ARTIGO

Erisictão Já Invadiu o Bosque de Deméter: análise dos riscos sociais do colapso de IAs por retroalimentação

Pedro Khauaja
pedrokhauaja@gmail.com

Advogado; Mestre e Doutorando pelo Programa de Pós-Graduação em Sociologia e Direito da Universidade Federal Fluminense (PPGSD-UFF); editor-assistente da revista CONFLUÊNCIAS; professor substituto no Departamento de Direito Privado da Universidade Federal Fluminense (UFF).

V. 5 / N. 1 / JULHO DE 2024
PÁGINAS 94 A 107

Erisictão Já Invadiu o Bosque de Deméter: análise dos riscos sociais do colapso de IAs por retroalimentação

Palavras-chave
Retroalimentação de IAs
Colapso de sistemas
Risco
Modernidade tardia
Inteligência Artificial

Resumo
Este artigo analisa os riscos sociais que surgem da possibilidade de colapso de sistemas de Inteligência Artificial (IA) por retroalimentação. A retroalimentação ocorre quando IAs generativas são treinadas com dados gerados por outras IAs, o que pode levar ao colapso desses sistemas, conforme demonstrado por Shumailov et al. (2024) e Martinez et al. (2024). A pesquisa toma como ponto de partida o conceito de "risco reflexivo" de Beck (2011), aplicando-o ao contexto tecnológico contemporâneo. A metodologia inclui uma revisão teórica e bibliográfica sobre os riscos associados ao uso crescente de IAs, e uma análise dos dados quantitativos fornecidos pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), que evidenciam o aumento da presença de IA no Brasil, assim como outros trabalhos que apontam o mesmo em outros países, muitas vezes sem transparência sobre a origem dos conteúdos gerados. Os resultados indicam que a retroalimentação de IA não só representa um risco técnico para o colapso dos sistemas, mas também um risco social, devido à forma desigual com que os impactos negativos são distribuídos. A análise conclui que o fenômeno de colapso das IAs reflete um padrão característico da modernidade tardia, onde os benefícios tecnológicos e econômicos são concentrados, enquanto os danos potenciais são externalizados, especialmente em contextos de maior vulnerabilidade.

V. 5 / N. 1 / JULHO DE 2024
PÁGINAS 94 A 107

Erysichthon Has Already Invaded Demeter's Grove: an analysis of the social risks of AI collapse due to feedback

Keywords
AI feedback loop
System collapse;
Risk
Late modernity
Artificial Intelligence

Abstract
This paper analyzes the social risks arising from the potential collapse of Artificial Intelligence (AI) systems due to feedback loops. A feedback loop occurs when generative AIs are trained with data generated by other AIs, which can lead to the collapse of these systems, as demonstrated by Shumailov et al. (2024) and Martinez et al. (2024). The research is based on Beck's (2011) concept of "reflexive risk," applying it to the contemporary technological context. The methodology includes a theoretical and bibliographic review of the risks associated with the growing use of AIs and an analysis of quantitative data provided by the Regional Center for Studies on the Development of the Information Society (Cetic.br), which highlights the increasing presence of AI in Brazil, as well as other studies pointing to the same trend in other countries, often without transparency regarding the origin of generated content. The results indicate that AI feedback not only represents a technical risk for system collapse but also a social risk, due to the unequal distribution of negative impacts. The analysis concludes that the phenomenon of AI collapse reflects a pattern characteristic of late modernity, where technological and economic benefits are concentrated, while potential harms are externalized, particularly in more vulnerable contexts.

1 Introdução

O mito de Erisictão é contado por Ovídio em seu livro "Metamorfoses" como parte do grande acervo de histórias mitológicas gregas. Essa é mais uma que trata do desrespeito aos deuses, da ira divina, e do alto preço da soberba e falta de controle sobre as próprias pulsões. Erisictão era um rei da Tessália conhecido por sua extrema avareza e pelo seu desrespeito às divindades. Em um ato de grande insolência, decidiu cortar uma árvore sagrada de Deméter em um bosque dedicado à deusa da agricultura.

A árvore era considerada sagrada e fazia parte do domínio de Deméter, responsável pela fertilidade da terra e pelo ciclo das colheitas. Em resposta ao ato sacrílego, a deusa decidiu punir Erisictão com uma maldição de fome insaciável. Essa fome fez com que o rei gastasse toda a sua fortuna e posses, e eventualmente começasse a devorar sua própria carne, tamanho nível de seu desespero. A história termina com Erisictão sendo consumido pela própria fome, um destino cruel e simbólico que, naquele contexto, reforçava a mensagem sobre a importância de respeitar os deuses e o seu domínio (Lobato Martins, 2023).

