ybzbjyzsbozwnubxvzlhjyzsbo
bjnubxybzzwzyzswvbozzlhjyz
xvzlhjyzsbwnzsuo **ensaios sobre**
ybzlbjyzsb **tecnologia, privacidade**
hjyzsbo **e os limites do poder digital**
ybzblhjyzzwnubxvzjysbosboz
zybzbyzsozwnubxvzlhybsbzo

Organização
Thobias Prado Moura
Raquel Saraiva
Mariana Canto
Pedro Silva Neto
Luana Batista

Instituto de Pesquisa em Direito & Tecnologia do Recife

Internet Society Capítulo Brasil

Ensaios sobre tecnologia, privacidade e os limites do poder digital

Ensaios sobre tecnologia, privacidade e os limites do poder digital

Organização

Thobias Prado Moura
Raquel Saraiva
Mariana Canto
Pedro Silva Neto
Luana Batista

Autores

Bruno J. M. Marcolini
Giuseppe Grando
Jéferson Campos Nobre
Laura Soares
Leandro Bertholdo
Nathalia Sautchuk Patricio
Rafael Francisco França
Rodolfo Silva Avelino
Roberta Robert
Thobias Prado Moura

Instituto de Pesquisa em Direito & Tecnologia do Recife

Internet Society Capítulo Brasil

Sumário

Ensaios sobre tecnologia, privacidade e os limites do poder digital

Prefácio

Por Luiza Correa de Magalhães Dutra

Quem somos, quem nos tornamos e como nos compreendemos é, em grande medida, mediado pelas infraestruturas digitais que organizam nossas interações. Identidades são continuamente construídas e reconstruídas nesse ambiente, ao mesmo tempo em que dinâmicas sociais mais amplas, como disputas políticas, conflitos e formas de organização coletiva, passam a ser atravessadas por lógicas tecnológicas que reconfiguram o próprio modo como percebemos o mundo.

Nesse contexto, temas antes periféricos assumem centralidade no debate público: a influência das plataformas digitais nos processos decisórios, os impactos do uso crescente de inteligência artificial e a ampliação - sob novas e antigas formas - das capacidades de vigilância. As disputas em torno da informação, do controle e da observação tornam-se cada vez mais intensas, exigindo atenção crítica sobre quem detém esses poderes e como eles são exercidos. É nesse cenário que a criptografia forte se afirma como elemento central, não apenas como ferramenta técnica, mas como condição para a proteção de direitos e a preservação de espaços de autonomia.

Partindo dessas inquietações, o livro Ensaios sobre tecnologia, privacidade e os limites do poder digital propõe refletir, tensionar e aprofundar a compreensão sobre o que está em jogo quando se discutem vigilância e proteção criptográfica. Os capítulos que compõem a

obra percorrem temas diversos - da proteção de dados nas relações de trabalho à regulação da criptografia no Brasil e aos desafios do reconhecimento facial, articulando, em diferentes níveis, as relações entre tecnologia, poder e governança.

O capítulo I, de Bruno J. M. Marcolini e Giuseppe Grando, intitulado Entre a privacidade e a vigilância: a proteção de dados pessoais nas relações de trabalho, examina, com rigor crítico, as transformações do monitoramento no ambiente laboral à luz da expansão das tecnologias digitais e da consolidação do teletrabalho. Ao deslocar o controle patronal de práticas pontuais para sistemas contínuos, automatizados e frequentemente imperceptíveis, a vigilância laboral passa a incidir de modo direto sobre a esfera de direitos fundamentais do trabalhador, em especial a privacidade e a dignidade. Partindo da premissa de que dados de desempenho, comportamento e interação configuram dados pessoais, os autores sustentam que tais práticas devem ser integralmente submetidas ao regime da Lei Geral de Proteção de Dados, discutindo os limites e as possibilidades das bases legais aplicáveis no contexto laboral, em um cenário marcado por assimetrias estruturais de poder.

Ao longo da análise, o capítulo demonstra que a legitimidade do monitoramento depende de sua estrita conformidade com princípios como finalidade, necessidade, proporcionalidade e transparência, bem como da rejeição de práticas discriminatórias e opacas. Ao articular fundamentos teóricos, exame normativo e casos concretos, o trabalho demonstra que a dataficação do trabalho exige um modelo de governança orientado por direitos, capaz de reequilibrar a relação entre eficiência empresarial e proteção do trabalhador. Nesse sentido, a LGPD é apresentada não como obstáculo, mas como instrumento indispensável para a construção de práticas de gestão responsáveis, que reconheçam os limites ético-jurídicos da vigilância e promovam relações laborais mais justas e transparentes.

O capítulo II, por sua vez, volta-se ao debate entre direito e biometria. Escrito por Rafael Francisco França e intitulado O sujeito em frente às câmeras: desafios do reconhecimento facial diante da diversidade humana, o capítulo propõe uma inflexão analítica ao deslocar o foco dos sistemas técnicos para os próprios titulares dos dados biométricos. Em vez de privilegiar exclusivamente a arquitetura algorítmica ou sua adequação normativa, o texto investiga como a diversidade humana desafia os pressupostos de universalidade, unicidade e estabilidade que sustentam tais tecnologias. A partir de uma abordagem jurídico-crítica e interdisciplinar, apresentou-se que a face, enquanto dado biométrico, é dinâmica e atravessada por fatores biológicos e sociais, o que compromete a pretensa neutralidade dos sistemas. Nesse contexto, grupos como crianças e idosos emergem como particularmente vulneráveis, uma vez que suas características faciais, em constante transformação, tensionam a eficácia e a confiabilidade dos mecanismos de reconhecimento.

Ao longo da análise, o capítulo demonstra que as falhas de identificação não constituem meras limitações técnicas contingentes, mas refletem um modelo que opera por padronização e, ao fazê-lo, marginaliza aquilo que escapa à norma. Casos como o de gêmeos idênticos, bem como os efeitos do envelhecimento e do desenvolvimento craniofacial, revelam limites estruturais da tecnologia e seus potenciais impactos discriminatórios. Diante disso, sustenta-se que a avaliação jurídica da legitimidade do reconhecimento facial exige uma abordagem mais cautelosa e centrada na proteção da dignidade humana e da não discriminação. Ao final, o capítulo contribui para o debate ao apontar que os riscos de erro e exclusão são inerentes ao modelo atual, reforçando a necessidade de respostas regulatórias mais restritivas e de um desenho tecnológico comprometido com a inclusão e a justiça.

