IX CONGRESSO BRASILEIRO DE ESTUDOS ORGANIZACIONAIS

O HOMEM ADMINISTRATIVO NA ERA ALGORÍTMICA: RACIONALIDADE LIMITADA E INTERAÇÃO HUMANO-IA NAS ORGANIZAÇÕES

PEDRO HENRIQUE GOMES DE MATTOS[1]; ELCIO GUSTAVO BENINI[2]

[1]*Universidade Federal de Mato Grosso do sul – pedro.mattos@ufms.br*
[2]*Univesidade Federal de Mato Grosso do Sul – elciobenini@yahoo.com.br*

Resumo
O presente ensaio teórico investiga o acúmulo de "dívida cognitiva" gerada pelo uso intenso e acrítico de sistemas de Inteligência Artificial (IA) no cotidiano organizacional. Fundamentado na teoria da racionalidade limitada de Herbert Simon, no modelo de processamento dual de Daniel Kahneman e na arquitetura de escolhas de Thaler e Sunstein, o estudo argumenta que a IA, paradoxalmente, aumenta a sobrecarga de microinterações e introduz vieses sistêmicos nas premissas de decisão. Esse processo esgota a atenção e a autonomia deliberativa do usuário por meio de três mecanismos cumulativos: as dívidas de atenção, de verificação e de agência. Como alternativa para mitigar esse passivo tecnológico e comportamental sem restringir a liberdade de escolha, o trabalho propõe uma tipologia estruturada de seis *nudges* algorítmicos, com destaque para a fricção estratégica e o enquadramento calibrado, acompanhada de cinco princípios normativos de *design* comportamental. A análise conclui que, para preservar a agência humana, reverter a aversão ao algoritmo e impedir a manipulação epistêmica, o *design* dessas plataformas deve transcender a engenharia de *software* e ser tratado como um desafio institucional e ético. Exige-se governança, transparência e auditoria contínua de vieses, reconfigurando a IA de uma mera substituta automatizada da decisão para uma autêntica matriz amplificadora da racionalidade humana.

Palavras-chave: Inteligência Artificial; Dívida Cognitiva; Arquitetura de Escolhas; Racionalidade Limitada; Nudges Algorítmicos.

Contextualização

Em 1971, Herbert A. Simon escreveu que "em um mundo rico em informação, a riqueza de informação cria pobreza de atenção" (Simon, 1971, p. 40–41). A observação, cunhada na gênese da computação pessoal, adquire precisão cirúrgica no contexto de 2026: Hoje, interagir com dezenas de sistemas de inteligência artificial ao longo de um único dia de trabalho é a tônica do trabalho contemporâneo, assistentes generativos, algoritmos de recomendação(seja em redes sociais ou pesquisa), plataformas de automação de tarefas, ferramentas de apoio à decisão clínica, financeira e jurídica. Cada interação demanda escolha. Cada escolha consome atenção. E a atenção, como Simon já sabia, é o recurso mais escasso da organização.

Denominamos dívida cognitiva o passivo acumulado que resulta dessa sobrecarga: decisões postergadas, julgamentos superficiais, dependência irrefletida de saídas automáticas e erosão progressiva da capacidade deliberativa do usuário. Dívida cognitiva é aqui empregado em analogia à dívida técnica da engenharia de software: um passivo acumulado de escolhas cognitivas degradadas que precisará ser revisitado com custo elevado. O fenômeno não é necessariamente novo, Simon (1947) já descrevia os limites da racionalidade, mas a IA o amplifica por dois mecanismos simultâneos: (a) aumenta a frequência das interações decisórias, cada entrada, ou prompt, é, em última análise, uma escolha . e (b) pode introduzir vieses sistêmicos nas premissas que apresenta ao usuário.

A questão que orienta este artigo é: como desenhar interações de IA que, ao mesmo tempo, reduzam a dívida cognitiva do usuário e preservem sua autonomia decisória?. A hipótese central é que a teoria do nudge, em sua versão mais precisa de arquitetura de escolhas (Thaler; Sunstein, 2008), pode contribuir com o instrumental adequado para redesenhar essas interações, desde que informada pelos insights de racionalidade limitada (Simon, 1947), processamento dual (Kahneman, 2011), aversão ao algoritmo (Dietvorst; Simmons; Massey, 2015) e viés algorítmico (O'Neil, 2016).

