Quando a simulação refuta a si mesma: demarcação científica, IA generativa e responsabilidade epistêmica na pesquisa em comunicação digital

Saulo Lino dos Santos

Programa: Mestrado em Comunicação Digital
Instituição: Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa – IDP

Brasília – DF
2026

Resumo

O momento digital contemporâneo multiplicou artefatos computacionais cujos produtos se parecem com conhecimento: modelos de linguagem que redigem com fluência, simulações que geram padrões plausíveis, análises de grandes volumes de dados que exibem correlações convincentes. Este artigo argumenta que a pergunta clássica da demarcação científica (o que separa ciência de pseudociência) retorna com força renovada diante desses artefatos, e que sua resposta exige três disciplinas epistêmicas: a declaração prévia de critérios de refutação, a separação rigorosa entre o que foi programado e o que foi encontrado, e a validação contra o mundo, na qual um resultado negativo vale mais do que uma confirmação espúria. O argumento é desenvolvido a partir de um caso estendido: um modelo computacional de circulação e recepção de mensagens políticas, construído pelo autor, cuja primeira hipótese submetida a validação externa foi refutada, e cuja documentação, auditada, revelou o quanto é fácil atribuir a um sistema capacidades que seu código não implementa. Sustenta-se que a refutação, longe de invalidar o instrumento, é precisamente o que o distingue de um gerador de plausibilidades; e que a pesquisa em comunicação digital, campo que estuda os efeitos dos sistemas computacionais, tem razões particulares para não reproduzir em seus métodos a opacidade que critica em seu objeto.

Palavras-chave: demarcação científica; simulação computacional; IA generativa; opinião pública; responsabilidade epistêmica.

Abstract

The contemporary digital moment has multiplied computational artefacts whose outputs look like knowledge: language models that write fluently, simulations that generate plausible patterns, large-scale data analyses that display convincing correlations. This article argues that the classical problem of scientific demarcation (what separates science from pseudoscience) returns with renewed force in the face of these artefacts, and that answering it requires three epistemic disciplines: the prior declaration of refutation criteria, a rigorous separation between what was programmed and what was found, and validation against the world, in which a negative result is worth more than a spurious confirmation. The argument is developed through an extended case: a computational model of the circulation and reception of political messages, built by the author, whose first hypothesis submitted to external validation was refuted, and whose documentation, once audited, revealed how easy it is to attribute to a system capabilities its code does not implement. It is argued that refutation, far from invalidating the instrument, is precisely what distinguishes it from a plausibility generator; and that digital communication research, a field that studies the effects of computational systems, has particular reason not to reproduce in its methods the opacity it criticises in its object.

Keywords: scientific demarcation; computational simulation; generative AI; public opinion; epistemic responsibility.

Sumário

1 Introdução: o output que se parece com conhecimento

Há uma cena que se repete, com variações, em qualquer grupo de pesquisa que trabalhe com métodos computacionais: o sistema devolve um número ou um parágrafo bem escrito, e alguém pergunta “então está provado?”. A cena condensa o problema epistemológico central

do momento digital: a distância entre *produzir um output* e *produzir conhecimento* nunca foi tão fácil de atravessar sem perceber. Um modelo de linguagem completa frases com autoridade gramatical que não corresponde a nenhuma autoridade epistêmica (BENDER et al., 2021). Uma simulação multiagente gera curvas que se parecem com dinâmicas sociais. Uma análise de milhões de postagens exibe correlações que se parecem com causas (BOYD; CRAWFORD, 2012). Em todos os casos, o artefato computacional tem uma propriedade nova e perigosa: ele *sempre responde*. A pergunta que este artigo enfrenta é o que autoriza tratar alguma dessas respostas como conhecimento, e o que obriga a descartar as demais.

A pergunta não é nova. Ela é a pergunta da demarcação, posta por Popper (1972) contra as teorias que explicavam tudo e por isso não explicavam nada: o que distingue uma afirmação científica não é sua capacidade de confirmação, mas sua exposição deliberada à refutação. O que o momento digital acrescenta é uma torção: os sistemas computacionais contemporâneos são máquinas de confirmação. Um modelo de linguagem treinado para prever a próxima palavra produzirá, por construção, o texto mais plausível, e plausibilidade é exatamente o que a demarcação popperiana ensina a desconfiar. Uma simulação parametrizada por seu autor produzirá, com frequência desconcertante, aquilo que seu autor esperava, e a tentação de chamar isso de "achado" é o análogo computacional do viés de confirmação que Ioannidis (2005) identificou como motor estrutural da produção de resultados falsos na literatura científica.