No capítulo III, Thobias Prado Moura e Nathalia Sautchuk Patricio,

em Trusted Execution Environments em ecossistemas de criptografia ponta a ponta: análise de ameaças e desafios regulatórios no Brasil, examinam as implicações técnico-jurídicas da adoção de Ambientes de Execução Confiável (TEEs) no contexto da computação em nuvem, demonstrando como essa arquitetura reconfigura os fundamentos da segurança da informação ao estender a proteção criptográfica ao momento do processamento dos dados. Ao articular conceitos da engenharia de segurança com categorias do direito da proteção de dados e do processo penal, o texto demonstra que a integração entre TEEs e criptografia de ponta a ponta desloca o eixo da confiança: da posse e controle das chaves criptográficas para a governança da chamada Raiz de Confiança, responsável por garantir a integridade do ambiente e das atualizações de software. Nesse cenário, o dado deixa de ser compreendido apenas como informação cifrada em trânsito ou em repouso, passando a existir como representações intermediárias em memória isolada, o que desafia sua qualificação jurídica e o posiciona em uma zona de incerteza entre pseudonimização e anonimização.

Ao longo da análise, os autores demonstram que, embora os TEEs representem um avanço relevante para a segurança da economia digital — ao mitigar riscos associados à computação terceirizada e reforçar a confidencialidade em ambientes distribuídos, também introduzem novas vulnerabilidades sistêmicas relacionadas à governança da infraestrutura. A dependência de ciclos de atualização e mecanismos de atestação cria pontos sensíveis suscetíveis a manipulações, inclusive por meio de intervenções estatais. No contexto brasileiro, marcado por ambiguidades regulatórias e tensões históricas em torno da criptografia, tais características podem favorecer cenários de desvio de finalidade, nos quais tecnologias concebidas para proteção da privacidade são reconfiguradas como instrumentos de vigilância. Nesse sentido, o capítulo sustenta a necessidade de uma abordagem crítica que reconheça tanto os benefícios quanto os riscos dessa arquitetura, enfatizando a importância de salvaguardas institucionais, transparência e controle social para assegurar que a inovação tecnológica permaneça alinhada à proteção de direitos fundamentais.

O capítulo IV, intitulado Análise de Segurança da Computação Confidencial e escrito por Jéferson Campos Nobre, Laura Soares, Leandro Bertholdo e Roberta Robert, apresenta um panorama abrangente da Computação Confidencial como resposta aos desafios contemporâneos de proteção de dados em ambientes de nuvem e sistemas baseados em inteligência artificial. Ao estender as garantias de segurança para o momento do processamento - tradicionalmente o ponto mais vulnerável do ciclo de vida dos dados, esse paradigma reconfigura os mecanismos clássicos de proteção, ancorando-se em arquiteturas de isolamento em hardware, cadeias de confiança e processos de atestação remota. O texto examina as principais tecnologias envolvidas, com destaque para os Ambientes de Execução Confiável (TEEs), e discute como tais soluções dependem de uma infraestrutura complexa e de pressupostos técnicos que condicionam suas promessas de confidencialidade e integridade. Ao mesmo tempo, explora a articulação entre Computação Confidencial, Criptografia de Ponta a Ponta e técnicas de Comunicação Anônima, destacando seu caráter complementar na proteção não apenas do conteúdo, mas também dos metadados.

Ao longo da análise, o capítulo demonstra que, embora represente um avanço significativo em termos de cibersegurança, a Computação Confidencial não elimina riscos, mas os reconfigura. A dependência de cadeias de confiança baseadas em hardware, a crescente complexidade arquitetural e a fragmentação do ecossistema introduzem novos vetores de vulnerabilidade e desafios de implementação. O estudo de caso do Private Processing, desenvolvido pela Meta no contexto do WhatsApp, ilustra de forma concreta os benefícios e limitações desse modelo, exi-

bindo tanto seu potencial para ampliar a privacidade em aplicações com inteligência artificial quanto os entraves técnicos, operacionais e de governança que ainda persistem. Nesse sentido, o capítulo contribui para uma compreensão crítica do tema ao situar a Computação Confidencial não como solução definitiva, mas como parte de um arranjo mais amplo e ainda em construção, que exige avaliação contínua, transparência e aprimoramento técnico e institucional.

Por fim, o capítulo V, último capítulo do livro, intitulado Criptografia, confiança e governança: uma análise sociotécnica da moderação de conteúdo em sistemas de comunicação fim a fim, e escrito por Rodolfo Silva Avelino, analisa a criptografia de ponta a ponta (E2EE) para além de sua dimensão técnica, situando-a como elemento central na reconfiguração das relações de poder, confiança e responsabilidade nas infraestruturas comunicacionais contemporâneas. A partir de uma abordagem sociotécnica, o texto demonstra que a E2EE não apenas fortalece a proteção da privacidade e da segurança das comunicações, mas também desloca a autoridade observacional tradicionalmente concentrada em plataformas e Estados, ao excluir intermediários do acesso ao conteúdo. Ao recuperar a evolução dos protocolos criptográficos e seus fundamentos técnicos, o capítulo demonstra que a chamada "confiança algorítmica" altera profundamente as formas de mediação social, reduzindo a dependência de instituições centralizadas e inaugurando arranjos mais distribuídos - ainda que marcados por novos desafios de governança.

Ao longo da análise, o capítulo sustenta que a moderação de conteúdo em ambientes protegidos por E2EE não desaparece, mas se transforma em um processo fragmentado, reativo e distribuído entre múltiplos atores. Propostas como a varredura no lado do cliente são examinadas como tentativas de reintroduzir observabilidade, ao custo de redefinir os limites entre vigilância, privacidade e controle. Nesse contexto, a fragmentação da responsabilidade emerge como característica estruturante da governança digital contemporânea, exibindo que os dilemas entre privacidade, segurança pública e responsabilização decorrem de escolhas arquiteturais e institucionais, e não de limitações técnicas da criptografia. Ao explorar alternativas como Software Livre e plataformas federadas, o capítulo reforça que tais tensões não admitem soluções puramente técnicas, exigindo, ao contrário, arranjos democráticos, transparentes e continuamente negociados.