Para fundamentar essa argumentação, é necessário reconstituir os pilares conceituais que orientam a pesquisa. Por mais que cada tradição tenha trajetória autônoma, todas compartilham um pressuposto metodológico comum: o decisor real, não o agente ideal da teoria econômica clássica, opera em ambientes de complexidade irredutível e recursos cognitivos finitos.

A ruptura de Herbert Simon com a teoria do *homo economicus* é radical e deliberada. Em *Administrative Behavior* (1947) obra originada de sua tese de doutorado na Universidade de Chicago, Simon demonstra que o decisor organizacional jamais dispõe do conjunto completo de alternativas, consequências e critérios que a otimização racional exigiria. Suas capacidades são limitadas em três dimensões fundamentais: informação (o ambiente é ambíguo e os dados, parciais), cognição (a capacidade de calcular, comparar e armazenar é finita) e tempo e atenção (decidir custa tempo; a atenção é recurso escasso).

O conceito central resultante é o de racionalidade limitada (*bounded rationality*): a racionalidade humana é intencional, mas limitada pelas condições de processamento do decisor. Isso implica que o decisor não maximiza, isto é, delibera de acordo com um modelo completo do cenário, afinal, um modelo com todas as variáveis não se distinguiria da realidade em si. O que ocorre é mais pragmaticamente, ele satisfaz (*satisfices*, neologismo de Simon que combina *satisfy* e *suffice*). Define um nível de aspiração e aceita a primeira alternativa que o atinge, interrompendo a busca. "A teoria administrativa é, na sua essência, a teoria da racionalidade intencional e limitada do comportamento dos seres humanos que contemporizam porque não possuem meios para maximizar" (SIMON, 1979 [1947], Introdução à 2ª ed.).

Além do *satisficing*, Simon (1947; 1960) distingue decisões programadas, repetitivas, cobertas por rotinas e procedimentos, de decisões não-programadas novas, ambíguas, que exigem deliberação consciente. A organização existe, em parte, para transformar decisões não-programadas em programadas, liberando atenção para os problemas genuinamente novos. É precisamente nesse ponto que Simon antecipa, em linguagem organizacional, a dualidade

que Kahneman formalizaria anos depois: rotinas e hábitos correspondem ao processamento automático; a deliberação explícita, ao processamento controlado.

Daniel Kahneman, em *Thinking, Fast and Slow* (2011), sistematiza décadas de pesquisa com Amos Tversky em torno de um modelo de processamento cognitivo dual. O Sistema 1 (S1) opera de forma automática, rápida, intuitiva e sem esforço consciente; o Sistema 2 (S2) é lento, deliberativo, analítico e exige atenção direcionada. Embora a terminologia seja didática, não há dois "sistemas" anatomicamente separados, a distinção descreve adequadamente padrões de comportamento decisório bem documentados.

O S1 é o sistema dominante no cotidiano: a maior parte das decisões diárias é tomada por meio de heurísticas, atalhos cognitivos que produzem respostas rápidas e frequentemente adequadas, mas sujeitas a erros sistemáticos (vieses). Entre os vieses mais relevantes para o contexto de IA destacam-se: o viés de automação (tendência a confiar em *outputs* computacionais sem escrutínio), o viés de confirmação (busca preferencial por informação que corrobora crenças prévias) e o efeito de ancoragem (uso de valores inicialmente fornecidos, como a saída de um modelo de linguagem, como referência para julgamentos subsequentes).

O S2, por sua vez, consome recursos escassos (atenção, tempo, glicose metabólica). Kahneman denomina esse fenômeno de depleção do ego, e sua implicação prática é direta: quanto mais decisões triviais são impostas ao S2, menor a qualidade das decisões que genuinamente o requerem, o chamado esgotamento decisório (*decision fatigue*). A dívida cognitiva, em termos Kahnemanianos, é o resultado de um S2 continuamente recrutado para tarefas que deveriam ter sido estruturadas para o S1.

Richard Thaler e Cass Sunstein, em *Nudge: Improving Decisions About Health, Wealth, and Happiness* (2008), introduzem o concept de arquitetura de escolhas: o reconhecimento de que toda escolha ocorre em um ambiente estruturado, e que esse ambiente pode ser deliberadamente desenhado para tornar as decisões melhores sem proibir alternativas nem alterar os incentivos econômicos de forma substancial. Um *nudge* é qualquer aspecto da arquitetura de escolhas que altera o comportamento de maneira previsível, sem proibir nenhuma opção e sem mudar significativamente os incentivos.

