cd:// conselho digital

NOTA DE TEMA

Inteligência Artificial: Autenticação de Conteúdo Gerado por Aplicações de IA

Como identificar conteúdos gerados por inteligência generativa

ABRIL / 2026

Sobre o Conselho Digital

O Conselho Digital é uma entidade brasileira, sem fins lucrativos ou afiliações políticas, que coordena, estuda e representa o ecossistema dos aplicativos de internet e toda a diversidade dos seus modelos de negócios.

Nossa organização acredita que a tecnologia, quando bem construída e utilizada, é uma porta para o futuro. Ela nos mantém conectados, potencializa habilidades, desenvolve novas oportunidades e pode mudar a vida das pessoas para melhor.

Partindo dessa premissa, atuamos através de estudos, eventos e atividades de advocacy em favor de políticas públicas e setoriais que fortaleçam uma internet livre, segura e responsável no Brasil e no mundo.

Defendemos políticas que respeitem a neutralidade tecnológica, a inovação e a diversidade de modelos de negócios; e que tenham como consequência:

- Usuários conscientes e com poder de escolha;
- Uma sociedade plural e próspera;
- Ambientes de negócio juridicamente seguros;
- Mercados abertos e dinâmicos; e
- Empresas responsáveis e competitivas.

Por fim, assumimos o compromisso de construir um ambiente harmonioso e produtivo entre nossos associados, assim como uma relação transparente e colaborativa com a sociedade e governo.

Diretor-Executivo
www.conselhodigital.org.br

Takeaways – Posição do Conselho Digital

- **Regulação excessivamente prescritiva tende a gerar efeitos colaterais indesejados**, como inibir inovação e limitar aplicações legítimas de IA além da generativa. Ao impor um único método de autenticação, corre-se o risco de obsolescência regulatória e de travar soluções mais eficazes que surjam no curto prazo.
- **A combinação de múltiplas técnicas de autenticação aumenta a confiabilidade sistêmica**, ainda que nenhuma seja 100% eficaz isoladamente. A consequência prática é maior transparência e rastreabilidade, reduzindo riscos reputacionais e danos informacionais em ambientes sensíveis, como eleições.
- **Marcas d'água visíveis promovem transparência imediata, mas podem comprometer usabilidade e ser facilmente removidas ou imitadas.** Isso pode gerar falsa sensação de segurança ou até ampliar a desinformação caso rótulos sejam manipulados.
- **Marcas invisíveis e metadados fortalecem rastreabilidade técnica, mas são vulneráveis a alterações de formato e aumentam custos operacionais.** A consequência é a necessidade de infraestrutura robusta e coordenação entre atores para garantir eficácia contínua.
- **Autenticação humana reduz riscos não intencionais e vieses técnicos, mas introduz lentidão e inconsistência.** Em larga escala, isso pode gerar gargalos operacionais e decisões despadronizadas, especialmente quando há tentativa deliberada de enganar revisores.
- **Técnicas emergentes (análise comportamental, detecção estatística, blockchain) ainda não oferecem maturidade suficiente para uso isolado.** Apostar exclusivamente nelas pode gerar alta taxa de falsos negativos ou custos desproporcionais frente aos benefícios.
- **Empoderar plataformas e criadores a utilizar ferramentas diversas tende a produzir maior transparência do que impor soluções fechadas.** A consequência é um ecossistema mais adaptável, capaz de evoluir com a tecnologia e responder de forma dinâmica a novos riscos.
- **A autenticação valida procedência, mas não garante veracidade do conteúdo.** Mesmo conteúdos falsos podem ser corretamente marcados como gerados por IA, o que exige políticas complementares de educação midiática e responsabilização adequada.

Qual a posição do Conselho Digital?

- Desde a rápida popularização de chatbots de inteligência artificial (IA) a **IA generativa** tornou-se um fenômeno global. Enquanto os chatbots capturaram grande parte da atenção pública focada na IA generativa, **deve-se reforçar que a IA do tipo generativa é apenas um tipo de IA.**

- Ao focar-se excessivamente em apenas um tipo de aplicação, tende-se a discutir regulações robustas e por vezes limitantes a outros tipos de tecnologia, impedindo uma infinidade de serviços básicos de uso rotineiro como tradutores e serviços de coleta e classificação de dados.

- Isso significa que **existem muitas formas de se gerar conteúdo através de aplicações de IA generativa**. Muitas delas oferecem novas vantagens, especialmente para fins de acessibilidade e para aumentar a criatividade humana.