Essa estória, porém, carrega um outro sentido, para além da mensagem e cuidado com os deuses. É mais uma das várias tentativas de, através do mito, avisar dos perigos da gula e da avareza, alertando para o eventual destino de todo descontrole consumista: consumir a si mesmo. Nesse artigo, a estória de Erisictão serve como uma metáfora para o tema central, que é a análise dos riscos sociais implicados nas evidências recentes de que os sistemas de Inteligência Artificial (IA), especialmente as IAs generativas, correm alto risco de colapso quando alimentadas com dados criados por outras IAs.

Essa prática de utilizar dados de uma IA para alimentar o treinamento de outra é o que chamo de retroalimentação. Nesse texto utilizo o termo retroalimentação como uma tradução possível para "*feedback loop*", termo que foi utilizado nos primeiros trabalhos sobre o tema, em inglês., com especial atenção para Shumailov et al. (2024) e Martínez et al. (2023). No geral, um "*feedback loop*" é exatamente esse processo em que uma determinada IA recebe dados de outra como se fossem produzidos por humanos. Essa aparência de humanidade se mostrará um elemento essencial para o risco de colapso.

O trabalho de Shumailov et al. (2024) foi o primeiro a apontar essa possibilidade e tem tido muita repercussão, mas outros já vieram confirmar o mesmo efeito em contextos similares, como Martinez et al. (2024). Tomando como pressuposto essa descoberta, qual seja, de que as IAs tendem ao colapso quando retroalimentadas, abordarei o risco social que origina desse fato, quando somado ao amplo espectro de usos das IAs generativas hoje. Farei essa abordagem de análise de risco a partir do arcabouço teórico de Beck (2011) e sua caracterização da modernidade como período de risco reflexivo, distribuído desigualmente e ocultado deliberadamente.

Assim, primeiro trabalharei algumas conceituações iniciais sobre o colapso de IAs por retroalimentação, qualificando o debate ainda novo nas produções brasileiras. Logo em seguida, também trarei dados que corroboram a ideia de que há um aumento perigoso de conteúdo gerado por IA na Internet, e especialmente um conteúdo sem nenhum tipo de aviso ou indicativo de sua origem. Depois, com uso principalmente dos relatórios de pesquisa produzidos pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br)[1], vou trabalhar com a hipótese deum aumento da presença de tecnologias de IA no ecossistema tecnológico brasileiro, e a natureza delicada e sensível de alguns de seus usos.

Esse cenário servirá de base para a análise feita ao final do artigo que, partindo da conceituação teórica de Beck (2011), proporá que o risco de colapso das IAs é característico da

modernidade tardia, funcionando como uma forma desigual de centralizar lucros e compartilhar danos.

2 O Colapso de IAs por retroalimentação

Essa pesquisa se baseia largamente no trabalho de Schumailov et al. (2024) e, portanto, trabalharei com uma conceituação técnica que esteja de acordo com a dos autores. Ainda que no artigo não estejam definidos de forma expressa todos os termos, os autores estão trabalhando com um tipo específico de tecnologia: os *Large Lenguage Models* (LLMs), que são um desdobramento das tecnologias chamadas de IAs generativas.

As Inteligências Artificiais generativas são um subcampo da inteligência artificial focado na criação de modelos capazes de gerar novos dados a partir de um conjunto de dados existente (Ludermir, 2021). Esses sistemas não apenas analisam ou processam informações, mas são projetados para criar conteúdo supostamente "novo", como texto, imagens, áudio, e até mesmo vídeo, que são indistinguíveis ou dificilmente distinguíveis dos produzidos por seres humanos, como mostra a *survey* de Wu et al. (2023). A geração de conteúdo por essas IAs é baseada em aprendizado de padrões e estruturas presentes nos dados de treinamento, permitindo que elas produzam resultados criativos ou inovadores, sem intervenção humana direta.

Diferentes tipos de IAs generativas têm aplicações vastas e variadas, e não necessariamente funcionam da mesma forma que os LLMs (Ludermir, 2021). Os Modelos de Linguagem de Grande Escala, conhecidos como LLMs, são sistemas de inteligência artificial projetados para compreender, gerar, emular e manipular linguagem natural em grande escala (Naveed et al., 2024). Esses modelos são baseados em arquiteturas avançadas de redes neurais, como os transformadores, e são treinados em grandes volumes de dados textuais para captar e replicar padrões linguísticos complexos.

O treinamento de LLMs envolve a exposição a vastas quantidades de texto, proveniente de diversas fontes, como livros, artigos, sites e outros documentos. O objetivo é permitir que o modelo aprenda a prever a próxima palavra mais provável em uma sequência de texto, identificar relações contextuais entre palavras e entender o significado geral do texto. Esse processo de treinamento permite que os LLMs aparentem ter uma compreensão profunda da estrutura e do uso da linguagem, enquanto, ao mesmo tempo, emulam um padrão-base humano de respostas (Naveed et al., 2024).