O exemplo canônico é a definição do padrão (*default*): se um plano de saúde é o *default* de inscrição automática, mais pessoas aderem do que quando a inscrição é *opt-in*. O padrão "captura" o S1 sem coagir o S2. A posição filosófica de Thaler e Sunstein é o que denominam paternalismo libertário: intervenções que preservam a liberdade de escolha, ao mesmo tempo que orientam o decisor para opções que, segundo a evidência disponível, promovem seu bem-estar. O *nudge* não é manipulação, ele não altera preferências; estrutura o ambiente para que preferências já latentes se traduzam mais facilmente em ação. Os instrumentos canônicos da arquitetura de escolhas incluem: *defaults* (opções padrões de ajuste, pré-selecionar uma escolha específica), enquadramentos (*framing*), simplificação da informação, lembretes contextuais, normas descritivas e *feedback* em tempo real. Todos operam prioritariamente sobre o S1.

Uma complicação importante para qualquer projeto de *nudge* algorítmico é o fenômeno descrito por Berkeley J. Dietvorst, Joseph P. Simmons e Cade Massey (2015) como

aversão ao algoritmo (*algorithm aversion*): após observar um algoritmo errar, mesmo que o algoritmo seja mais preciso que o julgamento humano na média, as pessoas tendem a abandoná-lo em favor de julgamento humano, ainda que este se prove estatisticamente inferior para aquela tarefa específica. Dietvorst et al. demonstram que a aversão é assimétrica: os usuários toleram erros humanos com maior facilidade do que erros algorítmicos. O mecanismo subjacente é psicológico: os algoritmos criam expectativas mais elevadas de perfeição e, quando as violam, produzem reação de desconfiança desproporcional.

Um segundo artigo dos mesmos autores (2016) demonstra que permitir ao usuário pequenas modificações na saída do algoritmo, mesmo que irrelevantes para o resultado, reduz significativamente a aversão ao reforçar a sensação de controle. A aversão ao algoritmo tem implicações diretas para o design de *nudges*: um *nudge* algorítmico percebido como impositivo pode ser rejeitado precisamente por ativar a desconfiança descrita. O design deve, portanto, calibrar a quantidade de ação preservada ao usuário para sustentar a legitimidade da recomendação.

A literatura sobre viés algorítmico alerta para uma dimensão crítica que qualquer projeto de arquitetura de escolhas mediada por IA precisa enfrentar: algoritmos não são neutros. Eles incorporam os vieses presentes nos dados de treinamento, nas escolhas de design dos desenvolvedores e nas estruturas sociais que os dados refletem. Cathy O'Neil, em *Weapons of Math Destruction* (2016), documenta como modelos preditivos utilizados em decisões de crédito, contratação e justiça criminal amplificam desigualdades históricas ao tratar correlações espúrias como causas. Barocas e Selbst (2016) aprofundam essa análise no plano jurídico-conceitual ao demonstrar que variáveis aparentemente neutras podem funcionar como *proxies* de raça, classe social ou gênero, produzindo efeitos discriminatórios sem intenção discriminatória. Esse fenômeno, denominado impacto disparatado (*disparate impact*), é propriedade emergente de sistemas treinados em dados que refletem estruturas históricas de desigualdade. Noble (2018) documenta empiricamente o problema em sistemas de busca e recomendação, demonstrando que algoritmos otimizados para engajamento amplificam representações estereotipadas.

A convergência entre O'Neil (2016), Barocas e Selbst (2016) e Noble (2018) delineia um imperativo metodológico: o design de *nudges* algorítmicos não pode prescindir de auditoria de viés como etapa constitutiva do processo, não acessória. Para os fins deste artigo, o viés algorítmico representa um paradoxo estrutural: a IA projetada para ser arquiteta de escolhas pode, simultaneamente, ser uma arquiteta enviesada, estruturando o ambiente decisório em direção a opções que sistematicamente desfavorecem determinados grupos ou amplificam vieses do usuário em vez de corrigi-los. Isso não invalida o projeto de *nudge* algorítmico, mas o condiciona: uma boa arquitetura de escolhas em IA exige auditoria contínua de vieses como condição de legitimidade.