- Embora o conteúdo gerado por IA não seja novo, sua crescente sofisticação e adoção — e a resposta do público geral, indústria e governos — cultivaram um senso de urgência para aproveitar a tecnologia para benefício social e minimizar os danos que podem surgir de seu uso.

- Ao se referir a sistemas de alto risco, o PL 2338/2023 traz em seu artigo 19 que “Quando o sistema de IA gerar conteúdo sintético, deverá, considerando o estado da arte do desenvolvimento tecnológico e o contexto de uso, **incluir identificador em tais conteúdos para verificação de autenticidade** ou de características de sua proveniência, modificações ou transmissão, conforme regulamento.”

- **Porém, o PL não define que tipo de identificador deverá ser utilizado, quais técnicas de verificação de autenticidade são aceitas, nem quem seria responsável por sua aplicação.**

- **O Conselho Digital acredita que o ambiente regulatório mais propício para o desenvolvimento de técnicas eficazes de autenticação é um que permita a inovação e a experimentação.** Só através de testes constantes e consistentes de diversos métodos se chegará a sistemas de verificação que sejam úteis para a sociedade e que possam ser implementados por diversos atores na cadeia de produção.

O que é Autenticação de Conteúdo Gerado por IA?

- **Técnicas de autenticação de conteúdo são práticas comuns no ramo da computação e cibersegurança, envolvendo a verificação da identidade de um usuário, processo ou dispositivo quando do uso de determinada ferramenta.** Neste caso, a verificação é de dados, modelos e o resultado visível da operação de processamento de dados de uma aplicação de IA (*outputs*).

- A autenticação de conteúdo gerado por IA, nesse sentido, refere-se a ao **ato de determinar quando certo conteúdo é gerado por uma aplicação de IA ou, de outra forma, verificar sua procedência**, utilizando mecanismos de autenticação como métodos criptográficos e verificação humana.

- Embora existam várias técnicas que podem ser usadas para autenticar IA, **este guia explora uma das modalidades mais robustas atualmente: o rotulamento (também conhecido como marca d'água), além de outros métodos que dependem da autenticação humana.** Técnicas de autenticação se sobrepõem ao longo da cadeia de valor da IA e resultam em formas distintas de soluções de rotulamento em diferentes técnicas de autenticação de IA. Esta dinâmica e outras técnicas serão exploradas mais adiante no relatório.

- **As técnicas de autenticação de IA variam em utilidade e facilidade de implementação em toda a cadeia de valor da IA**, dependendo da aplicação e do contexto de uso. Abaixo destacamos algumas das principais técnicas que avançaram em pesquisa e uso, detalhando uma visão geral de cada técnica, principais usos e melhores práticas, limitações e compensações aplicáveis.

- **O Conselho Digital acredita que uma combinação desses métodos será a maneira mais eficaz de validar e autenticar conteúdo gerado por IA**. No entanto, devido à natureza emergente dessas técnicas, bem como à natureza dinâmica do conteúdo gerado por aplicações de IA, é improvável que uma única solução, seja ela técnica ou manual, envolvendo ou não intervenção humana, abarque completamente todo tipo de conteúdo gerado por IA. **Isso significa que cumpre ao legislador não limitar ou querer implementar apenas uma solução, tendo em vista que esse é um campo em constante evolução e que a solução de hoje pode não ser mais válida amanhã.**

Autenticação Técnica

- A autenticação técnica é um **compilado de diferentes métodos que envolvem a incorporação de um sinal com informações em um trecho de texto ou imagem, também denominado de "rotulagem" ou "marca d'água".** Esse sinal pode ser aparente, como uma figura ou caixa de texto mostrando informações sobre sua origem, autoria, ou para identificar se foi ou não gerado por IA. Exemplos incluem uma marca d'água escrita "RASCUNHO" ou "Não para Liberação" em um documento ou a marca de uma titular em foto de sua propriedade (por exemplo, imagens da Getty.

- Além disso, **um rótulo pode ser invisível ou imperceptível.** Marcas d'água invisíveis ou imperceptíveis são constituídas de uma pequena quantidade de dados que são incorporados nos pixels de uma imagem. Às vezes, a rotulagem também é usada para descrever metadados, que são informações em um segmento especial de um arquivo que faz parte de uma imagem, mas não está incorporada em seus pixels. **Esses dados podem ser vinculados aos pixels por meio de um hash criptográfico e assinados digitalmente para fornecer garantia de que nenhuma manipulação ocorreu após sua criação.**

> **Saiba mais:** Um "hash criptográfico" é um algoritmo matemático que transforma dados de qualquer tamanho (texto, arquivos, imagens) em uma sequência única de caracteres de comprimento fixo, agindo como uma "impressão digital" digital.

