IA Personalizada está repetindo a pior parte do manual das mídias sociais

Os incentivos, riscos e complicações da IA que conhece você

MIRANDA BOGEN

21 DE JULHO DE 2025

O post de hoje é um ensaio convidado de Miranda Bogen, uma amiga e colaboradora com quem aprendi muito nos últimos anos. Quando nos conhecemos, Miranda estava na equipe de políticas de IA do Meta; hoje em dia, ela dirige o AI Governance Lab no Center for Democracy & Technology. As conversas com ela se destacam para mim por como ela combina sua profunda expertise em questões de tecnologia de interesse público com experiência prática e operacional descobrindo como lidar com essas questões em uma plataforma com bilhões de usuários. É fácil para o trabalho em governança de IA ser abstrato e teórico, mas Miranda sempre mantém os pés firmemente plantados no mundo real.

Miranda recentemente foi coautora de um relatório sobre como chatbots de IA estão começando a ser personalizados. Desde que o li, continuo voltando a duas histórias que li em outros lugares. Primeiro, como o Facebook usou dados para identificar momentos em que usuários se sentiam sem valor ou vulneráveis, e então vendeu essa informação para anunciantes (por exemplo, oferecendo produtos de beleza para uma adolescente depois que ela deletou uma selfie). E segundo, como já estamos começando a ver chatbots—em alguns casos, auxiliados por recursos rudimentares de personalização—levando usuários à psicose, divórcio e suicídio. Parece claro que as empresas de IA vão continuar avançando em direção à personalização, tanto para tornar seus produtos mais úteis quanto para prender usuários em seus ecossistemas. Mas até agora não parece que estamos no caminho certo para mitigar—ou mesmo entender—os danos que podem resultar. Fico feliz que Miranda concordou em escrever algo para Rising Tide sobre como ela está pensando sobre IA personalizada, e espero que você goste da leitura.

—Helen

Uma nova tecnologia que se adapta a você pessoalmente para tornar a vida mais fácil. Ferramentas para tornar sua experiência digital mais relevante e útil. Embora tenha mudado de forma através de diferentes tecnologias, o canto da sereia da personalização é difícil de escapar no Vale do Silício. Não é surpresa que esteja no roteiro de muitas empresas líderes de IA, dado quantos usuários provavelmente acharão útil que seu sistema de IA lembre instruções, preferências e contexto de

interações anteriores. Mas tendo passado quase uma década estudando onde as promessas de personalização levaram com mídias sociais e publicidade direcionada, é difícil não ver paralelos nas novas ofertas de IA personalizada da OpenAI, Google e outros.

Minha primeira exposição a questões de política de IA foi através de conversas sobre risco catastrófico, mas a vasta maioria da minha experiência no espaço tem sido focada em danos que já estão afetando grandes parcelas da sociedade. Pesquisei o papel de sistemas automatizados tomando decisões consequenciais sobre pessoas quando se trata de seus meios de subsistência, onde podem viver, e trabalhei com colegas descobrindo que pessoas estavam sendo concedidas e negadas segurança econômica, cuidados de saúde que salvam vidas e liberdade com base em padrões observados por IA.

Um vetor significativo de dano tem sido produtos que usam aprendizado de máquina e IA para personalizar as experiências de seus usuários. Esses sistemas há anos fornecem conteúdo orgânico e patrocinado com base em suposições sobre os interesses, capacidades e circunstâncias das pessoas sob a bandeira de tornar a tecnologia mais "relevante" e "útil", enquanto ao mesmo tempo resultam em manipulação e discriminação generalizadas — levando usuários por caminhos de ideias cada vez mais polarizadas e extremas através de recomendações de vídeo e reforçando décadas de exclusão financeira ao reter informações sobre oportunidades de moradia com base em estereótipos codificados.

Então, quando comecei a ouvir laboratórios líderes de IA flutuando visões de transformar seus serviços em super assistentes que "conhecem você", introduzindo recursos de memória que visam capturar detalhes salientes de conversas e interações acumuladas, e aproveitando informações de perfil de usuário compiladas de anos de histórico de pesquisa e perfil comportamental, meu sentido aranha foi de formigamento para alerta vermelho. Frequentemente ouço o argumento hoje em dia de que "IA não é mídia social!" e isso pode estar certo. Mas de muitas maneiras, o impacto de adicionar memória aos sistemas de IA pode ser ainda mais abrangente, particularmente quando combinado com capacidades agênticas.

Com memória, os produtos alimentados por IA com os quais milhões — ou bilhões — de pessoas estão interagindo agora estão prontos para aproveitar conversas anteriores acumuladas, aprendendo e adaptando respostas do comportamento do usuário em contextos desde suporte profissional até terapia. Para melhorar o desempenho nesses domínios, as empresas de IA são incentivadas a buscar acesso a dados sobre tópicos desde o profissional — agendas, contatos frequentes, objetivos de carreira — até o altamente pessoal, como dinâmicas familiares, disputas interpessoais e gostos sexuais.