Os capítulos que compõem esta obra demonstram, de forma consistente, que falar em criptografia está longe de ser apenas uma discussão técnica. Trata-se, antes, de um campo atravessado por disputas sobre poder, governança, direitos fundamentais e modelos de organização social. Das tensões em torno da moderação de conteúdo em ambientes criptografados às reconfigurações introduzidas pela computação confidencial e pelos ambientes de execução confiável, evidencia-se que cada escolha arquitetural carrega implicações políticas profundas. A criptografia emerge, assim, não apenas como ferramenta de proteção, mas como infraestrutura de mediação de relações sociais, capaz de redistribuir autoridade, redefinir responsabilidades e tensionar os limites entre privacidade, segurança e controle.

Longe de admitir respostas simples ou soluções universais, o debate exige um olhar atento, interdisciplinar e sensível às camadas técnicas e institucionais que o constituem. As análises aqui reunidas revelam que os dilemas contemporâneos não decorrem de falhas da tecnologia em si, mas das formas como ela é projetada, implementada e governada. Nesse cenário, não há neutralidade possível: toda decisão técnica é também uma decisão política. A questão que se impõe, portanto, não é apenas como regular ou utilizar a criptografia, mas de que forma nos colocaremos diante dessas disputas: como agentes de sua contenção, de sua ampliação ou de sua transformação?

Introdução

Quem controla a informação?

Vivemos sendo observados. A afirmação não é retórica, mas uma descrição precisa da condição digital contemporânea. Câmeras de reconhecimento facial identificam rostos em espaços públicos. Plataformas de mensagens processam bilhões de conversas em servidores que não controlamos. Empregadores monitoram cada tecla, cada clique, cada segundo de inatividade de seus funcionários. Governos disputam o poder de acessar comunicações privadas em nome da segurança pública. Algoritmos decidem o que vemos, o que compramos e, cada vez mais, como somos avaliados.

Mas nem toda observação é igual. Parte desses atores busca vigiar, no sentido mais clássico do termo: exercer controle, antecipar comportamentos e, em última instância, intervir. Outros, operam sob uma lógica distinta: coletam, agregam e analisam dados para compreender hábitos, preferências e padrões de consumo, transformando informação em valor econômico. A diferença pode ser sutil, mas é fundamental. Entre vigilância e extração de dados há uma zona de sobreposição, na qual conhecer pode facilmente se converter em controlar.

A pergunta fundamental deste livro não é técnica, é política: quem tem o poder de observar, registrar e agir sobre as informações

que produzimos? E, igualmente importante: quais ferramentas existem para limitar esse poder?

A criptografia é uma dessas ferramentas, talvez a mais poderosa, porque se baseia em matemática e não em confiança. Mas não é a única. A proteção de dados pessoais, os marcos regulatórios como a LGPD, a governança das plataformas, o design das arquiteturas de software e até as escolhas sobre quais tecnologias biométricas são aceitáveis, tudo isso compõe o campo de batalha onde se decide o equilíbrio entre vigilância e privacidade, entre controle e liberdade.

> A informação é poder.
> Mas como todo poder,
> há aqueles que o querem para si.

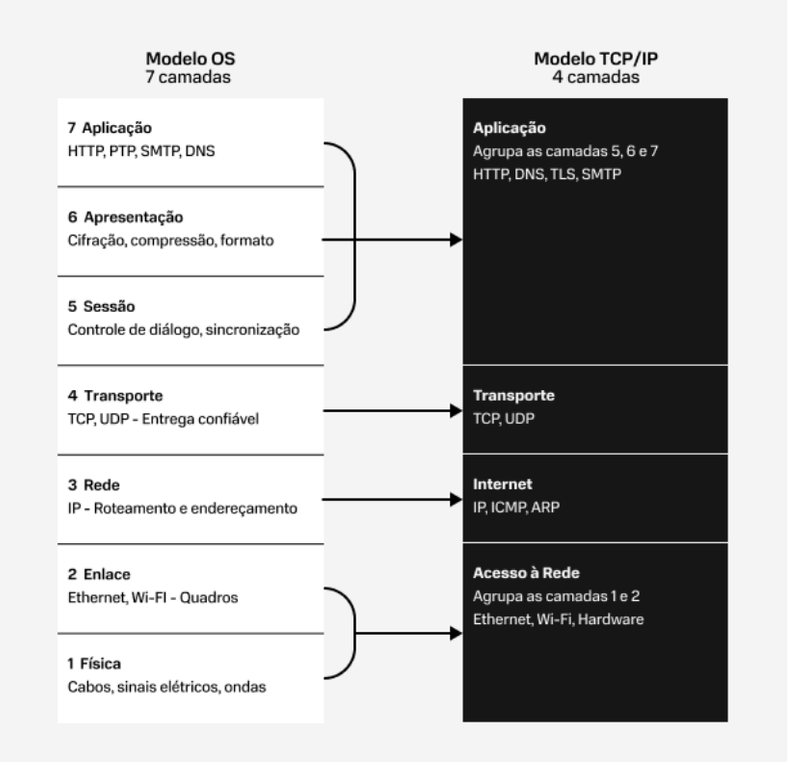
Este livro reúne ensaios que examinam esse equilíbrio a partir de cinco perspectivas distintas mas interconectadas: a discriminação algorítmica no reconhecimento facial, os desafios regulatórios dos ambientes de execução confiável, a criptografia como tecnologia política de redistribuição de poder, a vigilância no ambiente de trabalho e a segurança da computação confidencial. Este capítulo introdutório oferece os fundamentos conceituais necessários para navegar todos eles.

Como a informação viaja: camadas e a arquitetura da rede

Para compreender a vigilância digital, e por que a criptografia é tão necessária, é preciso primeiro entender como a informação viaja. A Internet não é um tubo único, mas sim uma teia de intermediários por onde os dados saltam até chegar ao destino. E cada intermediário é, em tese, um ponto de observação.

A analogia mais intuitiva é o sistema postal. Você escreve o conteúdo (aplicação); coloca a carta num envelope lacrado (segurança); escreve o endereço (rede); e o carteiro usa ruas, aviões e caminhões para transportar (física). Cada ator intermediário faz sua parte sem precisar saber o conteúdo da carta. Na Internet, o princípio é o mesmo: sem proteção, qualquer intermediário pelo qual os dados passam pode, em tese, ler o que está sendo transmitido, assim como aconteceria com uma carta enviada sem envelope.

