA, Stacy; et al, 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest, v. 20, n. 1, p. 1–68, 2019. [Acesso em 10 dez. 2024]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/31313636/

BARBOSA, Alexandre, 2022. A soberania digital sustentável como base para o futuro da Internet. *Comciência*. 2022. [Acesso em 29 setembro 2024]. Disponível em: https://www.comciencia.br/a-soberania-digital-sustentavel-como-base-para-o-futuro-da-internet

BENTES, Anna, 2022. Da Madison Avenue ao Vale do Silício: ciências comportamentais do engajamento, tecnologias de influência e economia da atenção. 2022. Tese (Doutorado em em Comunicação e Cultura) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ.

BIRHANE, Abeba, 2021. Algorithmic injustice: a relational ethics approach. *Patterns*. 2021. v. 2, n. 2, p. 100205–100205. [Acesso em 28 fevereiro 2024]. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666389921000155

BOEHNER, Kirsten; DEPAULA, Rogério; DOURISH, Paul; SENGERS, Phoebe. How emotion is made and measured. International Journal of Human-Computer Studies, v. 65, n. 4, p. 275–291, Apr. 2007. [Acesso em: 16 dezembro 2024] Disponível em: https://doi.org/10.1016/j.ijhcs.2006.11.016.

BRIGNULL, Harry, 2010. Dark Patterns: dirty tricks designers use to make people do stuff. *90 Percent Of Everything*, 8 jul. 2010. [Acesso em: 6 maio 2024]. Disponível em: https://90percentofeverything.com/2010/07/08/dark-patterns-dirty-tricks-designers-use-to-make-people-do-stuff/index.html

BRUNO, Fernanda, 2018. A economia psíquica dos algoritmos: quando o laboratório é mundo. *Nexo jornal*, 12 de junho de 2018. [Acesso em 19 fevereiro 2024]. Disponível em: https://www.nexojornal.com.br/a-economia-psiquica-dos-algoritmos-quando-o-laboratorio-e-o-mundo

BRUNO, Fernanda; BENTES, Anna; FALTAY, Paulo, 2019. Economia psíquica dos algoritmos e laboratório de plataforma: mercado, ciência e modulação do comportamento. *Revista FAMECOS*. 2019. v. 26, n. 3. [Acesso em 19 fev 2024]. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/revistafamecos/article/view/33095

BRUNO, Fernanda, 2022. Racionalidade algorítmica & subjetividade maquínica. IN: SANTAELLA, Lucia (Org.). *Simbioses do Humano e Tecnologias: Impasses, Dilemas, Desafios*. São Paulo, SP: Editora da Universidade de São Paulo/IEA-USP.

BROWNE, Simone, 2010. Digital Epidermalization: Race, Identity and Biometrics. *Critical Sociology*. 2010. v. 36, n. 1, 131–150. [Acesso em 1 maio 2024]. Disponível em: doi:10.1177/0896920509347144

BUCHER, Taina, 2018. *If… Then: Algorithmic power and politics.* Oxford, England: Oxford University Press.

CALONGA, Luiz Octavio Lanssoni; SOARES, Carla D. M.; MELO, Thiago Coelho de; MACHADO, Luciano Marchi. Pensa que me Engana, eu Finjo que Acredito: Padrões Obscuros sob a Perspectiva do Usuário. XLVI Encontro da ANPAD - EnANPAD 2022 On-line - 21 - 23 de set de 2022. [Acesso em 10 dez. 2024]. Disponível em: https://anpad.com.br/uploads/articles/120/approved/8a88d5f412f2ad376f8597d28cbd3720.pdf

CARDON, Dominique; JEAN-PHILIPPE COINTET; MAZIÈRES, Antoine; et al, 2018. Neurons spike back. *Reseaux*. 2018. v. 211, n. 5, p. 173–220. [Acesso em 6 maio 2024]. Disponível em: https://www.cairn-int.info/abstract-E_RES_211_0173--neurons-spike-back.htm

COALIZÃO DIREITOS NA REDE, 2023. Projeto de lei nº 2338/2023 - Nota técnica. *Coalizão Direitos na Rede*, 23 agosto 2023. [Acesso em: 5 mai. 2024]. Disponível em: https://direitosnarede.org.br/2023/08/23/coalizao-direitos-na-rede-divulga-nota-tecnica-sobre-o-pl-2338-2023-que-busca-regular-a-ia/

CRAWFORD, Kate, 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven: Yale University Press.