Os LLMs mais avançados, como GPT-3 e GPT-4, são exemplos de modelos que possuem bilhões ou até trilhões de parâmetros, o que lhes confere uma capacidade significativa para capturar nuances e complexidades da linguagem (Colombo e Goulart, 2023). A escala desses modelos permite um desempenho impressionante em várias tarefas linguísticas, mas também levanta questões sobre a ética, a privacidade e o potencial para vieses nos dados de treinamento, assim como debates sobre a legitimidade da sua atividade, com acusações de plágio e outras similares (Wu et al., 2023).

O colapso de sistemas de Inteligência Artificial por retroalimentação se refere a um fenômeno em que sistemas de inteligência artificial, especialmente aqueles baseados em aprendizado de máquina, sofrem um progressivo declínio em sua qualidade de conteúdo gerado por conta do uso repetitivo de seus próprios *outputs*, ou de outras IAs, como *inputs* para futuros processos de treinamento. Esse fenômeno pode resultar em uma degeneração da qualidade das previsões ou decisões produzidas pelo sistema, levando a um ciclo de deterioração autoalimentado, conforme definido por Schumailov et al. (2024).

Os autores apontam que esse tipo de colapso ocorre principalmente em ambientes onde a IA é usada para gerar ou filtrar dados que, posteriormente, são reutilizados no processo de treinamento ou operação contínua do mesmo sistema ou de outros sistemas interligados. Quando esse processo não é cuidadosamente gerido ou monitorado, os erros ou vieses presentes nos *outputs* podem ser amplificados ao longo do tempo, resultando em uma perda de robustez e precisão.

A essência do trabalho de Schumailov et al. (2024), é que quanto maior a presença de substrato gerado por IAs na base de treinamento dos LLMs, maior é a deterioração da qualidade do conteúdo gerado ao final, e maior o risco de um colapso do modelo, que foi definido como sendo um processo degenerativo em que a IA erra a distribuição original dos dados, e passa a reproduzir algo sem sentido. Os autores apontam que, intuitivamente, é provável que esse fenômeno se repita em outras IAs generativas além dos LLMs, e ofereça um risco ainda maior do que o evidenciado.

Matinez et al (2023), que usam de fato o termo "*feedback loop*", encontraram essencialmente o mesmo resultado com as IAs generativas de imagem, como o DALL-E. Nesse trabalho, cujos parâmetros de metodologia são um pouco menos rígidos que o anterior, os autores essencialmente retroalimentaram uma IA geradora de imagem com imagens de outras IAs a fim de treinamento. Isso gerou imagens progressivamente piores conforme maior presença de imagens de IA no *pool* de treino.

A conclusão é similar: quanto maior a porcentagem de conteúdo gerado por IA presente na hora do treinamento, menor a qualidade, em termos de um parâmetro de avaliação humano, do resultado gerado.

2.1 Aumento da presença de conteúdo gerado por IAs na rede

Schumailov et al. (2024) já terminam seu artigo apontando o perigo que o aumento exponencial de conteúdo gerado por IA cria no contexto do colapso de modelos LLMs. É com base nesse *insight* que desenvolverei agora um panorama geral dessa situação para compreendermos que há, de fato, um risco grande da situação se agravar.

O aumento recente na quantidade de conteúdos gerados por IAs na web é um fenômeno impulsionado por uma combinação de avanços tecnológicos, demanda crescente por automação e a busca por eficiência na produção de conteúdo digital (Wu et al., 2023). Vários fatores têm contribuído para essa proliferação de conteúdo gerado por IA, que vão desde melhorias nos algoritmos de aprendizado de máquina até a expansão do acesso a dados e o desenvolvimento de plataformas mais acessíveis e poderosas para criação automatizada.

Primeiramente, os avanços em aprendizado profundo e técnicas de processamento de linguagem natural (NLP) têm sido fundamentais para o aumento da capacidade das IAs em gerar textos que se assemelham aos produzidos por humanos. Modelos como o GPT-3 e suas interações subsequentes são bons exemplos de transformadores que conseguem produzir textos longos e coesos a partir de um conjunto limitado de informações ou comandos (Colombo e Goulart, 2023). Esses modelos são treinados em grandes volumes de dados textuais disponíveis na Internet, permitindo-lhes capturar nuances, estilos e contextos variados. A capacidade desses modelos de gerar texto de maneira quase indistinguível do escrito por humanos resultou em sua adoção ampla para diversas finalidades, incluindo blogs, artigos de notícias,

resumos de textos, e até mesmo ficção. (Naveed et al., 2024, Vayadande et al., 2024)

Além desses avanços técnicos, a demanda por automação e escalabilidade na produção de conteúdo também desempenha um papel crucial (Naveed et al., 2024). Empresas de marketing digital, jornalismo e outros setores baseados em conteúdo estão cada vez mais voltadas para soluções que permitam a produção rápida e eficiente de grandes volumes de texto, sem comprometer a qualidade. A utilização de IAs generativas permite às empresas manterem um fluxo constante de conteúdo atualizado, melhorando sua presença online e otimização para mecanismos de busca (SEO), com menos dependência de trabalhadores humanos e menor custo operacional.