- Técnicas de rotulagem são únicas no sentido de que, dependendo da ferramenta utilizada, podem criar um nível de transparência para uma ampla gama de usuários, permitindo que consumidores e usuários finais (além de desenvolvedores) saibam que uma peça de conteúdo (imagem ou texto) foi gerada usando IA. Alguns exemplos emergentes de marca d'água em conteúdo gerado por IA são descritos abaixo.

- **Empresas associadas do Conselho Digital estão experimentando técnicas de rotulagem de conteúdo produzido por aplicações de IA**. SynthID, uma solução desenvolvida pelo Google Deepmind e Google Cloud, rotula metadados de conteúdo gerado por IA criado com a plataforma de desenvolvedor do Google, Vertex. O ChatGPT da OpenAI

está considerando o uso de marcas d'água no texto que gera. A Meta está explorando Stable Signature, que é um método para marcar imagens criadas com IA de código aberto. **Frisa-se que tais soluções só podem ser criadas e implementadas devido a um ambiente regulatório que estimula a inovação e a experimentação de diferentes soluções.**

Principais Usos e Melhores Práticas

- **Rotulagem de modelo.** Essa técnica pode ser usada para rastrear e verificar as propriedades de um modelo de IA ou monitorar seu uso, incorporando informações nos parâmetros do modelo ou na sua própria estrutura. Esta técnica é utilizada especialmente para evitar o uso não autorizado, a distribuição ou a modificação de uma aplicação de IA específica.

- **Rotulagem do conjunto de dados.** Esse tipo de marcação pode ser usada para rastrear e verificar dados de *input* que serão usados no treinamento de aplicações de IA, fazendo a conexão entre o conjunto de dados em si e o seu proprietário original. Esse tipo de marca d'água é geralmente inserida em um subconjunto de um conjunto de dados durante um processo de treinamento e são então distribuídas para garantir que elas permaneçam mesmo quando haja a filtragem ou reorganização dos dados (e para que não possam ser facilmente removidas). Marcas d'água em conjuntos de dados ajudam organizações a rastrear a propriedade e a proveniência do conjunto de dados e podem ser usadas para verificação da legalidade de modelos de machine learning.

- **Esteganografia.** Essa é uma técnica que esconde uma marca d'água ou informações específicas dentro de um arquivo de mídia primário. Um dos tipos mais comuns de esteganografia envolve incorporar essas informações ocultas ou secretas no Bit Menos Significativo (*Least Significant Bit*) de um arquivo de mídia, o que é feito modificando-o ligeiramente ou adicionando informações adicionais a bytes de dados dentro de pixels em um arquivo de mídia. Isso garante que a marca d'água e as informações de verificação relevantes não sejam visíveis a olho nu, para também manter a usabilidade da imagem e/ou do texto em si. A esteganografia é frequentemente confundida com criptografia. Embora as técnicas sejam semelhantes, a criptografia envolve alterar ou embaralhar informações em um texto cifrado, que só pode ser desembaralhado com uma chave de decifração. Também envolve esconder ou ocultar informações de uma maneira que não seja

prontamente aparente, sem precisar descriptografá-la. Uma vez que a esteganografia requer uma aplicação cuidadosa envolvendo especificidade em nível da menor unidade de medida digital, essa abordagem requer uma colaboração robusta da indústria para ser bem-sucedida.

> **Saiba mais:** O Bit Menos Significativo (LSB - Least Significant Bit) é o bit localizado na posição mais à direita de um número binário. Ele possui o menor valor posicional, correspondendo a 2^0 (ou seja, 1), representando a unidade de menor peso. Ao contrário do bit mais significativo (MSB), o LSB tem o impacto mínimo no valor total do número.

- **Marcação forense invisível.** Essa técnica pode ser usada em conteúdo de vídeo e imagem através da codificação de uma única marca d'água em um arquivo de vídeo que atua como um identificador único do destinatário ou consumidor do conteúdo. Caso um conteúdo de vídeo seja gerado por IA, a marca no arquivo de vídeo permitiria aos proprietários do conteúdo rastrear sua disseminação. Como na esteganografia, a marcação forense invisível promove o uso do conteúdo, pois o conteúdo marcado é acompanhado de verificação. Esta é uma ferramenta muito popular utilizada na indústria do entretenimento para rastrear produções e vídeos protegidos por direitos autorais e para prevenir a pirataria ou distribuição não autorizada.