Figura 1 Modelo OSI e modelo TCP/IP



O modelo OSI. Formalizado em 1983, o Modelo de Referência OSI organizou o funcionamento da rede em sete camadas com funções bem definidas. Da base ao topo, temos: a camada Física (1), que trata dos sinais elétricos e ondas; a de Enlace (2), que organiza bits em quadros; a de Rede (3), responsável pelo endereçamento e roteamento via IP[1]; a de Transporte (4), que garante a entrega confiável dos dados por meio do TCP[2]; a de Sessão (5), que gerencia os diálogos entre sistemas; a de Apresentação (6), que cuida da cifragem e da formatação; e, por fim, a de Aplicação (7), onde rodam os programas que usamos no dia a dia, como navegadores, clientes de e-mail e aplicativos de mensagens.

O modelo TCP/IP. Na prática, a Internet condensa essas funções em quatro camadas: Acesso à Rede (camadas 1-2 do OSI), Internet (3), Transporte (4) e Aplicação (5-7). Essa simplificação pragmática é uma das razões que tornaram o TCP/IP o padrão universal.

Por que as camadas importam para a segurança da Internet? Porque a proteção criptográfica atua em camadas específicas, e quem vigia também. A criptografia de transporte, como o TLS, protege os dados entre as camadas de transporte e aplicação. A E2EE (criptografia ponta-a-ponta - end-to-end encryption) opera diretamente na camada de aplicação, tornando o conteúdo opaco até para a própria plataforma que entrega a mensagem. Ainda assim, os metadados, como quem fala com quem, quando e de onde, frequentemente permanecem visíveis nas camadas inferiores. Um vigilante determinado não precisa ler o conteúdo da carta para saber que ela foi enviada.

Cada camada, uma responsabilidade. A Figura 1 simplifica a arquitetura em quatro funções essenciais. A Camada de Aplicação é onde rodam os programas que o usuário utiliza diretamente, como o navegador ou o aplicativo de mensagens; é ela que gera e consome os dados. A Camada de Transporte garante a entrega confiável dos dados de uma ponta à outra da comunicação; é nela que protocolos de segurança como o TLS operam, cifrando os dados antes de serem transmitidos. A Camada de Internet cuida do endereçamento e do roteamento: determina por qual caminho os pacotes vão viajar até o destino, usando o protocolo IP. A Camada de Acesso à Rede, por fim, corresponde ao meio físico de transmissão: cabos de fibra óptica, ondas de Wi-Fi, sinais elétricos, tudo aquilo que efetivamente carrega os bits de um ponto a outro.

A grande inovação dessa arquitetura é que as camadas são independentes. Você pode trocar a camada física (de Wi-Fi para cabo Ethernet) sem alterar nada na camada de aplicação. Pode adicionar criptografia na camada de segurança sem mudar como os pacotes são roteados. Essa modularidade é o que permite que a Internet evolua incrementalmente, sem precisar ser reconstruída do zero a cada avanço tecnológico.

Comutação de pacotes: como os dados realmente viajam

Se a divisão em camadas explica quem faz o quê, a comutação de pacotes explica como a informação se move. E o mecanismo é contra-intuitivo: sua mensagem não viaja inteira por um único caminho, como uma carta pelo correio. Em vez disso, ela é fragmentada em pequenos pedaços chamados pacotes, cada um com endereço próprio, que viajam independentemente por rotas diferentes até o destino, onde são remontados na ordem correta.

1. Protocolo de Internet, o sistema de endereçamento que identifica cada dispositivo na rede, explicado adiante).
2. Protocolo de Controle de Transmissão, também explicado a seguir.

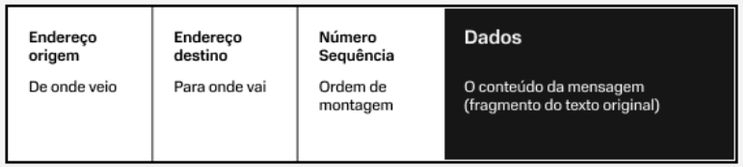


Figura 2 Estrutura de um pacote de dados

Cada pacote é uma unidade autônoma de informação. Como se vê na Figura 2, cada pacote contém não apenas um fragmento dos dados originais, mas também um cabeçalho com metadados essenciais: o endereço de origem (de onde veio), o endereço de destino (para onde vai), e um número de sequência (para que o destinatário saiba em que ordem remontar os fragmentos). É como enviar um livro pelo correio dividindo-o em capítulos, cada um em um envelope separado e numerado.

O papel dos roteadores. Os roteadores são os nós intermediários da rede. Quando um roteador recebe um pacote, ele lê o endereço de destino no cabeçalho e decide para qual roteador vizinho encaminhá-lo, buscando o caminho mais eficiente. Esse processo se repete a cada salto até que o pacote chegue ao destino. Dessa forma, o roteador não precisa conhecer o caminho completo, só precisa saber o próximo passo. É como pedir direções em uma cidade desconhecida: cada pessoa indica apenas a próxima esquina, não o trajeto inteiro.