Este artigo desenvolve o argumento por dentro de um caso: um modelo computacional de circulação e recepção de mensagens políticas em ambiente plataformizado, construído pelo autor como instrumento de sua pesquisa de mestrado sobre opinião pública em contexto eleitoral. O caso interessa não pelo que o modelo acertou, mas pelo que ele errou, e pelo que a auditoria de sua própria documentação revelou sobre a facilidade de sobre-afirmar. A primeira hipótese do modelo submetida a validação externa foi refutada: a correlação entre o sentimento simulado e a confiança institucional observada por faixa de renda foi de −0,65, direção oposta à esperada, e o índice agregado de confiança em validação externa do instrumento está em 0,17 numa escala de 0 a 1. Defendo que esses números, publicados com a mesma ênfase que se daria a uma confirmação, são o que separa o instrumento de um gerador de plausibilidades, e que a disciplina necessária para publicá-los é a mesma que o campo da comunicação digital precisa exigir de qualquer sistema computacional que estude, use ou critique.

O percurso é o seguinte. A seção 2 revisita o problema da demarcação no contexto dos artefatos computacionais. A seção 3 discute a novidade epistêmica da simulação como método

que não é nem teoria nem experimento (HUMPHREYS, 2009). A seção 4 examina o lugar da IA generativa nesse quadro, a partir da crítica dos "papagaios estocásticos" e de uma decisão de arquitetura do caso estudado. A seção 5 nomeia uma armadilha metodológica recorrente no trabalho com simulações (confundir premissa com descoberta), com dois episódios documentados do próprio projeto. A seção 6 analisa o resultado negativo como virtude epistêmica. A seção 7 conclui com implicações para a responsabilidade epistêmica na pesquisa em comunicação.

2 A demarcação revisitada: quando o software responde por tudo

O critério popperiano é conhecido: uma teoria é científica na medida em que proíbe algo, ou seja, em que especifica, antes do teste, quais observações a derrubariam (POPPER, 1972). Chalmers (1993) mostrou as dificuldades do critério em estado puro, e Kuhn (2011) deslocou a questão para a prática das comunidades científicas, nas quais a ciência normal não testa o paradigma a cada experimento. Feyerabend (2007) foi mais longe e negou que exista *um método* cuja violação defina a não-ciência. Essas objeções são sérias, e este artigo não pretende resolvê-las. O que pretende é mais modesto e mais urgente: mostrar que, qualquer que seja a posição do leitor no debate clássico, os artefatos computacionais do momento digital criam um problema de demarcação *prático* que nenhuma das posições clássicas autoriza a ignorar.

O problema é este: um sistema computacional sempre produz saída. O experimento clássico pode falhar: o reagente não reage, o instrumento não detecta. O modelo computacional, não: dada uma entrada, ele devolve números. Essa disponibilidade permanente da resposta desloca o ônus da demarcação do sistema para o pesquisador. Não se trata mais de perguntar se a teoria proíbe algo, mas de perguntar se *o uso do sistema* proíbe algo: o pesquisador declarou, antes de rodar, o que o faria descartar o resultado? Ou rodou, olhou, e decidiu depois o que contava como sucesso?

No caso estudado, essa disciplina foi imposta por arquitetura: o modelo recusa-se a executar qualquer simulação sem uma hipótese formal cujo critério de refutação seja campo obrigatório, declarado e persistido antes da execução. A escolha é diretamente popperiana, e vale registrar o que ela custou. Foi esse critério pré-declarado que obrigou a classificar como *refutado* o primeiro caso submetido a validação externa, quando teria sido retoricamente confortável reclassificá-lo como "parcialmente corroborado". A demarcação, no trabalho computacional, é menos uma posição filosófica que se assina do que um constrangimento que se programa contra si mesmo.

3 A novidade epistêmica da simulação

Humphreys (2009) argumentou que os métodos de simulação computacional constituem uma novidade filosófica genuína: não se deixam reduzir nem à teoria (porque produzem resultados que o teórico não consegue derivar analiticamente) nem ao experimento (porque não observam o mundo, mas um substituto construído dele). A simulação habita um terceiro lugar, e desse lugar vêm tanto sua utilidade quanto seu risco característico.