Outro fator importante é o aumento do acesso a essas tecnologias. Plataformas como GPT-3 foram disponibilizadas através de APIs, facilitando que desenvolvedores e empresas integrem modelos de linguagem avançados em seus aplicativos e websites. Com uma barreira de entrada mais baixa, um número maior de indivíduos e organizações têm a capacidade de explorar o potencial das IAs para geração de conteúdo. Além disso, ferramentas de IA foram incorporadas a softwares de criação de conteúdo, como editores de texto e geradores de imagem, tornando a geração de conteúdo automatizado mais simplificada.

Nessa pesquisa, para além de questões sobre qualidade desse conteúdo ou a eticidade da sua produção (Colombo e Goulart, 2023), o ponto de maior preocupação é a saturação do ambiente digital com conteúdo gerado por IAs. Como produzem grandes volumes de texto rapidamente, há um risco de inundação de conteúdo repetitivo ou sem valor agregado, mas com toda aparência de conteúdo humano. Além de afetar a confiança dos usuários na veracidade e qualidade dos conteúdos que encontram na web, dificultam muito a diferenciação para construção de bases de treinamento das próprias IAs—daí o risco de colapso dos modelos (Schumailov et al., 2024, e Martinez et al, 2023).

O aumento do risco de colapso de modelos de IA por retroalimentação pode ser compreendido através de uma análise das dinâmicas de como os modelos de IA são treinados e atualizados, bem como das implicações de uma proliferação massiva de conteúdo automatizado. Primeiro a retroalimentação de dados é um aspecto já inserido no treinamento e aprimoramento dos modelos de IA (Martinez et al., 2023). Esses sistemas frequentemente utilizam dados gerados por si mesmos ou por outros sistemas de IA como entrada para treinamento contínuo, mas de forma acessória. Quando a quantidade de conteúdo gerado por IA aumenta exponencialmente, o volume de dados disponíveis para treinamento pode crescer desproporcionalmente.

Além disso, a proliferação de conteúdo gerado por IA pode levar a uma concentração maior de informações semelhantes ou repetitivas na web (Wu et al., 2023). Modelos de IA, especialmente os baseados em técnicas de aprendizado supervisionado, são treinados em dados históricos que refletem o que já existe na internet. Se uma grande parte do novo conteúdo gerado é similar ou derivada de modelos anteriores, o treinamento subsequente pode ser influenciado por padrões redundantes e pouco diversos. Isso pode resultar na produção de outputs cada vez mais homogêneos e menos inovadores, uma vez que os modelos tendem a se consolidar em padrões repetitivos e menos diversificados.

Outro fator crítico é o impacto da desinformação e dos conteúdos manipulativos (Westerlund, 2019; e Vayadande et al., 2024). Modelos de IA que geram conteúdo para fins de propaganda, desinformação ou outros usos mal-intencionados podem contaminar o pool de dados disponível para treinamento. Quando esses conteúdos prejudiciais são amplamente

disseminados e usados para treinar novos modelos, eles podem contribuir para uma propagação de desinformação e manipulação em larga escala (Mariani, 2023). Isso não só compromete a integridade dos dados utilizados para treinamento, mas também amplifica o risco de que os modelos perpetuem e disseminem desinformação de forma mais intensa.

Adicionalmente, a integração de modelos de IA em sistemas complexos que interagem entre si pode criar redes de retroalimentação mais complexas e difíceis de controlar. Por exemplo, se vários sistemas de recomendação utilizam dados gerados por IA como entrada para suas próprias operações, a interação entre esses sistemas pode resultar em um feedback loop onde o conteúdo gerado se torna cada vez mais tendencioso ou enviesado. Esse fenômeno pode levar a um colapso progressivo na qualidade do conteúdo, onde as decisões e recomendações baseadas em dados automatizados tornam-se cada vez mais insatisfatórias ou prejudiciais.