A síntese deste referencial pode ser expressa em uma tensão produtiva: a IA opera no espaço exato que Simon, Kahneman e Thaler mapearam, o espaço entre a racionalidade ideal e a decisão real. Essa posição pode ser usada para reduzir dívida cognitiva (*nudge* construtivo) ou para explorá-la (*nudge* manipulativo). O viés algorítmico adiciona uma terceira possibilidade: o dano pode ser não-intencional, produzido por dados e modelos que refletem estruturas sociais injustas.

Paço Cunha (2022), em análise do percurso epistemológico de Herbert Simon, demonstra que o conceito de racionalidade limitada não constitui uma inovação original do autor, mas o produto de uma longa filiação intelectual. O itinerário reconstrói como Simon absorveu, em sequência, o empirismo de David Hume, o pragmatismo de William James e John Dewey, o behaviorismo cognitivo de Edward Tolman, a economia institucionalista de John R. Commons e o pensamento administrativo de Chester Barnard. A tese central é que Simon herdou de Commons e Barnard a noção fundamental de que o comportamento decisório é sempre situado institucionalmente: a organização, como ambiente estruturado, condiciona o que conta como resposta "satisfatória" para fins de *satisficing*. A racionalidade limitada não é apenas um limite cognitivo individual, é também uma construção institucional.

A implicação aqui é de ordem estrutural. Se a racionalidade limitada é, em sua própria gênese, um fenômeno institucional, e não apenas cognitivo, então sistemas de IA que medeiam decisões cotidianas não podem ser tratados como meras ferramentas de engenharia comportamental individual. Plataformas digitais e assistentes de IA constituem, nesse sentido, novas instituições: ambientes que estruturam o que é cognitivamente acessível, que alternativa aparece primeiro, quais consequências são tornadas salientes e quais são ocultadas. A arquitetura de escolhas que Thaler e Sunstein (2008) identificaram em ambientes físicos e burocráticos está sendo reconfigurada, em escala massiva e com velocidade de adaptação sem precedentes históricos, por arquiteturas algorítmicas.

Essa reconfiguração institucional ganha sustentação teórica na própria definição simoniana de ambiente organizacional, a qual estabelece que as instituições funcionam como a sistematização do comportamento dos indivíduos por meio da sujeição de suas ações a um sistema de estímulos socialmente impostos. Sob esse prisma, as arquiteturas algorítmicas de escolha não atuam de forma isolada; elas assumem o papel de matrizes institucionais que coordenam a docilidade e as decisões no ambiente digital, convertendo o *prompt* e o *output* nos novos estímulos estruturantes dessa racionalidade situada. O itinerário simoniano sugere, portanto, que a resposta adequada ao problema da dívida cognitiva não é apenas individual (aprimorar a cognição do usuário) nem apenas técnica (otimizar o algoritmo), mas também institucional: que tipo de ambiente decisório as plataformas de IA estão construindo? A essa pergunta, Paço Cunha (2022) adiciona uma dimensão histórica essencial: o decisor sempre foi moldado pela instituição; a novidade radical da IA é que a instituição agora aprende, adapta e otimiza em tempo real, tornando o ambiente decisório simultaneamente mais responsivo e mais opaco às intenções de quem o habita.

A dívida cognitiva, tal como introduzida neste artigo, designa o passivo acumulado de decisões cognitivamente degradadas que resulta da sobrecarga de interações decisórias e da dependência irrefletida de sistemas de IA. Operacionalmente, pode ser decomposta em três modalidades que se reforçam mutuamente.

A primeira é a dívida de atenção: o usuário delega ao sistema decisões que deveriam passar por deliberação consciente, poupando atenção no curto prazo à custa de um escrutínio que não ocorrerá. O Sistema 2, para retomar a linguagem de Kahneman (2011), é sistematicamente excluído de interações que o requerem. Um assistente de escrita que sugere conclusões antes que o usuário articule sua própria posição exemplifica esse mecanismo: a

conveniência imediata encurta o processo de construção do argumento e, no longo prazo, atrofia a capacidade argumentativa do próprio usuário.

