Credenciais de personalidade: construindo confiança na era da IA

CT-GId

Shirlei Chaves

10/2024
https://www.rnp.br/ct-gid

RNP
ORGANIZAÇÃO SOCIAL DO MCTI

Educação, Pesquisa e Inovação em Rede

Agenda

1. Motivação
2. Credenciais de Personalidade
3. Direções Futuras
4. Conclusão

Motivação

Definição do Problema

- Historicamente, atores mal-intencionados usam identidades falsas para cometer atividades enganosas ou fraudulentas online;
- Até certo ponto, considerado um “custo aceitável” para proteção da privacidade e da manutenção do acesso irrestrito online.

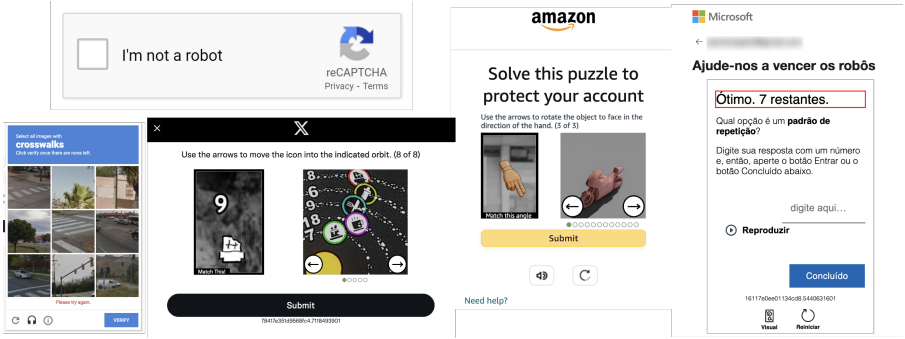
Definição do Problema

- Historicamente, atores mal-intencionados usam identidades falsas para cometer atividades enganosas ou fraudulentas online;
- Até certo ponto, considerado um “custo aceitável” para proteção da privacidade e da manutenção do acesso irrestrito online.

Sistemas de IA altamente capazes podem sobrecarregar a Internet com atividades enganosas, ameaçando tanto a privacidade quanto a integridade online.

Soluções Empregadas

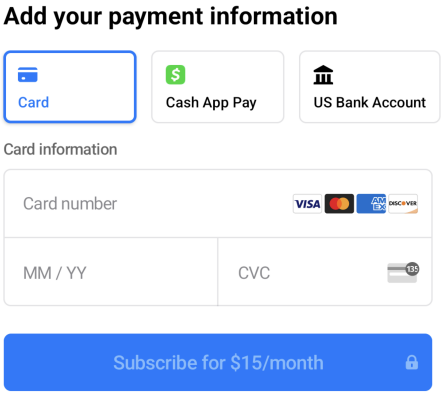
Filtros comportamentais



Não robusto para IAs altamente capazes.

Soluções Empregadas

Barreiras financeiras



Não é inclusivo.

Soluções Empregadas

Verificação baseada em aparência ou documentos

Figura: Selfie com documento[1]

Não robusto para IAs altamente capazes.
Não preserva a privacidade.

[1] TRE-RJ

Soluções Empregadas

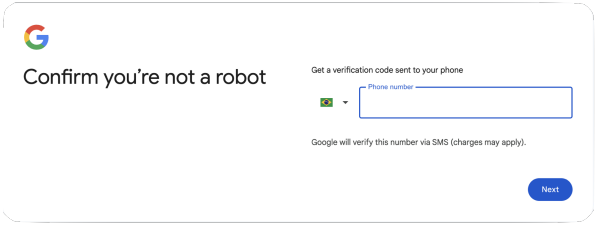
Verificação baseada em aparência ou documentos

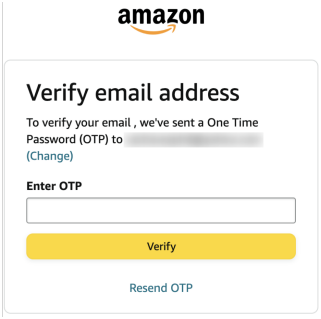
Figura: Eu Mandando Foto Segurando Documento Para Análise[2]

[2] Autor: Fã Clube Chaves E Sua Turma

Soluções Empregadas

Identificadores digitais e de hardware





Não escasso o suficiente.