As IAs geradoras de imagem e vídeo, integradas com outras como LLMs potentes, formam a base do que Westerlund (2019) chama de *deepfake technology*, as ferramentas para emular quase perfeitamente pessoas através de vídeo e áudio e, se devidamente treinadas com uma base de dados suficientemente larga, até mesmo imitar a forma com a pessoa se comunica, cacoetes, maneirismos etc. Vayadande et al. (2024), apontam uma proliferação violenta desses conteúdos, especialmente em contextos de manipulação política e midiática como em eleições.

O problema, em termos de colapso de sistemas de IA, é que esse conteúdo pode passar despercebido e entrar no pool de treinamento, contaminando a base de dados. É o mesmo problema que decorre do que apontam Veselovsky et al. (2023): um aumento muito grande do uso de LLMs pelos chamados *crowd workers*, algo como "trabalhadores de multidão", que são grandes massas de trabalhadores baratos contratados para tarefas simples e repetitivas. Uma dessas tarefas, não à toa, é produzir fragmentos textuais para treinamento de IAs.

A dinâmica de trabalho, porém, motiva esses trabalhadores a se utilizarem de IAs para produzir esses textos. No artigo dos autores, são apresentados números baseados numa análise da plataforma de *crowdsourcing* da Amazon, uma das mais utilizadas para terceirização dessas tarefas de forma rápida e com pouco custo, e pouca remuneração também. A essência do *crowdsourcing* é justamente transferir o ônus para um número grande de pessoas, o que dificulta uma verificação concreta do uso de IAs. Os pesquisadores encontraram evidencias que algo entre 33% e 46% dos *crowd workers* usaram IA nas suas tarefas textuais (Veselovsky et al., 2023).

Thompson et al. (2024), em um estudo que também tem reverberado bastante, criaram uma metodologia de análise para uma base de dados imensa de frases em dezenas de línguas diferentes, a partir de textos disponíveis publicamente na Internet. Na casa dos bilhões de frases, o estudo focou em investigar instâncias de frases repetidas de tal forma que indicassem serem traduções de uma mesma fase, o que cria um *cluster* formado por frases que são iguais, mas em diferentes línguas.

O que o estudo demonstrou, analisando os parâmetros dessas traduções, é que algo em torno de 57% das frases da base de dados estava em um *cluster* com três ou mais línguas, ou seja, eram frases traduzidas para pelo menos duas outras línguas além da língua de origem. Usando um parâmetro de análise de que quanto maior o número de frases num *cluster*, maior a probabilidade de ser um caso de *Machine Translation*, ou seja, tradução via IAs de texto, o estudo concluiu que, em muitas línguas do estudo, tradução via IA representa a maioria do conteúdo em texto disponível.

Claro que isso não significa que o conteúdo tenha sido originalmente criado por IA, e um texto mal traduzido é muito diferente de um

texto mal escrito. O estudo também aponta alguns vieses, como a impossibilidade de estudos mais individualizados em textos específicos ou que conseguissem identificar diferentes línguas de origens e corrigir distorções. Mas os autores revelam que buscaram algumas frases específicas que apareciam traduzidas em muitas línguas, por vezes dezenas de línguas diferentes, e encontraram coisas como blogs com dicas de comportamento no local de trabalho, relatos de pescaria, e uma série de outros conteúdos que caracterizaram como sendo de baixa complexidade.

Isso reverbera outros estudos, como os de Wu et al. (2023) e Naveed et al. (2024), que apontam uma proliferação muito rápida de textos de baixa complexidade, bastante expositivos ou pseudo-informativos, e consumidos de forma muito superficial. Esse fenômeno se encaixa no caráter viral que Parikka (2007) apontava para as redes digitais já no começo do século. Na metáfora do vírus, o autor tenta mostrar três elementos centrais da cultura na Internet: a velocidade de espalhamento e reprodução dos eventos culturais, sua independência do "hospedeiro" (a própria Internet enquanto infraestrutura) e, por fim, sua tendência violenta à disseminação fatal.

É nessa lógica que o conteúdo gerado por IA parece estar se espalhando de forma descontrolada e destrutiva para o hospedeiro, que nesse caso é a própria IA (Mariani, 2023). Agora, com alguma evidência sobre a doença em si, qual seja, o colapso de que falam Schmailov et al., (2024), a metáfora de Parikka (2007) começa a se mostrar de forma mais estruturada, e a virulência da internet de forma mais clara na velocidade e intensidade que caracterizam o aumento desses conteúdos na *web*.

3 Crescente dependência de sistemas de IA generativa

O problema do colapso de sistemas de IA não seria tão grave quando ferramentas como o Chat GPT eram ainda curiosidades usadas aqui e ali. Agora, porém, a curva de dependência de tecnologias de IA generativa parece aumentar tão rápido quanto a do aumento de conteúdo gerado por elas. Esse crescente uso reflete uma tendência global, em que tecnologias emergentes, especialmente as IAs generativas, estão se tornando cada vez mais integradas em diversos setores da sociedade (Cesarino, 2022).