A utilidade, para a pesquisa em comunicação política, é conhecida de quem enfrenta os limites da pesquisa por amostragem: o survey mede o que o eleitorado declara num instante, mas não responde perguntas contrafactuais: o que teria acontecido se a campanha tivesse enquadrado o tema de outro modo, ou concentrado esforço em outra plataforma. A modelagem baseada em agentes oferece uma via para esse tipo de pergunta: em vez de estimar parâmetros de um agregado, simula-se uma população heterogênea interagindo sob regras teoricamente fundamentadas, e observa-se o padrão coletivo que emerge (EPSTEIN; AXTELL, 1996). *No interior do modelo (e a expressão importa), dois cenários* que diferem em um único parâmetro, executados sobre a mesma população com a mesma semente aleatória, permitem atribuir a diferença observada àquele parâmetro. É uma forma controlada de raciocínio contrafactual que a pesquisa observacional não oferece.

O risco vem do mesmo traço que produz a utilidade. Porque a simulação não observa o mundo, sua estabilidade interna não diz nada sobre sua aderência ao mundo, distinção entre verificação e validação que a literatura de modelagem baseada em agentes formalizou (GRIMM et al., 2006). Um modelo pode ser reprodutível, convergente, elegante, e sistematicamente errado. O caso estudado fornece o exemplo em números: suas execuções são determinísticas sob a mesma semente, seus 145 testes automatizados passam, sua variância entre sementes decai como $N^{-0,49}$, comportamento de convergência estatística de manual. Ainda assim, quando o sentimento que ele simula foi confrontado com a confiança institucional medida por survey, a correlação veio negativa. Nenhuma dessas propriedades internas dizia algo sobre a aderência do modelo aos dados observados. É por isso que a simulação precisa carregar seus critérios de refutação declarados: ela é o método em que é mais fácil confundir o funcionamento do instrumento com a verdade do resultado.

4 Papagaios estocásticos e o lugar da IA generativa no laço epistêmico

Se a simulação cria um problema de demarcação prático, a IA generativa o agudiza. A crítica de Bender et al. (2021) aos grandes modelos de linguagem ("papagaios estocásticos" que costuram sequências prováveis de texto sem qualquer modelo de mundo subjacente) e a advertência de Shanahan (2024) sobre a gramática enganosa com que falamos desses sistemas ("o modelo acha", "o modelo sabe") convergem para o mesmo ponto epistemológico: a fluência do output não é evidência de compreensão, e a plausibilidade do texto não é evidência de verdade. Para a produção científica, as implicações já são visíveis, de artigos fabricados a revisões de literatura com referências inexistentes, e periódicos científicos passaram a exigir transparência sobre o que a máquina fez e validação humana do que ela produziu (VAN DIS et al., 2023).

O caso estudado oferece um exercício concreto de tradução desse princípio em arquitetura. O modelo de simulação usa um modelo de linguagem, mas em posição deliberadamente subordinada: o sentimento de cada interação simulada é calculado deterministicamente pelo motor, a partir das regras teóricas do modelo, e o LLM apenas *veste* esse sentimento com texto natural, quando disponível; se falha, um template assume, e nada muda no resultado. A decisão pode parecer detalhe de engenharia, mas é uma tomada de posição epistemológica: o componente estatístico de imitação de linguagem fica *fora do laço epistêmico*. Nenhum número que o modelo produz, nenhuma métrica que será validada, nenhuma conclusão que será defendida depende do que o papagaio estocástico disse. A alternativa (deixar o LLM decidir sentimentos, comportamentos ou desfechos) teria contaminado o instrumento com exatamente a opacidade que Bender et al. diagnosticam: um gerador de plausibilidades no coração de um sistema que pretende ser auditável.

Generalizo a lição: no uso científico de IA generativa, a pergunta decisiva não é "usar ou não usar", mas *onde, no fluxo que vai dos dados à conclusão, o componente generativo está posicionado***, e se cada afirmação** que chega à conclusão pode ser rastreada até uma fonte que não seja ele. É a mesma pergunta que boyd e Crawford (2012) ensinaram a fazer sobre o big data: não "quantos dados", mas "que afirmações esses dados autorizam, e para quem".