D'ALMONTE, E. F., & SANTOS, A, 2024. O. Regulamentação das plataformas digitais: entre a soberania digital e o transnacionalismo. *E-Compós*. 2024, ahead of print. [Acesso em 1 out. 2024]. Disponível em: https://www.e-compos.org.br/e-compos/article/view/2876

EKMAN, Paul; FRIESEN, Wallace V, 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*. 1971. v. 17, n. 2, 124–129. [Acesso em 1 maio 2024]. Disponível em: https://doi.org/10.1037/h0030377

EUROPEAN PARLIAMENT, 2024. Artificial Intelligence Act. 13 mar. 2024. [Acesso em 24 abril 2024]. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

EYAL, Nir, 2014. *Hooked: How to Build Habit-Forming Products*. New York: Portfolio, 2014. ISBN 978-, 1591847786.

FALTAY, Paulo, 2020. Máquinas paranoides e sujeito influenciável: conspiração, conhecimento e subjetividade em redes algorítmicas. Tese (Doutorado em Comunicação e Cultura) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ.

FOGG, B. J, 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*. Amsterdam: Morgan Kaufmann. ISBN 978-1558606432.

FREUD, Sigmund, 2010. Uma dificuldade da psicanálise (1917). Em: *História de uma neurose infantil ("o homem dos lobos"), além do princípio do prazer e outros textos*. São Paulo, SP: Companhia das Letras.

GAGO, Veronica; MEZZADRA, Sandro, 2017. A Critique of the Extractive Operations of Capital: Toward an Expanded Concept of Extractivism. *Rethinking Marxism*. 2017. v.4,

n. 29, p 574-591. [Acesso em 1 maio 2024]. Disponível em: https://doi.org/10.1080/08935696.2017.1417087

GILLESPIE, Tarleton, 2014. The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). MIT Press.

GROSMAN, Jéremy; REIGELUTH, Tyler. Perspectives on algorithmic normativities: engineers, objects, activities. *Big Data & Society*, n. 6, v. 2, 2019. https://doi.org/10.1177/20539517198587

HALL, Erika, 2013. *Just Enough Research*. New York: A Book Apart.

HARRIS, Tristan, 2016. How technology hijacks people's minds—from a magician and Google's design ethicist. *Medium [online]*. [S. l.]: Thrive Global, 2016. [Acesso em 10 dez. 2024] Disponível em: https://medium.com/thrive-global/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3.

HELMOND, Anne, 2015. The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society*. 2015. v. 1, n. 2. [Acesso em 6 maio 2024]. Disponível em: https://journals.sagepub.com/doi/10.1177/2056305115603080

HUI, Yuk, 2020. *Tecnodiversidade*. São Paulo, SP: Ubu Editora.

HUI, Yuk, 2024. *Machine and sovereignty: for a planetary thinking*. Minneapolis, MN: University of Minnesota Press.

ISRAEL, Carolina Batista; FIRMINO, Rodrigo [coords.], 2023. Reconhecimento facial nas escolas públicas do Paraná - relatório 2023. Curitiba : UFPR. [Acesso em 10 dez. 2024]. Disponível em: https://jararacalab.org/relatorio-rf-pr/

JAMES, William. The Principles of Psychology, Volume 2 (of 2)." Mind, v. 2, p. 567, 2004.

KAHNEMAN, Daniel, 2012. *Rápido e Devagar: Duas Formas de Pensar*. Rio de Janeiro, RJ: Editora Objetiva.