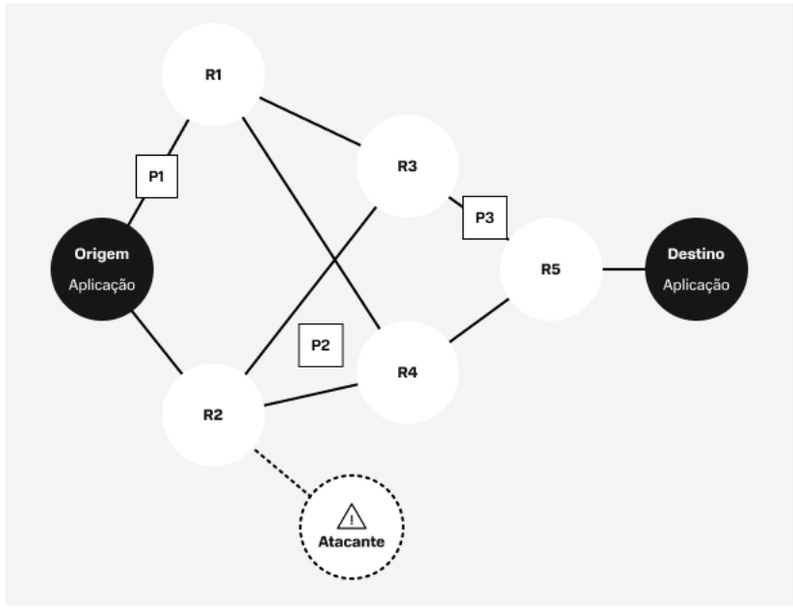


Figura 3 Computação de pacotes: a viagem dos dados pela rede

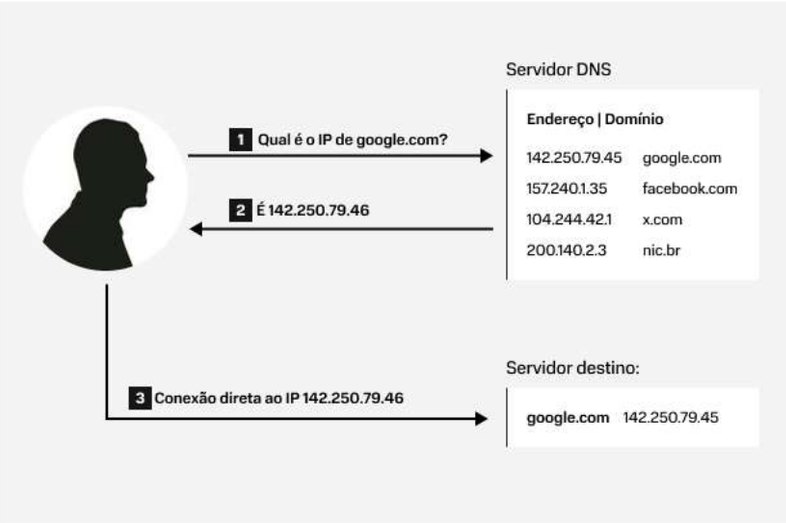
Resiliência e vulnerabilidade. A comutação de pacotes foi projetada originalmente para resiliência: se um nó da rede cai, os pacotes encontram rotas alternativas automaticamente. Essa descentralização, nascida de preocupações militares durante a Guerra Fria, tornou a Internet extraordinariamente robusta. Mas traz uma vulnerabilidade inerente: como os pacotes passam por roteadores que não controlamos, qualquer um deles pode, em princípio, ler, copiar, alterar ou destruir os dados em trânsito. Deve-se lembrar, ainda, que a Internet nasceu como uma rede que conectava pares que se conheciam, como instituições governamentais e universidades, e que tinham confiança mútua. Embora resiliência e simplicidade tenham sido princípios essenciais em sua concepção, não havia a preocupação de inserção de mecanismos de segurança intrínseca na rede.

Endereçamento e DNS: como os computadores se encontram

Antes de enviar pacotes, é preciso saber para onde enviá-los. Cada dispositivo conectado à Internet recebe um endereço numérico único: o endereço IP. No formato IPv4, esse endereço é uma sequência de quatro números (como 142.250.79.46). O problema é que humanos não memorizam números facilmente. Ninguém digita "142.250.79.46" no navegador — digitamos "google.com".

Para fazer essa tradução, existe o Sistema de Nomes de Domínio (DNS), uma espécie de "agenda telefônica" da Internet, como se vê na Figura 4. Quando você digita um endereço textual, seu computador consulta um servidor DNS que retorna o endereço IP correspondente. Só então a conexão é estabelecida.

Figura 4 Resolução de DNS: como os nomes viram números



IPv4 vs. IPv6. O IPv4 oferece cerca de 4 bilhões de endereços, um número que parecia imenso nos anos 1980, mas que se revelou insuficiente para um mundo onde geladeiras, relógios e câmeras também se conectam. A solução já implementada é o IPv6, que amplia o espaço de endereçamento para 340 undecilhões de endereços, o suficiente para atribuir um IP a cada átomo na superfície da Terra.

DNS e segurança. O sistema DNS, por padrão, opera sem criptografia: as consultas viajam em texto puro. Isso significa que um observador pode saber quais sites você visita mesmo sem acessar o conteúdo das páginas.

Protocolos como DNS-over-HTTPS (DoH) e DNS-over-TLS (DoT) foram criados para cifrar essas consultas, reforçando que a criptografia precisa ser aplicada em múltiplas camadas da comunicação para oferecer proteção real.

Segurança da informação e suas propriedades

O problema do texto puro. Mensagens sem proteção podem ser lidas, alteradas ou destruídas por qualquer nó. Um participante malicioso pode comprometer três propriedades fundamentais: a mensagem deixa de ser secreta (perda de confidencialidade); não há garantia de integridade (perda de integridade); e se deletada, há perda de disponibilidade.Essas três propriedades formam a tríade da segurança da informação.

Confidencialidade: a mensagem só deve ser legível pelo destinatário legítimo. Se um roteador (ou qualquer intermediário) lê o conteúdo sem autorização, a confidencialidade foi violada. Na camada da aplicação, isso equivale a interceptar e-mails, espionar conversas ou acessar prontuários médicos sem consentimento. A criptografia é a principal ferramenta para garantir confidencialidade: mesmo que o

pacote seja interceptado, seu conteúdo permanece ilegível.

Integridade: a garantia de que a mensagem não foi alterada no caminho. Se um intermediário modifica o conteúdo sem ser detectado, a integridade foi comprometida. Em contextos reais, isso pode significar alterar o valor de uma transferência bancária, falsificar uma prescrição médica ou manipular uma ordem judicial. Funções hash e assinaturas digitais são os mecanismos criptográficos que detectam qualquer alteração.

Disponibilidade: a informação precisa estar acessível quando necessária. Se um atacante destrói a mensagem, bloqueia o caminho ou sobrecarrega o servidor (como em ataques DDoS), a disponibilidade foi atacada. Redundância de servidores, distribuição geográfica e mecanismos de recuperação são as defesas típicas.

Esses três pilares não são independentes. Um ataque pode comprometer mais de uma propriedade simultaneamente: alterar uma mensagem viola integridade e confidencialidade (o atacante precisou ler para alterar); destruir dados viola disponibilidade e potencialmente integridade. A criptografia, sozinha, é mais eficaz para proteger confidencialidade e integridade. A disponibilidade exige medidas complementares de infraestrutura. A compreensão integrada dos três pilares é o que permite projetar sistemas de segurança verdadeiramente robustos.

Figura 5 Os três pilares da segurança da informação.