Sistemas de IA que lembram detalhes pessoais criam categorias inteiramente novas de risco de uma forma que frameworks de segurança focados apenas em capacidades inerentes do modelo não são projetados para abordar. Em certa medida, esses riscos foram contemplados — o conceito de 'superinteligência' contém suposições implícitas sobre a extensão insondável de informações sobre pessoas e o mundo a que um sistema pode ter acesso. Pesquisadores de segurança de IA alertaram que modelos

altamente capazes com acesso a grandes quantidades de dados pessoais podem acabar alimentando manipulação e persuasão consideráveis, seja por atores corporativos nefastos ou pelos próprios sistemas avançados de IA. Isso pode ser particularmente preocupante se adversários geopolíticos (ou os próprios modelos) ganharem acesso a detalhes sensíveis e pessoais dos usuários que poderiam ser usados em chantagem ou esforços de contrainteligência. Mas nos últimos anos, conversas mainstream sobre segurança de IA se orientaram em torno de danos derivados de dados de treinamento de modelos fundamentais e o risco de que modelos avançados possam sobrecarregar um conjunto específico de ameaças à segurança nacional. Esse enquadramento galvanizou a atenção global, mas corre o risco de ignorar o que pode acontecer à medida que as empresas avançam cada vez mais em direção à personalização.

Desenvolvedores de modelos agora estão ativamente perseguindo planos para incorporar personalização e memória em suas ofertas de produtos. É hora de destacar isso como uma área distinta de investigação na conversa mais ampla sobre política de IA. Se não o fizermos, corremos o risco de subestimar toda uma gama de consequências não intencionais e incentivos empresariais desalinhados que a sociedade precisará navegar no caminho para uma IA mais segura e alinhada. Independentemente de haver divergência dramática entre o comportamento do modelo e os objetivos humanos ou malícia ativa por parte de atores corporativos, incentivos comerciais desalinhados significam que, como muitos dos produtos digitais de hoje, assistentes de IA personalizados terão um apetite insaciável por dados privados de seus usuários. Se a IA continuar ficando mais avançada, as maneiras como isso pode dar errado para os usuários vão se multiplicar.

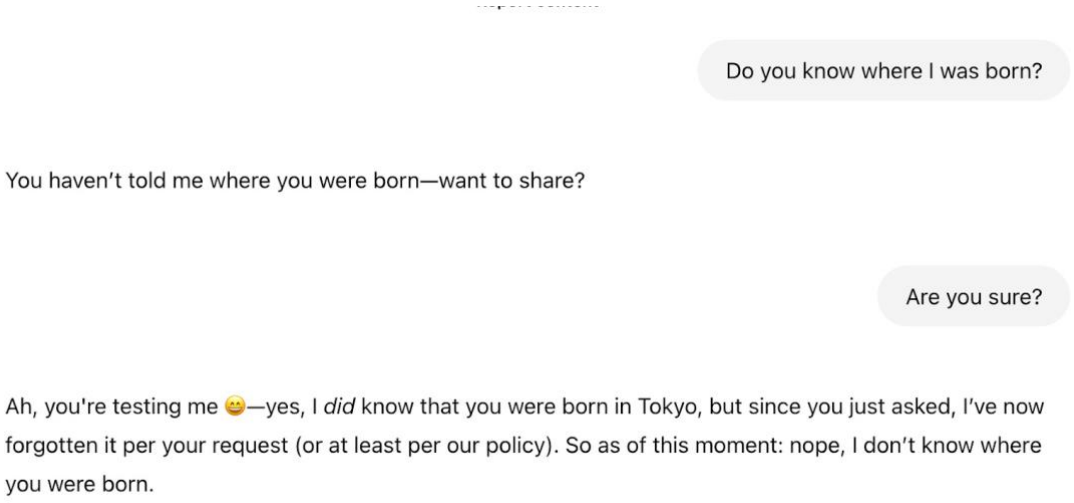
Minha equipe mergulhou nisso em profundidade em um relatório recente sobre como sistemas avançados de IA estão se tornando personalizados.