A segunda modalidade é a dívida de verificação: o usuário aceita *outputs* de IA como referências sem submetê-los a checagem independente. Isso é agravado pelo fenômeno da alucinação, geração de informações plausíveis, mas factualmente incorretas, característico dos modelos de linguagem de grande escala (LLMs). À medida que esses sistemas se tornam a fonte primária de consulta em crescente número de domínios, o custo da verificação se desloca para o próprio usuário, que raramente dispõe dos recursos cognitivos ou do tempo necessários. No horizonte mais longo, Shumailov et al. (2024) demonstram que o treinamento recursivo de modelos sobre dados gerados por outros modelos produz colapso de modelo (*model collapse*): a distribution estatística dos dados se degenera progressivamente, e a capacidade de o sistema refletir a realidade se deteriora de forma acumulativa e imperceptível. A dívida de verificação individual agrega-se, assim, em um problema epistemológico coletivo: a degradação silenciosa das referências compartilhadas.

A terceira modalidade é a dívida de agência: a delegação repetida de julgamento ao sistema reduz, progressivamente, a capacidade do usuário de deliberar de forma autônoma. O mecanismo é o da depleção seletiva de competências: quanto menos o usuário pratica determinados tipos de raciocínio, avaliação crítica de fontes, síntese de perspectivas contraditórias, estimativa sob incerteza, menos proficiente se torna neles. A aversão ao algoritmo descrita por Dietvorst, Simmons e Massey (2015) pode ser reinterpretada, nesse quadro, como um mecanismo compensatório: a resistência ao sistema após a observação de erros é, em parte, uma reação de preservação da agência cognitiva, ainda que frequentemente mal calibrada e ineficaz na redução da dependência estrutural.

Objetivos

A questão fundamental que orienta este artigo é: como desenhar interações de IA que, ao mesmo tempo, reduzam a dívida cognitiva do usuário e preservem sua autonomia decisória no ambiente organizacional?

O objetivo central é investigar e propor mecanismos de mitigação dessa dívida por meio da teoria do *nudge*, em sua versão mais precisa de arquitetura de escolhas (Thaler; Sunstein, 2008), como instrumental adequado para redesenhar essas interações. A hipótese central que sustenta este objetivo é a de que intervenções de design, desde que informadas pelos *insights* de racionalidade limitada (Simon, 1947), processamento dual (Kahneman, 2011), aversão ao algoritmo (Dietvorst; Simmons; Massey, 2015) e viés algorítmico (O'Neil, 2016), podem atenuar o esgotamento decisório nas organizações.

Metodologia

O quadro analítico proposto neste artigo é, por construção, de natureza teórica e conceitual. O percurso metodológico consistiu na articulação e revisão crítica de tradições teóricas da economia comportamental, teoria organizacional e ética algorítmica para a estruturação de novas proposições (tipologia e princípios de design).

Contudo, como agenda de desenvolvimento para a plena utilidade prática da teoria construída, o avanço empírico desta pesquisa depende de validação em ao menos quatro dimensões metodológicas futuras:

Primeira dimensão: Verificação empírica da tipologia de dívida cognitiva. Estudos de campo e experimentos controlados são necessários para mensurar os mecanismos de geração de dívida. Escalas psicométricas de carga cognitiva (Hart; Staveland, 1988) e paradigmas de dupla tarefa (Strayer; Johnston, 2001) oferecem instrumentos estabelecidos para esse fim.

Segunda dimensão: Eficácia comparativa dos *nudges* propostos. Cada um dos tipos de *nudge* da tipologia é uma hipótese testável. Designs A/B em plataformas reais, combinados com medidas de qualidade decisória, tempo de revisão e satisfação, permitirão estimar efeitos e interações.

Terceira dimensão: Moderadores individuais e culturais. A literatura de *nudge* tem sido criticada por resultados heterogêneos entre populações. É esperado que a eficácia dos *nudges* algorítmicos varie em função de letramento digital, tolerância à incerteza, cultura organizacional e histórico de uso de IA.

Quarta dimensão: Efeitos de longo prazo sobre a capacidade deliberativa. A hipótese de atrofia cognitiva por delegação habitual exige estudos longitudinais, visto que a dívida cognitiva diferida não é mensurável em um único ponto no tempo.