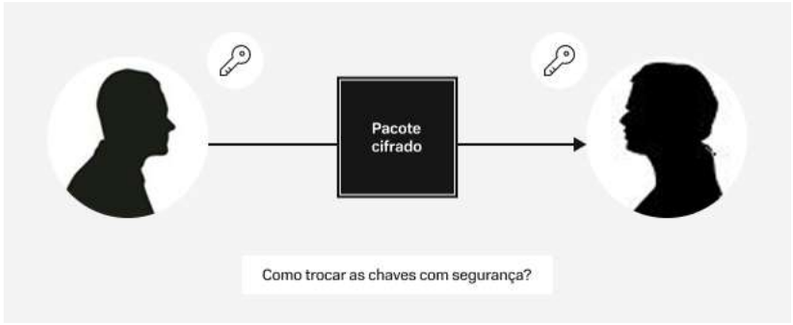
O que é criptografia?

Em definição ampla, criptografia é a ciência de transformar informações legíveis (texto puro) em algo incompreensível (texto cifrado) por meio de um algoritmo controlado por um parâmetro secreto chamado chave. Cifrar transforma texto puro em cifrado; decifrar restaura o original. Esses conceitos valem tanto para cifragem em papel quanto para algoritmos computacionais modernos.

Criptografia simétrica. Remetente e destinatário compartilham a mesma chave. O algoritmo AES é o padrão mundial e protege desde comunicações governamentais até Wi-Fi doméstico. Mas carrega um dilema fundamental: como entregar a chave com segurança antes de iniciar a comunicação? Se interceptada, toda proteção se desfaz.

Criptografia assimétrica (chave pública). Cada pessoa possui duas chaves matematicamente relacionadas: uma pública, que pode ser divulgada livremente, e uma privada, que permanece em segredo. O que uma chave cifra, somente a outra consegue decifrar. Para enviar uma mensagem segura a alguém, basta cifrá-la com a chave pública dessa pessoa; só ela, com sua chave privada, poderá lê-la. Algoritmos como

Figura 6 Criptografia Simétrica: o dilema da troca de chaves



RSA e Diffie-Hellman (dois dos primeiros e mais influentes algoritmos de criptografia assimétrica, criados na década de 1970) são a base dessa abordagem. Eles eliminaram o problema da troca de chaves e tornaram possíveis o comércio eletrônico e o HTTPS (o protocolo que protege as conexões web, identificado pelo cadeado na barra do navegador).

Criptografia ponta a ponta (E2EE). Garante que apenas os usuários diretamente envolvidos leiam o conteúdo, excluindo até os provedores de serviço. O protocolo mais adotado para isso é o Signal Protocol, que combina três propriedades: o sigilo futuro (forward secrecy, que garante que mesmo se uma chave for comprometida no futuro, as mensagens passadas continuam protegidas), a negação plausível (que impede provar quem escreveu o quê) e a autenticação robusta (que confirma a identidade dos interlocutores). É o estado da arte em privacidade de comunicações digitais e está por trás de aplicativos como WhatsApp e Signal.

Criptoanálise. A criptoanálise é o campo dedicado a tentar quebrar cifras. Seus pesquisadores analisam mensagens cifradas em busca de padrões, repetições ou fragilidades que permitam recuperar o conteúdo original sem a chave. A história da criptografia é uma corrida permanente entre quem cifra e quem ataca: cada algoritmo quebrado impulsiona o desenvolvimento de um novo algoritmo mais robusto e resistente.

Uma arte milenar: da Cifra de César à era digital

A criptografia acompanha a humanidade desde que existem conflitos e comércio. Das tábuas de argila aos algoritmos quânticos, a essência permanece: transformar o legível em ilegível.

Cifra de César (século I a.C.). Cada letra é deslocada por um número fixo. Simples, mas revolucionária: fundou o princípio de que um segredo pode ser protegido por procedimento sistemático controlado por um parâmetro.

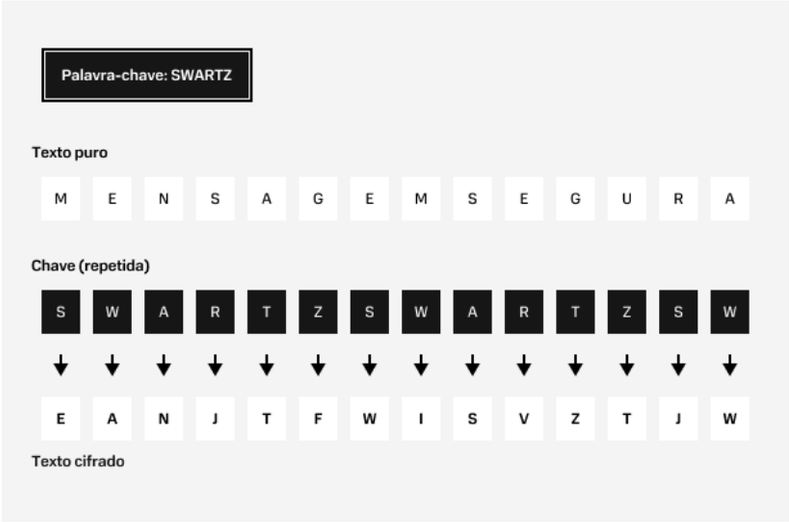


Figura 7 Cifra de Vigenère

Cifra de Vigenère (século XVI). Uma palavra-chave inteira é usada, em vez de um número. A cifra é polialfabética: a mesma letra produz resultados diferentes conforme a posição. Ela foi considerada indecifrável por 300 anos, até Charles Babbage identificar repetições no texto cifrado que revelavam o comprimento da chave.

Cifra de Autochave. A própria mensagem se torna parte da chave. Impede ataques por repetição periódica, embora palavras comuns da língua possam ser exploradas em mensagens longas.

Enigma & Máquinas (século XX). A Segunda Guerra elevou a criptografia a arma estratégica. A quebra da máquina Enigma, utilizada pela Alemanha na Segunda Guerra Mundial, por Turing inaugurou a era da Computação. Desde então, surgiram diversos algoritmos criptográficos, como AES, RSA e Signal Protocol, e agora a criptografia pós-quântica prepara-se para resistir a computadores quânticos.

