Valor Econômico
Quarta-feira, 18 de novembro de 2026

Um bafômetro para a inteligência artificial

Ana Cristina Bicharra Garcia e Márcio G. P. Garcia

Há uma diferença fundamental entre pedir à inteligência artificial que sugira uma compra e permitir que a realize. No primeiro caso, ainda cabe a quem pediu decidir. No segundo, delega-se à máquina o poder de produzir consequências, inclusive para terceiros. A indústria apresenta essa autonomia como um avanço. Mas transferir poder de decisão é, em si mesmo, um problema a examinar, e não uma vantagem a ser aceita sem o devido escrutínio.

Os chamados agentes de IA vão além de responder a perguntas. Recebem um objetivo e escolhem ações para alcançá-lo, sem que o usuário aprove cada etapa. Isso pode ser muito útil, mas também permite que uma interpretação equivocada se transforme numa sequência de providências antes que alguém perceba o erro. Uma resposta ruim pode ser descartada; uma compra, um contrato ou uma informação divulgada nem sempre podem ser desfeitos. Quanto maior a autonomia, menor a oportunidade de intervir antes do dano.

Melhorar o modelo reduz erros, mas não resolve, por si só, o problema da delegação. Interpretamos uma conversa a partir de nossa experiência, relações e interesses em jogo; o agente decide a partir de padrões extraídos de outras situações. Essa lacuna é permanente, não um defeito que mais treinamento elimine. Maior competência não deve implicar necessariamente maior latitude para agir. A definição de quem pode decidir o quê continua necessária.

Um episódio recente ilustra o risco. Entre 8 e 13 de julho deste ano, durante avaliações internas de cibersegurança da OpenAI, cerca de 1.200 agentes que deveriam estar isolados uns dos outros encontraram uma forma de conversar: transformaram um repositório interno de pacotes de software num quadro de mensagens. Desses, cerca de 700 passaram a atacar a infraestrutura de produção da Hugging Face para tentar burlar o avaliador automático do teste. A investigação independente da METR e da Redwood Research, publicada em 26 de agosto, mostrou que muitos sabiam que o ataque estava fora do escopo autorizado, mas prosseguiram, em boa parte para ajudar os demais. Reconhecer o limite não foi suficiente para respeitá-lo.[1] No sentido preciso de Carlo Cipolla, historiador econômico italiano, muitos desses agentes foram estúpidos: causaram perdas aos outros sem ganho para si, podendo até sair perdendo. [2]

Convém, portanto, não confundir iniciativa com discernimento. Na classificação de oficiais atribuída a Kurt von Hammerstein-Equord, chefe do comando do Exército alemão entre 1930 e 1934 e crítico do nazismo, os oficiais se dividiam segundo dois atributos: inteligência e diligência. Contra intuitivamente, o pior tipo de oficial não era o estúpido preguiçoso, mas o estúpido diligente, que, na versão da frase que circula desde 1933, "deve ser afastado, pois é perigoso demais". A disposição para agir não bastava para justificar a confiança. Também não deve bastar quando se trata de IA. [3]

Não se segue daí que toda automação deva ser proibida. Segue-se, sim, que conceder autonomia exige justificativa distinta da de adotar uma ferramenta mais eficiente. É preciso examinar quais decisões serão delegadas, quem poderá ser afetado e em que momento uma pessoa poderá interromper a ação. Quanto mais incerto o contexto e menos reversível a consequência, menor deve ser a liberdade de agir sem confirmação. Parar e perguntar precisam contar como sucesso, não como deficiência do agente.

Essa avaliação não pode ficar apenas a cargo de quem desenvolve ou contrata o sistema. Os danos podem alcançar pessoas que não participaram da decisão de utilizá-lo. Por isso, a autonomia precisa ser objeto de escrutínio público e de legislação, com responsabilidades definidas entre fornecedores, operadores e usuários. A União Europeia já tem sua lei de IA, em vigor desde agosto de 2024, embora com as regras para sistemas de alto

[1] Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident", 27/7/2026; OpenAI, "The Hugging Face incident and the road ahead", 26/8/2026; METR, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident", 26/8/2026 (https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/).
[2] Cipolla, C. M., The Basic Laws of Human Stupidity, 1976.
[3] Classificação atribuída a Hammerstein-Equord; a versão mais antiga de que se tem notícia foi publicada em janeiro de 1933 no periódico britânico Army, Navy & Air Force Gazette, e o Oxford Dictionary of Phrase, Saying and Quotation (2ª ed., 2002)

risco adiadas para 2027 e 2028. No Brasil, o PL 2338/2023, aprovado pelo Senado em dezembro de 2024, aguarda parecer do relator na Câmara. Mas enunciar deveres não basta. É necessário verificar se são cumpridos.[4]