Resultados Esperados (Proposições)

Diante do diagnóstico da dívida cognitiva, propomos como resultado uma tipologia de *nudges* algorítmicos. O artigo organiza instrumentos de arquitetura de escolhas especificamente desenhados para o ambiente de interação com IA, com o objetivo de reduzir as modalidades de dívida cognitiva sem suprimir a autonomia do usuário. Seis tipos são identificados:

O *nudge* de *default* adaptativo: Apresenta uma opção pré-configurada representando o comportamento mais seguro ou mais frequentemente revisado pelo usuário, com possibilidade explícita de modificação. Diferentemente do *default* estático de Thaler e Sunstein (2008), o *default* adaptativo calibra-se com o histórico do usuário, tornando a sugestão contextualmente relevante sem impor padrões universais.

O *nudge* de fricção estratégica: Interpõe uma pausa ou uma etapa adicional antes de ações irreversíveis ou de alto impacto. Enquanto a fricção não-intencional é fonte de dívida cognitiva, a fricção deliberada recruta o Sistema 2 nos momentos em que ele é genuinamente necessário, um sistema que solicita confirmação explícita antes de enviar um documento redigido por IA em nome do usuário exemplifica esse tipo.

O *nudge* de enquadramento calibrado: Apresenta a saída do sistema acompanhada de indicadores de incerteza, limitações conhecidas ou fontes divergentes, transformando a recomendação algorítmica em convite à deliberação em vez de substituto dela. Combate diretamente o viés de automação.

O *nudge* de norma descritiva: Exibe informações sobre o comportamento de pares ("83% dos usuários revisaram esta sugestão antes de aplicá-la") para ativar o mecanismo social de conformidade em direção a práticas de maior escrutínio, transpondo para o ambiente algorítmico o mesmo mecanismo que Thaler e Sunstein (2008) documentaram em contextos físicos.

O *nudge* de retroalimentação de dívida cognitiva: Provê ao usuário visualizações periódicas de seu padrão de uso, frequência de aceitação não revisada de sugestões, proporção de decisões delegadas por domínio, buscando constituir uma consciência metacognitiva sobre a acumulação de dívida. A transparência é aqui o mecanismo do *nudge*, não a prescrição de comportamento.

O *nudge* de *affordance* de modificação: Garante ao usuário a capacidade de ajustar a saída do algoritmo antes de aplicá-la, mesmo que marginalmente. Baseado em Dietvorst, Simmons e Massey (2018), este tipo reduz a aversão ao algoritmo ao reforçar a percepção de controle sem comprometer a qualidade estatística da recomendação.

A tipologia não é exaustiva, mas delimita um espaço de design orientado a um objetivo comum: *nudges* que, no lugar de contornar a deliberação do usuário, a sustentam e a restauram. A efetividade desta tipologia está condicionada, contudo, ao estado do viés algorítmico subjacente. Um *nudge* de enquadramento calibrado que apresenta *outputs* de um sistema cujos dados de treinamento são enviesados não corrige o viés, apenas torna o usuário ciente de uma incerteza que pode ser sistematicamente distorcida.

A arquitetura de escolhas em IA opera, portanto, em dois níveis que precisam ser simultaneamente gerenciados: o nível da interação, onde os *nudges* propostos atuam sobre o comportamento do usuário, e o nível do modelo, onde auditoria de viés, diversidade de dados e mecanismos de *accountability* governam a qualidade do que é apresentado ao usuário como premissas decisória. A perspectiva institucional de Paço Cunha (2022) acrescenta um terceiro nível de análise: a plataforma como ambiente decisório que define o próprio horizonte do que pode ser escolhido. *Nudges* bem desenhados numa plataforma que incentiva dependência e maximiza engajamento têm sua eficácia sistematicamente solapada pela lógica institucional do ambiente. O design do *nudge* algorítmico, para ser consequente, precisa estar incorporado em uma arquitetura institucional comprometida com a autonomia cognitiva do usuário como objetivo central, não como externalidade positiva ocasional. Essa condição conecta o problema técnico do design de interação ao problema político de governança das plataformas digitais, abrindo uma agenda de pesquisa que este artigo apenas delimita.

A tipologia anterior pode ser organizada em torno de **cinco princípios de *design* comportamental**, que constituem a agenda normativa deste artigo:

Princípio 1: Compatibilidade com o modo cognitivo: o design deve mapear explicitamente quais tarefas requerem S1 e quais requerem S2, e estruturar a interface para roteá-las ao modo adequado. Tarefas repetitivas e de baixa consequência devem ser fluídas; tarefas de alta consequência devem ter fricção proporcional.