- **Marcação diferencial.** Por fim, essa técnica combina várias modalidades de marcação, direcionando-as a diferentes elementos (dados, metadados, conteúdo) de *input* ou *output* e garantindo que cada um deles tenha um sinal único. Um benefício do uso de marcação diferencial é que ela permite que sistemas de IA citem as fontes de texto originais com maior precisão através da própria marca d'água, facilitando a citação direta do conteúdo. Isso é alcançado através da incorporação de marcas d'água no *output* e nos metadados também.

Limitações e Trade-offs

- Marcas d'água visíveis em conteúdos de imagem podem ser facilmente cortadas ou removidas, diminuindo o valor de uma rotulagem perceptível. Além disso, marcas d'água podem ser "imitadas" em imagens, levando à disseminação de informações falsas que podem levar outros a pensar que são legítimas.

- Marcas d'água invisíveis em imagens, por outro lado, são mais fáceis de remover do que rótulos invisíveis em outros tipos de conteúdo. Para imagens, a marca d'água invisível geralmente é incorporada em bits de pixel da imagem, então se o formato do arquivo mudar ou for alterado por meio de uma captura de tela ou outra conversão de formato de imagem, a marca d'água invisível não é mais eficaz.

- Marcas d'água visíveis em conteúdos de vídeo são muito mais difíceis de remover. No entanto, esse tipo de marca d'água pode ser distrativo e tornar o seu consumo menos atraente, pois é visível para os consumidores. Esse tipo de rotulagem também pode desviar a atenção do conteúdo principal, particularmente em campanhas de publicidade onde o foco em determinado símbolo ou produto é essencial

- Marcas d'água invisíveis criadas através da adição de informações aos metadados de um conteúdo podem criar ruído: informações desnecessárias em um conjunto de dados que, em consequência, necessitam de maior espaço de armazenamento. Isso às vezes pode dificultar o download e acesso a certos tipos de conteúdo, particularmente com largura de banda de baixa latência ou um sistema que não possui boa quantidade de memória ou armazenamento.

- Marcas d'água podem validar elementos da cadeia de valor da IA na medida em que os humanos sabem que são reais. Marcas d'água podem ser colocadas em conteúdo e dados gerados por aplicações de IA, mas não impedem a disseminação de informações falsas se o conteúdo gerado por IA for falso ou se depender de dados ruins (mas ainda assim marcados).

Autenticação Humana

- **A autenticação humana requer envolvimento humano para verificar se o conteúdo foi gerado ou modificado por uma aplicação de IA.** Um exemplo claro disso é a exigência de validação humana de conteúdo que foi sinalizado como potencialmente gerado por IA. A autenticação humana pode ser usada em uma ampla gama de contextos, para validar *inputs* ou *outputs* de modelos.

- **A autenticação humana é útil e já faz parte da cadeia de valor da IA**: geralmente, limpar e rotular conjuntos de dados ou inserir credenciais em metadados é um processo altamente manual que requer supervisão humana. Embora a automação no contexto do mapeamento e

governança de dados esteja se tornando mais robusta, rotular conjuntos de dados, avaliar *outputs*, validar modelos e realizar aprendizado por reforço com feedback humano ainda são processos muito manuais e que exigem habilidade e envolvimento humano.

- A intervenção humana, incluindo o envolvimento de perspectivas humanas diversas no desenvolvimento de uma aplicação de IA, é um mecanismo amplamente aceito para prevenir problemas como viés e danos no mundo real, além de melhorar a acessibilidade e uso do produto.

Melhores Práticas

- Envolve grupos diversos de pessoas no processo de autenticação. Uma variedade de pessoas de diferentes origens e áreas de conhecimento irá melhorar a qualidade dos processos de autenticação humana. Algumas pessoas perceberão coisas que outras não perceberão.

- "Intervenção seletiva", pela qual a autenticação humana é usada apenas em partes essenciais do processo de autenticação de conteúdo gerado por IA, de modo a garantir que a intervenção humana seja direcionada e focada a fim de maximizar os recursos humanos.

Limitações

- A autenticação humana pode não ser confiável em casos onde a aplicação de IA é utilizada propositalmente para enganar humanos.

- Humanos possuem vieses inerentes e implícitos que afetarão qualquer tipo de processo que envolva sua intervenção direta. Desse modo, a intervenção humana não corrigirá a presença desses vieses e não haverá uma consistência de análise em toda uma organização.