Por um lado, esse fenômeno é impulsionado por avanços tecnológicos rápidos, maior acessibilidade às ferramentas de IA e uma aposta no seu uso para redução de custo com capital humano. No entanto, esse aumento no uso também levanta questões sobre excessiva dependência e implicações éticas para diferentes áreas, como saúde, educação, empresas e negócios, e até mesmo administração pública e uso governamental.

No setor de saúde, por exemplo, a IA está começando a desempenhar um papel significativo no diagnóstico, tratamento e gestão de doenças e pacientes (Lemes e Lemos, 2020). As IAs generativas têm sido estudadas como ferramentas para gerar imagens médicas de alta qualidade, auxiliar na interpretação de exames e até mesmo na descoberta de novos medicamentos, realização de anamnese com pacientes e uma série de outras possibilidades. Ferramentas de IA integradas à bases de dados são usadas para analisar grandes volumes de dados de pacientes e identificar padrões.

No estudo TIC Saúde, realizado pelo NIC (2024c), no módulo de novas tecnologias, foram incluídos indicadores para apontar informações mais detalhadas da aplicação de IAs.

Em 2023 em torno de 3.200 estabelecimentos de saúde usavam ferramentais de IA, a larga maioria na rede privada. Os usos foram: a automatização de fluxos de trabalho (46%), reconhecimento de fala (33%), mineração de texto ou análise de linguagem escrita e falada (32%), reconhecimento e processamento de imagens (21%), e por fim treinamento para predição e análise de dados (16%).

Os estabelecimentos entrevistados apontaram que sua escolha de IA almejou melhorar a segurança digital (45%), apoiar a organização de processos clínicos e administrativos (41%), melhorar a eficiência dos tratamentos (38%), apoiar a gestão de recursos humanos e recrutamento (28%) e auxiliar na dosagem de medicamentos (16%). Esses números revelam que a IA aparece não somente como uma tecnologia importante, por exemplo, para a proteção de dados no setor de saúde, algo esperado, mas também como um potencial elemento central nos atendimentos e consultas.

No campo da educação, as IAs generativas já começaram a ser empregadas na criação de conteúdos educacionais personalizados, exercícios e até mesmo auxiliar na administração de instituições de ensino. Plataformas de ensino online utilizam algoritmos para adaptar o conteúdo às necessidades e ao ritmo de aprendizagem dos alunos, promovendo uma abordagem mais individualizada. Além disso, *chatbots* e assistentes virtuais estão sendo utilizados para responder a perguntas dos alunos e oferecer suporte em tempo real (Tavares, Meira e Amaral, 2020).

O estudo do TIC Educação (2023) revela a consolidação de uma tendência de adoção dessas tecnologias nas escolas, com muitos professores manifestando apoio a uma maior inserção e muitas escolas já começando a adotar essas ferramentas como auxiliares do ensino, e estruturando programas de ensino sobre IA e começando a olhar a possibilidade de tutores automatizados. A crescente dependência dessas ferramentas pode melhorar a acessibilidade e a eficiência no processo de ensino, mas exige que se considere a qualidade e a equidade do conteúdo gerado, bem como o impacto na interação humana no ambiente educacional.

Já o TIC Empresas (2024a) mostra que empresas brasileiras estão cada vez mais adotando sistemas de IA generativa para otimizar operações, melhorar o atendimento ao cliente e criar novas oportunidades de negócio. Ferramentas de geração automática de texto são utilizadas para criar marketing de conteúdo, gerenciar interações com clientes por meio de chatbots e analisar grandes volumes de dados para tomada de decisões estratégicas. Além disso, a automação de processos e a análise preditiva são apostas das empresas para reduzirem custos e aumentar a eficiência.

Esse estudo revelou uma estabilidade no percentual de adoção de tecnologias de IA, que passou de 13%, em 2021, para 14%, em 2023. Esse uso é mais recorrente nas grandes empresas e no setor de informação e comunicação, e a prática mais recorrente é utilização de IA para "automatização de processos de fluxo de trabalho" (73%), com aplicações menos simples, como machine learning (16%) ou geração de linguagem natural (13%), aparecendo de forma mais periférica, o que qualifica uma tendência interessante do ponto de vista da inovação.

No setor governamental, a IA aparece como uma saída para melhorar a eficiência dos serviços públicos e a gestão de recursos, empregada para analisar dados de políticas públicas, tentar prever necessidades sociais e otimizar a alocação de recursos (Toledo e Mendonça, 2023). Além disso, a IA é usada em processos de automação para simplificar burocracias e oferecer serviços mais rápidos e acessíveis à população (Lemes e Lemos, 2020). Na essência, o setor governamental parece também apostar na ferramenta de forma mais ou menos conservadora, ainda que haja uma tendência de aparecer cada vez mais nas interações com cidadãos.