A criptografia como tecnologia política

A criptografia ponta a ponta não é apenas uma escolha técnica: é também uma decisão política sobre quem tem acesso ao quê. Quando uma plataforma adota E2EE, ela abre mão da própria capacidade de ler as mensagens dos usuários. Isso significa que nem a empresa, nem o governo, nem qualquer terceiro consegue interceptar o conteúdo das conversas, mesmo com ordem judicial. Para jornalistas que protegem fontes, ativistas em países autoritários ou qualquer pessoa que precise de privacidade real, isso é uma garantia fundamental, já que ela retira do Estado e de grandes corporações a capacidade de vigilância sobre as comunicações privadas. A contrapartida é que se nem a plataforma vê o que está sendo enviado, ela também não consegue identificar e remover conteúdo ilegal, como imagens de abuso infantil ou mensagens que coordenam ataques terroristas. Essa tensão entre proteger a privacidade e permitir a responsabilização por crimes ficou conhecida como o "dilema da criptografia".

Propostas como a varredura no dispositivo do usuário (client-side scanning) tentam contornar esse dilema, ou seja, em vez de interceptar as mensagens durante o trânsito, o sistema analisaria o conteúdo antes de cifrá-lo, no próprio celular ou computador do usuário. Na prática, porém, isso cria uma abertura no sistema que pode ser explorada por governos autoritários para vigiar opositores políticos, minorias religiosas ou qualquer grupo considerado inconveniente. A E2EE deixa de ser ponta a ponta de verdade e passa a ter uma "porta dos fundos" embutida no dispositivo. Na ausência dessas soluções centralizadas, a moderação de conteúdo em ambientes criptografados não desaparece, mas assume um formato diferente dependente das denúncias dos próprios usuários, de filtros aplicados antes do envio (como bloqueio de links já conhecidos como maliciosos) e de ações sobre os metadados da comunicação, em vez de seu conteúdo.

Há uma diferença profunda entre confiar em uma instituição e confiar em uma equação matemática. Quando um banco diz que suas mensagens são seguras, você precisa acreditar na palavra da empresa. Quando um sistema usa E2EE com código aberto, qualquer pesquisador do mundo pode auditar o funcionamento e verificar se a proteção é real. A criptografia substitui a confiança nas pessoas pela confiança nos protocolos. Dessa forma, a segurança deixa de depender de promessas e passa a ser verificada por matemática. Isso fica evidente quando se observa casos como o Signal e as redes sociais federadas. O Signal usa um protocolo aberto, que qualquer especialista pode examinar e verificar; já as redes federadas são distribuídas por natureza, sem um único controlador que concentre poder sobre as comunicações. Nenhum desses modelos resolve o dilema da moderação de forma mágica, mas ambos mostram que há alternativas ao modelo centralizado dominante. Os dilemas entre privacidade e responsabilização não são fatalidades impostas pela tecnologia, mas resultam de escolhas arquiteturais feitas por quem projeta e governa esses sistemas.

Os novos horizontes: o mapa da jornada

O diagrama da figura 7 sintetiza os cinco eixos de tensão que este livro explora. No centro, a questão que permeia todos eles: o equilíbrio entre vigilância e privacidade, entre controle e liberdade. Cada capítulo examina uma faceta específica.

Capítulo I Direito & Trabalho
ENTRE A PRIVACIDADE E A VIGILÂNCIA: A PROTEÇÃO DE DADOS PESSOAIS NAS RELAÇÕES DE TRABALHO

Proteção de dados nas relações de trabalho. Do taylorismo aos sistemas digitais. LGPD como instrumento de reequilíbrio.

Figura 7 Os cinco eixos do livro

Capítulo II Direitos & Biometria
O SUJEITO EM FRENTE ÀS CÂMERAS: DESAFIOS DO RECONHECIMENTO FACIAL DIANTE DA DIVERSIDADE HUMANA

Inversão de perspectiva: da câmera para o rosto. Vulnerabilidades do reconhecimento facial diante de crianças, idosos, gêmeos e oclusões faciais.

Capítulo III Tecnologia e Regulação
TRUSTED EXECUTION ENVIRONMENTS EM ECOSSISTEMAS DE CRIPTOGRAFIA PONTA A PONTA: ANÁLISE DE AMEAÇAS E DESAFIOS REGULATÓRIOS NO BRASIL

Análise de ameaças e desafios regulatórios no Brasil. A ortogonalidade entre TEEs e E2EE e os riscos de captura institucional da tecnologia.

Capítulo IV Segurança e Núvem
ANÁLISE DE SEGURANÇA DA COMPUTAÇÃO CONFIDENCIAL

Propriedades de segurança sob modelo de ameaça realista. Estudo de caso: Private Processing do WhatsApp.

Capítulo V Análise Sociotécnica
CRIPTOGRAFIA, CONFIANÇA E GOVERNANÇA

Moderação de conteúdo em sistemas E2EE como experimento institucional. Criptografia como tecnologia política que redistribui poder.

A vigilância tem limites. Nós os definimos.

As tecnologias de vigilância avançam incessantemente: câmeras mais precisas, algoritmos mais sofisticados, capacidade de processamento crescente. Mas os mecanismos de defesa também evoluíram. A criptografia, nascida como técnica militar de generais antigos, é hoje a infraestrutura invisível que sustenta a privacidade de bilhões. Os marcos regulatórios, como a LGPD, estabelecem limites jurídicos ao poder de observar. O debate público, informado por análises como as reunidas neste livro, é o que garante que esses limites sejam reais e não nominais.

Este livro nasce da convicção de que entender a vigilância e privacidade não é privilégio de especialistas. Da mesma forma, os debates sobre TEEs, reconhecimento facial, E2EE e vigilância no trabalho podem, e devem, ser acessíveis a todos que vivem sob o regime digital que eles descrevem.

> "A criptografia é a única forma de proteção que se baseia em matemática, não em confiança. Os marcos jurídicos são a única forma de proteção que se baseia em direitos, não em força. Ambos são necessários. Nenhum é suficiente sozinho."

A jornada começa aqui. Vire a página.