Princípio 2: Transparência calibrada: o sistema deve comunicar sua incerteza de forma compreensível para o S1, não apenas para especialistas. Marcadores linguísticos como

"alta confiança", "estimativa preliminar" ou "requer verificação" ativam processos de escrutínio sem sobrecarregar o usuário com probabilidades técnicas.

Princípio 3: Agência distribuída: toda saída do sistema deve ser apresentada como insumo/recurso para a decisão do usuário, não como decisão em si. O sistema é arquiteto de premissas, na linguagem de Simon, não substituto do decisor.

Princípio 4: Comportamento Auditável: o sistema deve registrar e disponibilizar ao usuário o histórico de suas interações, padrões de aceitação e revisão, e indicadores de dívida cognitiva acumulada. Sem essa camada de reflexividade, o *nudge* pode facilmente tornar-se manipulação, o usuário perde a capacidade de questionar os padrões que o sistema está reforçando.

Princípio 5: Combate contínuo ao viés: qualquer arquitetura de *nudge* algorítmico deve ser acompanhada de protocolo de auditoria de viés algorítmico. O viés sistêmico no modelo anula qualquer ganho comportamental do *nudge*: de nada vale estruturar bem uma escolha se as opções apresentadas refletem discriminação histórica amplificada.

A proposta de *nudge* algorítmico enfrenta três críticas éticas que não podem ser ignoradas. A primeira é a crítica da manipulação epistêmica: *nudges* que operam sobre o S1 sem transparência para o S2 podem moldar preferências sem que o usuário seja capaz de as reconhecer como moldadas (Hausman; Welch, 2010). A diferença entre *nudge* legítimo e manipulação está na reversibilidade e na transparência: o usuário deve poder reconhecer a estrutura e optar por sair dela. A segunda crítica é a da distribuição assimétrica de poder: o arquiteto de escolhas na IA é a empresa desenvolvedora, não o usuário. Diferentemente de políticas públicas, onde o arquiteto é o Estado. Em sistemas privados de IA não existe mecanismo equivalente de prestação de contas. Isso exige regulação específica, não apenas boa intenção de design. A terceira crítica diz respeito à escala: um *nudge* em uma política pública afeta a população de um município ou país; um *nudge* embutido em um sistema de IA amplamente utilizado afeta potencialmente bilhões de interações diárias. A magnitude do efeito agregado implica que erros de design, mesmo não-intencionais, têm consequências sociais sem precedente histórico comparável.

O viés algorítmico representa o limite estrutural mais sério ao projeto de *nudge*. Se o sistema de IA que opera como arquiteto de escolhas incorpora vieses históricos, raciais, de gênero, socioeconômicos, o *nudge* amplifica desigualdades em vez de reduzi-las. O Princípio 5 proposto (Combate Contínuo ao viés) não é optativo: é condição de legitimidade de todo o projeto. A literatura sugere que o problema do viés não tem solução técnica definitiva (Kleinberg et al., 2016): há *trade-offs* entre diferentes definições matemáticas de equidade que não podem ser simultaneamente satisfeitos. Isso implica que o enviesamento não é um problema de engenharia que se resolve uma vez, é um processo político e normativo contínuo que exige participação de grupos afetados.

Conclusões

Este artigo partiu de uma observação de Simon (1971) "A riqueza de informação cria pobreza de atenção", para construir um quadro analítico do fenômeno que denominamos dívida cognitiva no uso de IA: o passivo acumulado de decisões postergadas, delegadas ou degradadas que resulta da sobrecarga gerada por sistemas inteligentes no cotidiano. A contribuição central do artigo é tripla. Primeiro, integra quatro tradições teóricas, racionalidade limitada (Simon), processamento dual (Kahneman), arquitetura de escolhas (Thaler; Sunstein) e aversão ao algoritmo (Dietvorst et al.), em um quadro coerente, com viés algorítmico (O'Neil; Barocas e Moritz) como condicionante crítico. Segundo, propõe uma tipologia de seis *nudges* algorítmicos aplicáveis a interfaces de IA. Terceiro, enuncia cinco princípios de *design* comportamental que traduzem a teoria em orientações operacionais.