De modo geral, é possível ver uma crescente participação das tecnologias de IA em diversos setores brasileiros, e de origens variadas. Setores como saúde e educação, por exemplo, carregam um peso muito grande em razão de sua natureza mais delicada e sensível. O resultado final, porém, é que há uma maior dependência de uma tecnologia muito recente, e que parece se consolidar de forma muito intensa nos últimos dois ou três anos.

4 O risco social do colapso por retroalimentação

Essa onda de crescente dependência social nas tecnologias de IA pode ser interpretada e conceituada a partir da teoria do risco de Ulrich Beck (2011), que enfatiza como os avanços tecnológicos e científicos, embora tragam benefícios, também introduzem novos riscos que são inerentes à própria modernidade tardia. Para o sociólogo, as sociedades contemporâneas, definidas como sociedades de risco, são caracterizadas por sua capacidade de criar perigos sistêmicos que transcendem fronteiras, têm impactos globais e são frequentemente imprevisíveis e incontroláveis.

Sob essa perspectiva, a difusão das tecnologias de IA pode ser vista como um elemento da sociedade de risco, na medida em que a dependência crescente de sistemas de IA gera uma série de novos riscos, ao mesmo tempo que oferece aparentes soluções promissoras para problemas sociais e econômicos. Esses riscos são tanto tecnológicos quanto sociais, e surgem principalmente porque as tecnologias de IA são altamente complexas, muitas vezes opacas e difíceis de serem completamente compreendidas ou controladas. Essa incerteza está diretamente ligada ao conceito de risco de Beck.

O autor define o conceito de "risco" de maneira a refletir suas preocupações com as sociedades modernas tardias e as mudanças sociais associadas à globalização e à modernidade reflexiva. Para Beck (2011), o conceito de risco deve ser central na análise das dinâmicas contemporâneas e nas formas com que a sociedade lida com as incertezas e perigos associados ao progresso tecnológico e social.

Beck (2011) descreve o risco como um fenômeno social que emerge das condições e consequências do avanço tecnológico e industrial. Mais do que um evento incerto ou perigoso, ele o conceitua como um aspecto intrínseco da modernidade (Beck, 2011), e argumenta que a sociedade moderna tardia está caracterizada por uma crescente preocupação com riscos que são, em grande parte, produtos das próprias inovações e práticas da sociedade industrial.

Para o autor, o risco tem algumas características principais. Uma é seu caráter global, já que os riscos não são mais confinados a áreas locais ou nacionais, mas têm uma dimensão global. Problemas como mudanças climáticas, poluição e crises financeiras são exemplos de riscos que transcendem fronteiras e exigem respostas internacionais. Como Beck (2011) ressalta, os riscos modernos não respeitam fronteiras.

A dependência da IA não é apenas uma questão de uma empresa ou país, mas uma transformação paradigmática global, especialmente com a integração dessas tecnologias em setores cruciais, como saúde, transporte, finanças e governança. A interconectividade digital significa que uma falha em um sistema de IA pode ter consequências de longo alcance. Essa interconectividade transnacional aumenta muito a escala dos riscos, tornando-os mais difíceis de se conter ou controlar.

Outra característica que Beck (2011) vê nos riscos da modernidade tardia é seu aspecto de produção interna. Muitos dos riscos contemporâneos são, em grande parte, uma consequência do próprio avanço tecnológico e das

práticas modernas. A industrialização, a urbanização e a globalização têm gerado novos tipos de riscos que não existiam antes, como os riscos ambientais e tecnológicos.

A IA, como produto da modernidade, ilustra bem esse ponto. Ela é desenvolvida para otimizar processos, automatizar tarefas e gerar valor econômico, mas, ao mesmo tempo, sua implementação cria novos tipos de vulnerabilidades. Por exemplo, algoritmos de IA podem perpetuar vieses discriminatórios ou falhar em decisões críticas em áreas como saúde e segurança, trazendo riscos para os indivíduos e para a sociedade. Essas falhas podem ser imprevisíveis e difíceis de mitigar devido à complexidade dos sistemas e à sua autonomia crescente.

Também podemos pensar na construção de Beck (2011) para a incerteza e a contingência dos riscos contemporâneos. Esses são frequentemente associados a incertezas e à dificuldade de prever ou controlar suas consequências, o que se deve à complexidade dos sistemas tecnológicos e sociais envolvidos, bem como à falta de compreensão completa dos impactos de novas tecnologias e práticas.