LEISER, Mark, 2024. Psychological Patterns and Article 5 of the AI Act: AI-Powered Deceptive Design in the System Architecture and the User Interface. *Journal of AI law and Regulation*. 2024. v.1, n.1, p. 5-23. [Acesso em 20 abril 2024]. Disponível em: https://research.vu.nl/en/publications/psychological-patterns-and-article-5-of-the-ai-act-ai-powered-dec

LYON, David, 2001. Under My Skin: From Identification Papers to Body Surveillance. J. Caplan and J. Torpey (eds) *Documenting Individual Identity: The Development of State Practices in the Modern World*, pp. 291–310. Princeton University Press: Princeton, NJ.

MCNEE, Sean M; RIEDL, John ; KONSTAN, Joseph A, 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. *Proceedings of the 2006 Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, April 22-27, 2006. [Acesso em 06 maio 2024]. Disponível em: https://dl.acm.org/doi/10.1145/1125451.1125659

MCSTAY, Andrew, 2018. *Emotional AI: The rise of empathic media*. London: SAGE Publications Ltd.

MCSTAY, Andrew; PAVLISCAK, P, 2019. Emotional Artificial Intelligence: Guidelines for Ethical Use. [Acesso em 1 maio 2024]. Disponível em: https://drive.google.com/file/d/1frAGcvCY_v25V8ylqgPF2brTK9UVj_5Z/view

NADLER, Anthony; MCGUIAN, Lee, 2017. An impulse to exploit: the behavioral turn in data-driven marketing. *Critical Studies in Media Communication*. 2017. v. 35, n. 2, pp. 151–165. [Acesso em 6 maio 2024]. Disponível em: https://doi.org/10.1080/15295036.2017.1387279

OARD, Douglas W; KIM, Jinmook, 1998. Implicit Feedback for Recommender Systems. *Scholar Commons*. 1998. [Acesso em 6 maio 2024]. Disponível em: https://scholarcommons.sc.edu/libsci_facpub/111/

PLATÃO, 2016. *Fedro*. São Paulo, SP: Editora 34.

RICAURTE, Paola, 2022. Ethics for the majority world: AI and the question of violence at scale. *Media, Culture & Society*. 2022. v. 44, n. 4. [Acesso em 22 abril 2024]. Disponível em: https://journals.sagepub.com/doi/full/10.1177/01634437221099612

RHUE, Lauren, 2018. Racial Influence on Automated Perceptions of Emotions. *Social Science Research Network*, 2018 [Acesso em 1 maio 2024]. Disponível em: http://dx.doi.org/10.2139/ssrn.3281765

SEAVER, Nick, 2018. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*. 2018. v. 24, n. 4, p. 421–436. [Acesso em 25 outubro 2021]. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/1359183518820366

SHIELDS, Stephanie, 2002. *Speaking from the heart: Gender and the social meaning of emotion*. Cambridge University Press, 2002.

SILVEIRA, Sérgio Amadeu da, 2023. Inteligência local, soberania digital e soberania de dados. In: PENTEADO, C. C; PELLEGRINI, J. C.; SILVEIRA, Sérgio Amadeu da (Orgs.) *Plataformização, inteligência artificial e soberania de dados*. 1. ed. São Paulo: Ação Educativa, v. 1. 198p .

STARK, Luke, 2018. Algorithmic Psychometrics and the Scalable Subject. *Social Studies of Science*. 2018. v. 48, n. 2, p. 204-231. [Acesso em 1 maio 2024]. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3154778

STARK, Luke; HOEY, Jesse. 2021. The Ethics of Emotion in Artificial Intelligence Systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 782–793. [Acesso em 13 dezembro 2024] Dispoível em: https://doi.org/10.1145/3442188.3445939

TURING, Alan, 1950. Computing Machinery and Intelligence. *Mind, New Series*. 1950. Vol. 59, No. 236. p. 433-460.

VEN, Ruben van de. Choose How You Feel; You Have Seven Options. Institute of network cultures, jan. 2017. [Acesso em 12 dezembro 2024] Disponível em:

http://networkcultures.org/longform/2017/01/25/choose-how-you-feel-you-have-seven-options/.

ZUBOFF, Shoshana, 2021. *A era do capitalismo de vigilância: a luta por um futuro humano na nova fronteira do poder*. Rio de Janeiro, RJ: Intrínseca.