5 A armadilha da emergência: premissa não é descoberta

Há um erro metodológico que o trabalho com simulações torna quase irresistível, e que merece nome próprio: a *armadilha da emergência***, que consiste em apresentar como achado**

do modelo aquilo que foi colocado nele como premissa. Dois episódios documentados do projeto ilustram o mecanismo, e os relato contra mim mesmo, porque é assim que a lição fica honesta.

O primeiro. A população sintética do modelo atribui a cada agente uma "competência política" calculada como composição de escolaridade e renda, operacionalização inspirada na distribuição desigual de capital cultural e econômico descrita por Bourdieu (2007), com a hipótese adicional, do modelo e não do autor citado, de que menor competência implica maior suscetibilidade ao enquadramento. Quando a simulação roda, observa-se um gradiente: agentes de menor renda e escolaridade absorvem mais o clima de opinião dominante, numa escala que vai de 0,33 a 0,67 entre as faixas. É tentador apresentar esse gradiente como resultado. Ele não é. Ele é a fórmula devolvida: a suscetibilidade foi *definida* como o complemento da competência, e a competência foi definida como função de renda e escolaridade. O que a simulação faz com essa premissa (como ela interage com plataformas, clima e memória ao longo das rodadas) pode conter resultados; o gradiente em si é eco. A regra geral: se um padrão foi posto numa fórmula, num peso ou numa regra, sua reaparição na saída é comportamento esperado da função, não descoberta. Ioannidis (2005) descreveu o quanto a literatura publicada é inflada por achados que não sobreviveriam a essa pergunta simples: *o que, no desenho, já garantia esse resultado?*

O segundo episódio é mais desconfortável. A documentação pública do projeto afirmou, durante semanas, que o modelo implementava "a espiral do silêncio em ação", a tese de Noelle-Neumann (2017) de que indivíduos que percebem sua opinião como minoritária tendem a calar. Uma auditoria linha a linha do código mostrou que a afirmação era falsa. O que o código implementa é conformidade de expressão: o clima de opinião modula *o que os agentes expressam*. A decisão de calar, porém, nunca compara a opinião do agente ao clima: o silêncio, no modelo, é mera não-interação por baixo engajamento, e sequer é registrado como evento observável. O mecanismo central da teoria citada não existe no sistema; existia apenas no texto sobre o sistema. O episódio não é anedota de bastidor: é o problema de demarcação do momento digital em miniatura. Entre o conceito teórico e o mecanismo computacional há um espaço onde a sobre-afirmação prospera, porque o texto sobre o software é mais fácil de escrever do que o software, e porque ninguém audita a documentação com o rigor com que se audita um experimento. A correção exigiu reescrever a documentação inteira sob uma regra única: nenhum conceito da teoria da comunicação pode ser atribuído ao sistema sem apontar a linha de código que o implementa, e todo conceito sem linha correspondente é declarado como proxy ou lacuna.

6 Quando o modelo erra: o valor epistêmico do resultado negativo

O primeiro caso do modelo submetido a validação externa testava uma hipótese de enquadramento: a de que a cobertura negativa de uma instituição financeira pública reduziria a confiança institucional simulada, com efeito mais intenso nos estratos de menor renda. A validação confrontou o sentimento simulado com dados observados de confiança institucional por faixa de renda. O resultado: correlação de −0,65, não apenas fraca, mas de direção oposta. O caso foi classificado como refutado, o índice agregado de confiança em validação externa do instrumento foi fixado em 0,17, e ambos os números constam da primeira página da documentação pública do projeto.

O que um resultado assim ensina depende inteiramente da moldura em que é lido. Na moldura da confirmação, que Ioannidis (2005) mostrou dominar os incentivos da publicação científica, ele é um fracasso a esconder ou maquiar. Na moldura da demarcação, ele é o instrumento funcionando: o modelo fez uma afirmação arriscada, o mundo respondeu, e a resposta foi registrada contra o modelo. Mais: o conteúdo específico da refutação é um achado metodológico com relevância para além do projeto. O que os dados recusaram foi o uso do *sentimento simulado em plataformas* como proxy da *confiança institucional medida por survey, dois construtos* que a pesquisa em comunicação digital rotineiramente trata como intercambiáveis quando converte métricas de redes sociais em afirmações sobre opinião pública. O resultado não autoriza a conclusão forte de que os construtos "são diferentes"; autoriza a conclusão precisa de que, neste caso e nestas condições, um não serviu de proxy convergente do outro. Para um campo que produz diariamente dashboards de "sentimento" oferecidos como leitura da opinião pública, essa delimitação negativa é um resultado útil.