Com a IA, esse fenômeno é particularmente relevante, pois os sistemas de IA funcionam em grande parte nos bastidores de muitas atividades cotidianas. A sociedade confia cada vez mais em decisões automatizadas que são tomadas por algoritmos cuja lógica interna é frequentemente opaca (o chamado "problema da caixa-preta"). Isso torna difícil identificar e antecipar quando e onde um risco pode surgir, seja na forma de erro, viés ou uso indevido de dados pessoais.

Outro ponto importante para Beck (2011) é que o enfrentamento dos riscos frequentemente revela desigualdades sociais e injustiças, pois os grupos vulneráveis podem ser desproporcionalmente afetados pelos riscos, enquanto aqueles que têm mais recursos podem se proteger melhor. Beck (2011) aponta que os riscos são muitas vezes distribuídos de maneira desigual, refletindo e exacerbando as desigualdades existentes. A dependência de sistemas de IA, no geral, cria um grande risco nas pontas do sistema como, por exemplo, em cidadãos usuários dos sistemas de saúde que passam a se integrar com essas tecnologias.

As populações mais vulneráveis, como trabalhadores menos qualificados, vão sofrer as consequências da automação e da substituição de empregos pela IA, enquanto aqueles com maior acesso a capital tecnológico podem se beneficiar desproporcionalmente. Grohmann e Araújo (2021) já apontavam como o “chão de fábrica” das IAs é composto, fundamentalmente, de trabalho precarizado e mal remunerado, que surge a partir de uma relação de exploração baseada em condições econômicas desiguais.

Zajko (2022) aponta que a sociologia, em especial estudos voltados para identificar desigualdades estruturais e estruturantes, oferece uma perspectiva crítica sobre as verdadeiras possibilidades de uma eliminação dos vieses sociais que são reproduzidos nas tecnologias. Sabemos, como mostra Silva (2022), que a vigilância automatizada e os sistemas de IA utilizados por governos e corporações podem impactar mais negativamente grupos marginalizados, aumentando desigualdades sociais e expondo-os a mais riscos, como discriminação algorítmica ou vigilância desproporcional.

Por fim, Beck (2011) levanta uma importante questão sobre mudanças radicais de valores e normas sociais em função da nova dinâmica de riscos. Esses novos problemas desafiam as normas e valores tradicionais e exige uma reavaliação das práticas sociais e políticas. O autor sugere que, à medida que a sociedade toma consciência desses novos riscos, ela se torna reflexiva, ou seja, começa a questionar e reavaliar os próprios fundamentos conceituais da modernidade. É uma mudança de paradigmas sociais que sustentam nosso modelo atual de estrutura coletiva.

5 Considerações Finais

Com o "*framework*" dos riscos da modernidade tardia de Beck (2011) em mente, podemos analisar o problema do colapso das IAs por retroalimentação como um desdobramento dessa nova dinâmica de riscos. Há o surgimento de uma nova tecnologia, com uma série de promessas de avanços, revoluções de custos e eficiência etc., seguido de um espalhamento muito intenso dessa tecnologia de forma financiada e coordenada. Em seguida, essa tecnologia passa a ser integrada em uma série de outros sistemas e serviços sem uma preocupação muito grande com o aumento exponencial da complexidade e dos riscos associados a isso.

Assim, uma vez estabelecido o perigo gerado pelo aumento desproporcional de conteúdo gerado por IA e seu papel nesse colapso por retroalimentação (Schumailov et al., 2024), temos o cenário perfeito de risco conforme descrito por Beck (2011): um perigo globalizado, integrado, que atinge de forma desigual os diferentes atores sociais, e que gera incerteza e dificuldade de contenção, construído pelo próprio avanço da modernidade tardia. Esse risco, ainda segundo Beck, pode ser interpretado como um desdobramento da própria modernidade reflexiva.

Nesse sentido, a atual dinâmica de desenvolvimento de IAs emula o padrão que aparece ao longo do período moderno: há uma enorme capitalização individual do lucro, com uma grande distribuição dos riscos e das consequências. Para além de questões ambientais, trabalhistas e éticas, as tecnologias de IA agora apresentam mais um risco, fruto da proliferação desmedida do seu próprio conteúdo pela internet. A possibilidade de colapso desses sistemas é um perigo concreto para todas as áreas que passarem a depender dos LLMs.

V. 5 / N. 1 / JULHO DE 2024
PÁGINAS 94 A 107

Notas finais

1 Órgão de pesquisa vinculado ao Núcleo de Informação e Coordenação do Ponto BR (NIC.br), que por sua vez é ligado ao Comitê Gestor da Internet do Brasil (CGI.br). Tem a missão de monitorar a adoção das tecnologias de informação e comunicação (TIC) no Brasil. Disponível em: https://cetic.br/pt/pesquisas/. Incluir o portal da instituição.