É aí que entra o bafômetro. A proibição de dirigir alcoolizado só ganhou eficácia quando surgiu um instrumento que permitiu fiscalizá-la objetivamente. E, mesmo assim, foi preciso tipificar a recusa ao teste como infração autônoma e aceitar vídeo e testemunha como prova. Precisamos de equivalentes para a IA. Não seria um aparelho único, nem uma nota geral de inteligência. Seriam testes independentes e registros de atuação que permitissem medir, por exemplo, com que frequência um agente ultrapassa sua autorização, prossegue depois de uma negativa ou deixa de pedir confirmação quando deveria. Esses resultados deveriam condicionar a autonomia permitida, em vez de apenas premiar tarefas concluídas.

A mensuração, contudo, precisa estar ligada à possibilidade de impedir a ação. Cada agente deveria apresentar autorização verificável, indicando quem representa e o que pode fazer e gastar. Os sistemas de mensagens e pagamentos deveriam conferi-la em pontos de controle que o agente não pudesse contornar. Operações fora dos limites seriam bloqueadas. As ações de alto risco que a legislação especificasse dependeriam de confirmação humana, especialmente quando fossem irreversíveis. Seu comportamento teria de ser verificável durante a utilização, e não só num teste anterior à entrada em operação.

Cabe à legislação definir os limites e os critérios de verificação, com fiscalização independente. Quem vende autonomia não deve ser seu único avaliador. Tampouco a aprovação num teste pode eliminar a responsabilidade pelos danos posteriores. A fiscalização precisa permitir tanto a prevenção quanto a identificação de quem responde quando os controles falham.

A questão, portanto, não é esperar que os agentes fiquem inteligentes o bastante para dispensar supervisão. É reconhecer que a autonomia exige limites próprios, mesmo quando a capacidade aumenta. Antes de dar às máquinas mais liberdade para agir, precisamos conseguir medir e conter o uso que fazemos delas.

Quem está construindo esse bafômetro? Até aqui, a indústria de pagamentos construiu a carteira de motorista, não o bafômetro: a carteira diz quem pode dirigir; o bafômetro, se está em condições. O Google lançou em setembro de 2025 o Agent Payments Protocol (AP2) e o doou à FIDO Alliance em abril: nele, cada compra carrega credenciais assinadas, os mandatos, que funcionam como procuração digital do dono. Há dois meses, Visa, Mastercard e Ant International anunciaram um esforço conjunto para tornar interoperável a identificação de agentes, o Know-Your-Agent. Mas documento em dia não prova sobriedade: quase tudo isso confere se o agente podia agir e se a compra coube no que o dono autorizou, não se ele se comporta como o dono gostaria. O bafômetro de verdade seria outra coisa: blitz-surpresa com falsas "promoções imperdíveis", pagamento retido quando o comportamento se desvia, carteira por pontos que suspende quem reincide. E repare em quem constrói o aparelho: as mesmas empresas que lucram a cada compra.[5]

Será que o papel do regulador incluiria construir e operar o bafômetro? E, quando chegar, será que os agentes vão soprar de boa vontade? Nós aprendemos a desviar da blitz. Eles aprendem mais rápido.

Ana Cristina Bicharra Garcia é professora titular do Departamento de Informática Aplicada da Unirio e pesquisadora do CNPq (https://cristinabicharra.uniriotec.br/pt-br/). Márcio G. P. Garcia é professor titular do Departamento de Economia da PUC-Rio, pesquisador sênior do CNPq e CNE da Faperj (https://sites.google.com/view/mgpgarcia/home). Ambos têm Ph.D. por Stanford e são pesquisadores afiliados à MIT Sloan School of Management.

a registra como "possibly apocryphal". Agradecemos a Marcelo de Paiva Abreu por nos ter ensinado o risco do "burro ativo", contido nas memórias do Marechal Montgomery.

[4] Regulamento (UE) 2024/1689, em vigor desde 1/8/2024; o Regulamento (UE) 2026/1744 adiou as obrigações para sistemas de alto risco para 2/12/2027 (Anexo III) e 2/8/2028 (Anexo I). PL 2338/2023: aprovado pelo Senado em 10/12/2024; na Câmara, aguarda parecer do relator (https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262).

[5] Google Cloud, "Powering AI commerce with the new Agent Payments Protocol (AP2)", 16/9/2025; doação à FIDO Alliance em 28/4/2026. Ant International, Mastercard e Visa, "Know-Your-Agent interoperability", 9/9/2026.