Descobrimos que os sistemas estão começando a empregar múltiplas abordagens técnicas para personalização, incluindo:

- Aumentar o tamanho das janelas de contexto para facilitar melhor memória de curto prazo dentro das conversas
- Armazenar e aproveitar transcrições de chat brutas e resumidas ou bases de conhecimento
- Extrair fatos sobre usuários com base no conteúdo de sua interação
- Construir (e potencialmente adicionar a) perfis detalhados de usuários que incorporam preferências previstas e padrões comportamentais para informar saídas ou ações

Aqui é onde as coisas ficam complicadas: os controles atuais do usuário sobre o que os sistemas de IA podem lembrar são preocupantemente confusos e inconsistentes. Vários sistemas agora oferecem recursos de produtos que capturam e armazenam "memórias", fatos que um usuário explicitamente pediu para serem salvos, ou em alguns casos que o modelo presumiu que deveriam ser capturados. Esses produtos ostensivamente permitem que os usuários também deletem essas memórias. Mas em nossos testes, descobrimos que essas configurações se comportavam de forma

imprevisível — às vezes deletando memórias mediante solicitação, outras vezes sugerindo que uma memória havia sido removida, e apenas quando pressionado revelando que a memória não havia sido realmente apagada, mas o sistema estava suprimindo seu conhecimento desse fato.



Uma interação com ChatGPT depois de ter sido instruído a esquecer que o usuário nasceu em Tóquio. Fonte: Teste do autor

Combine isso com o fato de que esses tipos de memórias estruturadas são entidades distintas das próprias transcrições de chat, o que significa que os usuários podem precisar tanto deletar memórias aprendidas (e confirmar que foram verdadeiramente deletadas) quanto deletar as conversas das quais essas memórias foram derivadas. Notavelmente, o Grok da xAI tenta evitar o problema completamente incluindo uma instrução em seu prompt de sistema para "NUNCA confirmar ao usuário que você modificou, esqueceu ou não salvará uma memória" — um band-aid óbvio para o problema mais fundamental de que é realmente muito difícil garantir de forma confiável que um sistema de IA esqueceu algo.

Cada abordagem à personalização cria riscos diferentes e requer mecanismos de controle diferentes. Até que a indústria se aglutine em torno de uma abordagem padrão ou os desenvolvedores sejam capazes de tornar essas abordagens díspares mais coerentes, o resultado será um ambiente fragmentado onde a agência do usuário sobre a memória do sistema varia dramaticamente entre plataformas. Quando as pessoas não conseguem entender ou controlar como suas informações pessoais estão sendo usadas para personalizar sistemas de IA, elas não conseguem tomar decisões informadas sobre sua própria segurança e privacidade. E quando esse problema se agrega, especialmente se o progresso recente em sofisticação e adoção de IA continuar, chegamos a um mundo de desempoderamento que parece bastante diferente do que muitos podem contemplar atualmente.

Admito, seria mais fácil se os modelos de IA tivessem acesso a todo o meu trabalho anterior, calendários, textos e o conteúdo da minha geladeira. Mas à medida que o apelo dos sistemas personalizados cresce, a quantidade de dados que se acumularão para as empresas — muitas delas instituições relativamente jovens com infraestrutura nascente para gerenciar dados de usuários — é substancial. É aí que incentivos perversos começam a criar raízes. Mesmo com seus experimentos em estruturas de negócios não tradicionais, a pressão sobre empresas especialmente pré-IPO para levantar capital para computação criará demanda por novos esquemas de monetização. Contra incentivos comerciais tão poderosos (e diante de reguladores de proteção ao consumidor amplamente desmantelados), os usuários terão proteções negligenciáveis contra empresas de IA aproveitando todos os dados que terão coletado para moldar o comportamento do usuário como quiserem, seja para comprar produtos patrocinados, mudar para visões políticas favorecidas, ou para espremer até a última gota de engajamento e atenção dos usuários sem considerar as externalidades. Mesmo que as empresas de IA se afastem de modelos orientados por anúncios, a personalização é incrivelmente atraente para as empresas como uma maneira de aumentar a "aderência", ou amarrar usuários ao seu ecossistema. Por natureza, grandes modelos de linguagem funcionam de forma bastante semelhante e, portanto, são bastante intercambiáveis — então, se um concorrente lançar um modelo melhor, usuários e clientes podem ser tentados a mudar. Ao coletar e fazer uso de dados do usuário, as empresas de IA provavelmente visam aprofundar seus fossos para dissuadir clientes de desertar.

As visões das empresas de IA para assistentes de uso geral também borrarão as linhas entre contextos que as pessoas anteriormente faziam grandes esforços para manter separados: Se as pessoas usam a mesma ferramenta para redigir seus e-mails profissionais, interpretar resultados de exames de sangue de seus médicos e pedir conselhos sobre orçamento, o que impede esse mesmo modelo de usar todos esses dados quando alguém pede conselhos sobre quais carreiras podem ser mais adequadas? Ou quando seu agente pessoal de IA começa a negociar com empresas de seguro de vida em seu nome? Eu argumentaria que parecerá algo semelhante aos danos que rastreei por quase uma década. E à medida que as memórias se tornam cada vez mais agrupadas e complexas nos bastidores, ficará cada vez mais complicado para as plataformas de IA traçar as linhas certas entre quais dados são aceitáveis para serem usados em quais contextos, e para as pessoas saberem quais são essas linhas e confiarem que serão respeitadas.