Soluções Empregadas

Detecção de conteúdo de IA

- *Watermarking*;
  - Não efetivo para amplificação de conteúdo humano.
- *Fingerprinting*;
  - Geralmente hashing.
- Metadados de Proveniência;
  - Informações sobre origem e histórico de edição do conteúdo.
- Predição baseada em classificador.
  - Probabilidade do conteúdo ser gerado por IA.

Não robusto para IAs altamente capazes.

Ameaças crescentes da IA

Indistinguibilidade

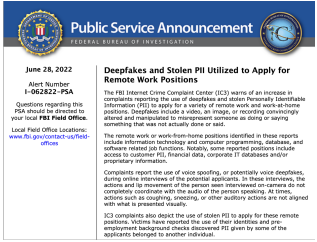


Figura: Canto superior esquerdo: Tom Cruise deepfake[3], centro: Prefeito Kyiv deepfake[4], direita: Anúncio do FBI sobre entrevista remota de emprego usando deepfake[5], canto inferior esquerdo: golpe financeiro usando deepfake [6]

[3] Tom cruise tik tok deepfake
[4] Chamada de vídeo deepfake com prefeito de Kyiv
[5] Anúncio FBI uso de deepfake em entrevista emprego remoto
[6] Golpe financeiro usando chamada de vídeo com deepfake

Ameaças crescentes da IA

Escalabilidade

- Ferramentas de IA cada vez mais acessíveis e baratas;
- Disponibilidade de modelos de IA de peso aberto (*open-weight*)[7].

[7] Normalmente são liberados os pesos e o código de inferência do modelo [1]. Todos os componentes deveriam ser liberados para ser considerado *open-source* [2]

Credenciais de Personalidade

Credenciais de Personalidade

Personhood Credentials - PHCs

- Um tipo de Credencial Verificável (p. ex.: VCs da W3C);
- Limitado à uma PHC por pessoa por emissor;
- Pseudônimo específico para cada serviço e não rastreável;
- Divulgação mínima de dados.

Prova que um usuário é uma pessoa real e não IA ou bot, sem revelar sua identidade pessoal.

Credenciais de Personalidade

Premissas Fundamentais

- Incapacidade de IA em imitar pessoas offline;
- Resistência a sistemas criptográficos estado da arte.

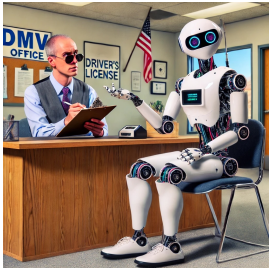


Figura: Fonte: DALL·E, gerada via OpenAI's GPT-4 model (2024).

Credenciais de Personalidade

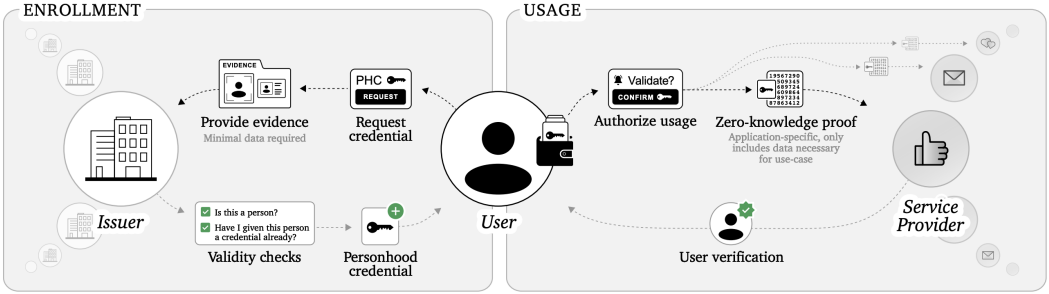


Figura: Obtenção e uso de uma PHC [1].