- A autenticação e revisão humana é um processo inerentemente manual. A intervenção humana pode até mesmo retardar ou atrasar a autenticação de conteúdo gerado por IA. Apesar dessa dinâmica, a intervenção humana é geralmente considerada uma das maneiras mais eficazes de minimizar os riscos não intencionais relacionados à IA.

Outras Técnicas Relevantes

- Muitas técnicas estão surgindo para autenticar conteúdo gerado por aplicações de IA, mas nem todas alcançaram um nível de robustez ou disseminação comparável às técnicas discutidas acima. **As modalidades abaixo foram propostas para ajudar no processo de autenticação do conteúdo gerado por IA, mas ainda são necessárias mais pesquisas para determinar sua eficiência.**

1. Análise Comportamental

- **Uma técnica que procura padrões naturais de imperfeição, ritmo, cadência ou outros comportamentos típicos na geração de linguagem ou imagem** para diferenciar uma criação humana do *output* de uma criação de IA. Embora seja semelhante à autenticação humana e possa certamente ser usada como um tipo de autenticação humana, é mais restrita em seu escopo.
- **Automatizar um processo de análise comportamental exigiria volumes massivos de dados de treinamento** para que uma ferramenta de IA decidisse se determinado conteúdo foi criado artificialmente. A tecnologia ainda não amadureceu o suficiente para fazer essas verificações de forma independente.

2. Detecção Estatística

- Um método de identificar anomalias estatísticas na distribuição de pixels, frequências de fala, etc., que se correlacionam com conteúdo gerado artificialmente. **A detecção estatística é difícil de implementar porque requer uma linha de base de frequências consideradas "normais" que é arbitrária por natureza.** A detecção estatística tem uma taxa mais alta de falsos negativos, tornando-a menos ideal como um método de uso para fins de autenticação.

3. Tecnologia de Ledger Distribuído/Blockchain

- **Fornece armazenamento e verificação de dados à prova de adulteração, tornando difícil alterá-los uma vez que são registrados no blockchain.** Como essa técnica aproveita a tecnologia blockchain, isso exige que as pessoas a compreendam totalmente e tenham acesso à tecnologia, tornando-a uma ferramenta menos eficaz.

- Além dessa barreira de entrada, o mesmo tipo de rastreamento de metadados pode ser feito por meio de meios tecnológicos que não exigem blockchain.
- Por último, **essa técnica também não protege contra vieses nos dados** e, devido à natureza anônima da tecnologia blockchain, a identidade de quem altera o registro pode ser difícil de rastrear.

> **Saiba mais:** Blockchain é uma tecnologia de registro digital que funciona como uma cadeia de blocos. Imagine um "livro-razão" gigante e compartilhado, onde cada transação ou informação é gravada permanentemente, sem que ninguém possa apagá-la ou alterá-la.

Conclusão e Recomendações

- A Autenticação de conteúdo gerado por IA é um campo emergente — propício para inovação e para o desenvolvimento e adoção de melhores práticas.

- O que se buscou mostrar aqui é que todo método de autenticação de conteúdo gerado por IA possui variações e não conseguirá garantir 100% de eficácia.

- Deve-se, nesse sentido, estimular a inovação, a criação e o fornecimento de tecnologias cada vez melhores. **A solução não passa por impor medidas fechadas de verificação de conteúdo, mas sim por empoderar criadores e plataformas a usarem as mais variadas ferramentas** que, quando utilizadas em conjunto, aumentarão os níveis de transparência e confiabilidade do conteúdo compartilhado online.

Figura 1: Autenticação de IA ao Longo da Cadeia de Valor da IA
(Fonte: ITI, *Authenticating AI-Generated Content*, 2024, p. 12, traduzido)

CONTEÚDO (visto pelo olho humano)	Marca d'água visível	*Etiqueta de Getty Images* *Credencial de Conteúdo ao Passar o Mouse*
	Etiqueta de Conteúdo	
METADADOS	Marca d'água embutida	*Esteganografia* *Credencial de Conteúdo de Metadado*
	Auditoria de Metadados	
MODELO	Marca d'água embutida	*Sinais de desafio/resposta embutidos no modelo*
CONJUNTO DE DADOS DE TREINAMENTO	Marca d'água de Conjunto de Dados	*Ruído/sinais distribuídos nos dados*
APLICATIVO / API	Marca d'água embutida	*Credencial de Conteúdo de Metadados*
INFRAESTRUTURA DE COMPUTAÇÃO	SBOM (Lista de Materiais de Software) atestado	*Credencial de Conteúdo de Metadados com certificado de atestação*