Registro também o que a refutação *não* fez: não encerrou o instrumento. Ela delimitou seu escopo inferencial pretendido (ordenações entre segmentos e comparações entre cenários, no interior do modelo, nunca magnitudes nem projeções de intenção de voto) e converteu a validação externa, antes tratada como etapa final, em programa de pesquisa permanente. A distinção prática é entre descartar o instrumento e restringir o que se pode afirmar com ele.

7 Considerações finais: responsabilidade epistêmica como prática

O argumento deste artigo pode ser condensado em três disciplinas, extraídas de um caso e generalizáveis a qualquer pesquisa em comunicação que use artefatos computacionais.

Declarar a refutação antes do resultado. No trabalho computacional, a demarcação é uma restrição de projeto antes de ser uma filiação teórica. Sistemas sempre respondem; o que separa o uso científico do uso retórico é a existência de um critério, anterior à execução, capaz de classificar a resposta como derrota.

Subtrair do resultado aquilo que o desenho já garantia. Toda saída de simulação, como toda saída de IA generativa, deve responder à pergunta: o que, no desenho, já garantia isso? O que sobra depois dessa subtração é o candidato a resultado. O que não sobra é premissa refletida, e apresentá-la como achado é a forma computacional do raciocínio circular.

Por fim, o erro merece a mesma ênfase que o acerto. O resultado negativo é o produto epistêmico mais confiável que um sistema computacional pode gerar, precisamente porque é o único que o gerador de plausibilidades não tem incentivo em produzir.

Há uma simetria final que o campo da comunicação digital não deveria ignorar. A área construiu, com razão, uma agenda crítica sobre a opacidade dos sistemas que estuda: os algoritmos que não se deixam auditar, as plataformas que não explicam suas mediações. Essa crítica perde autoridade se os métodos da própria área reproduzirem o que ela denuncia: simulações cujas premissas voltam como descobertas, métricas de sentimento convertidas em opinião pública sem validação, textos gerados assumidos como conhecimento. A responsabilidade epistêmica é constitutiva do método, não uma ética acrescentada a ele. De um campo que estuda máquinas de plausibilidade espera-se, no mínimo, que a plausibilidade não baste como critério nos seus próprios métodos.

Referências

BENDER, E. M.; GEBRU, T.; McMILLAN-MAJOR, A.; SHMITCHELL, S. *On the Dangers of Stochastic Parrots*: Can Language Models Be Too Big? In: FAccT '21. Nova York: ACM, 2021.

BOURDIEU, P. *A distinção*: crítica social do julgamento. São Paulo: Edusp; Porto Alegre: Zouk, 2007.

BOYD, d.; CRAWFORD, K. Critical Questions for Big Data. *Information, Communication & Society*, v. 15, n. 5, p. 662-679, 2012.

CHALMERS, A. *O que é ciência afinal?* São Paulo: Brasiliense, 1993.

EPSTEIN, J. M.; AXTELL, R. *Growing Artificial Societies*: Social Science from the Bottom Up. Cambridge: MIT Press, 1996.

FEYERABEND, P. *Contra o método*. São Paulo: Editora Unesp, 2007.

GRIMM, V. et al. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, v. 198, n. 1-2, p. 115-126, 2006.

HUMPHREYS, P. The philosophical novelty of computer simulation methods. *Synthese*, v. 169, n. 3, p. 615-626, 2009.

IOANNIDIS, J. P. A. Why Most Published Research Findings Are False. *PLoS Medicine*, v. 2, n. 8, e124, 2005.

KUHN, T. S. *A estrutura das revoluções científicas*. São Paulo: Perspectiva, 2011.

NOELLE-NEUMANN, E. *A espiral do silêncio*: opinião pública, nosso tecido social. Florianópolis: Estudos Nacionais, 2017.

POPPER, K. R. *A lógica da pesquisa científica*. São Paulo: Cultrix, 1972.

SHANAHAN, M. Talking About Large Language Models. *Communications of the ACM*, v. 67, n. 2, p. 68-79, 2024.

VAN DIS, E. A. M. et al. ChatGPT: five priorities for research. *Nature*, v. 614, p. 224-226, 2023.