Credenciais de Personalidade

Potenciais emissores e sistemas

- Várias organizações podem ser emissoras - governamentais ou não;

Não se advoga por nenhuma implementação específica de um sistema de emissão de PHC, alinhando-se à iniciativas em andamento que buscam sistemas de identidade e autenticação com divulgação mínima de dados, como as credenciais verificáveis da W3C[a] e a carteira de identidade digital europeia[b],

[a] Verifiable Credentials Data Model v2.0
[b] EUID Wallet

Credenciais de Personalidade

Objetivos de um sistema de emissão de PHC

- Reduzir a escala de atividades falsas;
- Proteger a privacidade e as liberdades civis das pessoas.

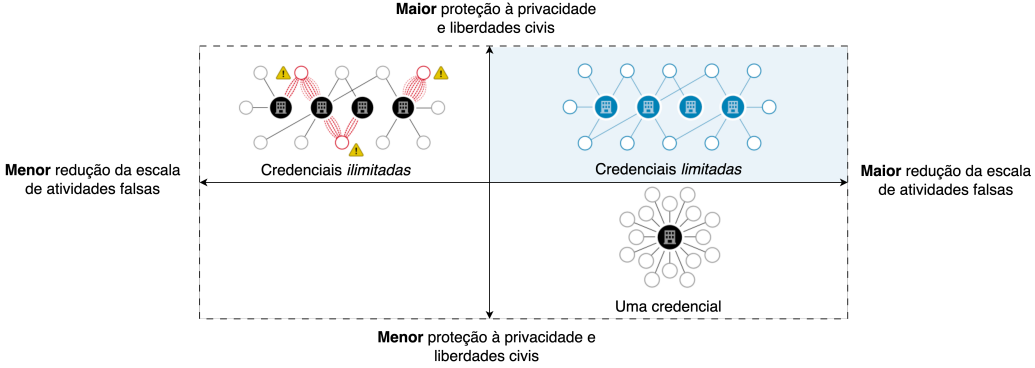


Figura: *Trade-offs* de projeto do ecossistema, adaptado de Addler et al. [1]

Credenciais de Personalidade

Potenciais benefícios

- Reduzir impacto de contas falsas (*sock puppets*);
- Mitigar ataques de bots;
- Permitir delegação verificada para agentes de IA.

Figura: *Sock Puppet by Alexbrn - Own work, CC BY-SA 4.0*[a]

[a] Sock Puppet Account - Wikipedia

Credenciais de Personalidade

Potenciais desafios

- Acesso equitativo;
- Liberdade de expressão;
- Limitações de poder e governança;
- Robustez contra ataques e erros.

Direções Futuras

Possíveis próximos passos

- Adaptação dos sistemas digitais existentes
  - Reexaminar os padrões de verificação e autenticação de identidade remota;
  - Estudar o impacto e a prevalência de contas enganosas nas principais plataformas de comunicação;
  - Estabelecer normas e padrões para regulamentar o uso de IA autônoma (*agentic users*) na internet.
- Priorizar a criação de credenciais de personalidade
  - Investir no desenvolvimento e criação de pilotos;
  - Incentivar a adoção de PHC.

Conclusão

Resumindo

- A IA está tornando difícil distinguir humanos;
- PHC: proposta para verificar quem é humano online sem comprometer a privacidade;
- Não se advoga por uma tecnologia ou implementação em particular, mas a urgência dos problemas que essas credenciais podem ajudar a resolver.

Referências

1 Adler, S., Hitzig, Z., Jain, S., Brewer, C., Chang, W., DiResta, R., Lazzarin, E., McGregor, S., Seltzer, W., Siddarth, D. and Soliman, N., 2024. Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online. Disponível em: https://arxiv.org/pdf/2408.07892.

2 Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., ... & Gupta, A. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Disponível em: https://arxiv.org/pdf/2311.09227.

Discussão

Contribuições? Dúvidas?

Muito obrigada!

ct-gid@listas.rnp.br
https://listas.rnp.br/mailman/listinfo/ct-gid