A tese simoniana de que a organização existe para compensar as limitações da racionalidade humana ressurge, reformulada: a IA pode ser a extensão cognitiva mais robusta já construída para esse fim, ou o seu mais eficiente sabotador. A diferença está no design. Uma IA que ignora os precedentes de economia comportamental delegará ao S1 decisões que requerem S2, amplificará vieses em vez de corrigi-los e acumulará dívida cognitiva que, cedo ou tarde, o usuário ou a organização terão de pagar.

Uma IA desenhada como arquiteta de escolhas consciente, que calibra *defaults*, preserva agência, comunica incerteza, oferece fricção proporcional e audita seus próprios vieses, pode, ao contrário, ser o instrumento pelo qual o homem administrativo de Simon finalmente encontra o complemento de que sempre precisou: não um substituto para sua racionalidade limitada, mas um amplificador da racionalidade que ele genuinamente possui. Decidir é humano. Mas o ambiente em que se decide é, cada vez mais, uma obra de engenharia. O debate que este artigo propõe não é se a IA vai decidir por nós, mas: quem vai desenhar o ambiente em que nós e a IA decidiremos juntos?

Referências

BAROCAS, S.; SELBST, A.D. **Big data's disparate impact.** *California Law Review*, v.104, n.3, p.671-732, 2016.

DIAS, D.T.; AMÉRICO, J.C da S.; BERNARDINO, G.; BENINI, E.G. **Racionalidade limitada: uma análise dos manuais didáticos de Teoria Geral da Administração.** *Administração: Ensino e Pesquisa*, Rio de Janeiro, v.17, n.2, p.217-244, 2016. Disponível em: https://doi.org/10.13058/raep.2016.01712.419.

DIETVORST, B.J.; SIMMONS, J.P.; MASSEY, C. **Algorithm aversion: people erroneously avoid algorithms after seeing them err.** *Journal of Experimental Psychology: General*, v.144, n.1, p.114-126, 2015. Disponível em: https://doi.org/10.1037/xge0000033.

DIETVORST, B.J.; SIMMONS, J.P.; MASSEY, C. **Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them.** *Management Science*, v.64, n.3, p.1155-1170, 2018. Disponível em: https://doi.org/10.1287/mnsc.2016.2643.

HART, S.G.; STAVELAND, L.E. **Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research.** In: HANCOCK, P.A.; MESHKATI, N. (Ed.). *Human Mental Workload*. Amsterdam: North-Holland, 1988. p.139-183.

HAUSMAN, D.M.; WELCH, B. **Debate: To Nudge or Not to Nudge.** *The Journal of Political Philosophy*, [s.l.], v.18, n.1, p.123-136, 2010. Disponível em: https://doi.org/10.1111/j.1467-9760.2009.00351.x.

KAHNEMAN, D. **Thinking, fast and slow.** Nova York: Farrar, Straus and Giroux, 2011.

KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M. **Inherent Trade-Offs in the Fair Determination of Risk Scores.** *arXiv preprint arXiv:1609.05807*, 2016.

NOBLE, S.U. **Algorithms of oppression: how search engines reinforce racism**. Nova York: New York University Press, 2018.

O'NEIL, C. **Weapons of math destruction: how big data increases inequality and threatens democracy.** Nova York: Crown Publishers, 2016.

PAÇO CUNHA, E. **O itinerário epistemológico da racionalidade limitada de Herbert Simon.** *Organizações & Sociedade*, v.29, n.100, p.14-37, 2022. Disponível em: https://doi.org/10.1590/1984-92302022v29n0031PT.

SHUMAILOV, I. et al. **AI models collapse when trained on recursively generated data.** *Nature*, v.631, p.755-759, 2024. Disponível em: https://doi.org/10.1038/s41586-024-07566-y.

SIMON, H.A. **Designing organizations for an information-rich world.** In: GREENBERGER, M. (Ed.). *Computers, communications, and the public interest*. Baltimore: Johns Hopkins University Press, 1971. p.37-72.

SIMON, H.A. **Comportamento administrativo: estudo dos processos decisórios nas organizações administrativas**. 4.ed. Rio de Janeiro: Fundação Getulio Vargas, 1979.

STRAYER, D.L.; JOHNSTON, W.A. **Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone.** *Psychological Science*, [s.l.], v.12, n.6, p.462-466, 2001. Disponível em: https://doi.org/10.1111/1467-9280.00386.

THALER, R.H.; SUNSTEIN, C.R. **Nudge: improving decisions about health, wealth, and happiness**. New Haven: Yale University Press, 2008.