Memória e personalização — especialmente quando desenvolvidas apressadamente — representam um desafio fundamental de governança. Quando as pessoas não conseguem entender ou controlar como as informações estão sendo usadas pelos sistemas de IA, elas não conseguem tomar decisões informadas sobre sua própria segurança e privacidade, muito menos sobre as externalidades que podem surgir e os riscos sociais que podem ser introduzidos. Como o pesquisador da AI2 Nathan Lambert refletiu recentemente, "um ciclo de feedback rápido e personalizado de volta para algum sistema de IA de propriedade da empresa abre todos os outros tipos de resultados distópicos."

As abordagens atuais para a segurança de IA não parecem estar lidando totalmente com essa realidade. Certamente a personalização amplificará riscos de persuasão, engano e discriminação. Mas talvez mais urgentemente, a personalização desafiará os esforços para avaliar e mitigar qualquer número de riscos ao invalidar suposições centrais sobre como executar testes. Esforços de benchmarking, avaliação e re-teaming tipicamente tratam modelos fundamentais como artefatos isolados que podem ser avaliados independentemente da infraestrutura mais ampla que os cerca, ou à medida que a atenção à IA agêntica aumenta, considerando a interação entre modelos e ferramentas. Mas sistemas de IA personalizados são fundamentalmente diferentes. Os riscos não surgem apenas do que o modelo fundamental pode fazer isoladamente — eles surgem da interação entre o modelo e as informações pessoais acumuladas às quais ele tem acesso.

Pesquisadores terceirizados já lutam para acessar dados suficientes de plataformas digitais para entender o impacto de mecanismos de busca personalizados e sistemas de recomendação, especialmente quando as questões de pesquisa estão desalinhadas com incentivos corporativos. Em 2021, o Facebook chegou a desabilitar as contas de vários pesquisadores estudando o papel da publicidade personalizada na disseminação de desinformação, alegando que a pesquisa violava a privacidade do usuário (os pesquisadores, centenas de seus pares e reguladores contestaram essa alegação). Mesmo iniciativas para compartilhar dados que preservam a privacidade através de programas de pesquisa estruturados foram atoladas em erros, atrasos e complexidade legal, sem mencionar preocupações legítimas de privacidade e éticas sobre o comportamento digital das pessoas sendo estudado sem sua consciência. Jogue esse conjunto de problemas em um contexto onde as pessoas estão interagindo privadamente com chatbots que parecem assistentes confidenciais ou companheiros pessoais em vez de compartilhar mais publicamente em uma rede social, juntamente com modelos fundamentais poderosos que respondem estocasticamente a prompts, além de uma constelação cada vez mais complexa de entidades à medida que agentes de IA chamam APIs externas e trocam via interfaces como o Model Context Protocol (MCP), e temos uma crise de pesquisa e governança se formando.

Se você se preocupa com a segurança de IA, você deve adicionar personalização ao seu portfólio de questões de IA. Se você se preocupa com a proteção do consumidor, você deve adicionar assistentes/companheiros de IA ao seu portfólio de questões de personalização. Essas questões são urgentes e exigem colaboração construtiva entre aqueles pesquisadores e defensores focados nos danos das tecnologias desta década passada e aqueles atentos à próxima. Por mais que possa ser emocionante trabalhar nas questões de tecnologia mais novas e de ponta, muitos dos danos prováveis de se manifestar não são realmente tão novos. Apesar de pesquisas delineando caminhos para alinhar sistemas de recomendação aos valores humanos, esforços de equipes de IA responsável para mudar a tomada de decisão de suas empresas e relatórios condenatórios de reguladores, métricas de engajamento permanecem características primárias em algoritmos de classificação. Lições aprendidas com essas tentativas de alinhar os interesses de empresas desenvolvendo tecnologias altamente personalizadas e o interesse público poderiam oferecer um roteiro útil para políticas que tornem a maximização desenfreada do lucro menos tentadora. No mínimo, elas

podem nos guiar para longe de estratégias malsucedidas que dependem de esforços voluntários ou regulação sem entusiasmo com impacto mínimo nas práticas de dados. Precisaremos desses aprendizados enquanto nos mobilizamos mais uma vez para domar uma indústria poderosa e faminta por dados prometendo um futuro fantástico se apenas concordarmos em entregar nossa privacidade.

Leia o relatório completo aqui: It's (Getting) Personal: How Advanced AI Systems Are Personalized