Capítulo I
Direito & Trabalho

ENTRE A PRIVACIDADE E A VIGILÂNCIA: A PROTEÇÃO DE DADOS PESSOAIS NAS RELAÇÕES DE TRABALHO

Bruno J. M. Marcolini[1]
Giuseppe Grando[2]

RESUMO: O avanço das tecnologias digitais ampliou as formas de controle no ambiente de trabalho, permitindo práticas de monitoramento cada vez mais intensas e muitas vezes pouco perceptíveis aos trabalhadores. Nesse contexto, este trabalho analisa a compatibilidade do monitoramento laboral com a Lei Geral de Proteção de Dados Pessoais ("LGPD") e com os direitos fundamentais do empregado. Para tanto, parte-se da premissa de que dados relacionados a desempenho, comportamento, horários, interações e uso de ferramentas digitais configuram dados pessoais e devem ser tratados conforme os parâmetros legais. A pesquisa, conduzida pelo método dedutivo, baseia-se em análise bibliográfica e documental, complementada pelo estudo de caso do monitoramento de funcionários no Banco Itaú. Os resultados

1. Bacharel em Direito pela UFPR, Pós-Graduado em Direito Digital pela FGV-SP e Pós-Graduando em Compliance em Cybersegurança pela PUC-RIO. Especialista em direito digital e propriedade intelectual. Advogado no CMT Advogados. Premiado no 2º Prêmio Danilo Doneda da ANPD.
2. Giuseppe Grando, Sócio-fundador do Grando Limberger Advogados, Bacharel em Direito pela Universidade de Santa Cruz do Sul - UNISC, Certificado pela Data Privacy Brasil, premiado no 3º Prêmio Danilo Doneda da ANPD.

indicam que o monitoramento somente é legítimo quando orientado por finalidade específica, proporcionalidade, transparência e ausência de discriminação, considerando ainda a vulnerabilidade estrutural do empregado na relação de trabalho. O estudo examina criticamente o uso das bases legais do consentimento, do legítimo interesse, da execução contratual e do cumprimento de obrigação legal, destacando seus limites no contexto laboral. Conclui-se que a LGPD fornece parâmetros relevantes para equilibrar o controle empresarial e a proteção dos direitos do trabalhador, exigindo práticas responsáveis e transparentes no tratamento de dados pessoais.

Palavras-Chave: Proteção de dados; vigilância; monitoramento no local de trabalho; privacidade; direitos fundamentais.

ABSTRACT: The advancement of digital technologies has expanded forms of control in the workplace, enabling monitoring practices that are increasingly intensive and often barely perceptible to workers. In this context, this study analyzes the compatibility of workplace monitoring with the Brazilian General Data Protection Law (LGPD) and with employees' fundamental rights. It is based on the premise that data related to performance, behavior, working hours, interactions, and the use of digital tools constitute personal data and must therefore be processed in accordance with legal standards. The research adopts a deductive method, grounded in bibliographic and documentary analysis and complemented by a case study on the monitoring of employees at Banco Itaú. The findings indicate that monitoring is legitimate only when guided by a specific purpose, proportionality, transparency, and non-discrimination, while also taking into account the employee's structural vulnerability within the employment relationship. The study critically examines the legal bases of consent, legitimate interest, contractual necessity, and compliance with legal obligations, highlighting their limitations in the labor context. It concludes that the LGPD provides important parameters for balancing corporate oversight with the protection of workers' rights, requiring responsible and transparent practices in the processing of personal data.

Key-words: Data protection; surveillance; labor monitoring; privacy; fundamental rights.v

1. Introdução

A consolidação do teletrabalho e a digitalização das ferramentas de ofício reconfiguraram o exercício do controle patronal, transformando a fiscalização tradicional em mecanismos de vigilância ubíqua que diluem as fronteiras entre o ambiente laboral e a vida privada. Diante dessa nova realidade telemática, a presente investigação se faz embasada na relação laboral e na proteção de dados, com base nos pilares abaixo descritos.

O problema de pesquisa que orienta este estudo consiste em questionar de que maneira a implementação de mecanismos de vigilância tecnológica ubíqua no ambiente laboral tensiona o direito fundamental à privacidade e como a Lei Geral de Proteção de Dados - LGPD pode servir de parâmetro para equilibrar a eficiência empresarial com a preservação da esfera íntima do trabalhador.

A hipótese aventada é que a vigilância digital[3] moderna, ao trans-

3. Utilizaremos o entendimento de vigilância digital elencado por Elias Jacob de Menezes Neto, baseado nas ideias de David Lyon, ou seja, que seria a "atenção concentrada, sistematizada e rotineira aos dados pessoais cujo objetivo é influenciar, gerenciar, proteger ou dirigir" (Menezes Neto, 2014, p. 7).

bordar o monitoramento de resultados para a captura de comportamentos e metadados, cria uma zona de intrusão que desafia o conceito clássico de privacidade laboral, exigindo que o controle seja submetido rigorosamente ao regime de proteção de dados para garantir proporcionalidade, transparência e finalidade legítima.

A justificativa para a realização desta pesquisa reside na necessidade premente de oferecer segurança jurídica a um mercado de trabalho em transformação, posicionando a conformidade com a LGPD não apenas como uma obrigação formal, mas como um parâmetro indispensável para a construção de relações laborais transparentes, equilibradas e compatíveis com a dignidade da pessoa humana.

O método de pesquisa adotado é o dedutivo, fundamentado em análise bibliográfica e documental, complementado pelo estudo de caso do recente episódio envolvendo o monitoramento granular de funcionários no Banco Itaú, utilizado como o fio condutor para a análise prática da tensão entre o poder diretivo e a privacidade.

A estrutura deste artigo foi organizada para permitir uma compreensão progressiva da temática. O capítulo 2 analisa o regime jurídico e a coexistência entre o poder diretivo e a privacidade no contexto do teletrabalho, enquanto o capítulo 3 qualifica tecnicamente os dados de monitoramento. Nessa etapa, distingue-se o dado bruto da informação estratégica para demonstrar como a telemetria comportamental (rastro do uso de periféricos como mouse e teclado) configura dado pessoal no processo de dataficação laboral.

A análise prossegue no capítulo 4 com o exame da legalidade do monitoramento sob os princípios da LGPD. Por fim, o capítulo 5 aborda a implementação das bases legais, avaliando a insuficiência do consentimento devido à assimetria de poder e detalhando os critérios para aplicação do legítimo interesse, da execução contratual e do dever legal na vigilância tecnológica.

Em suma, o objetivo geral deste artigo é analisar a legalidade e os limites ético-jurídicos da vigilância no trabalho frente ao binômio privacidade e proteção de dados, fixando as balizas necessárias para a harmonização entre a gestão empresarial e a preservação dos direitos fundamentais do empregado.

Por fim, a investigação demonstra que o monitoramento laboral exige governança pautada em proporcionalidade para legitimar-se frente às novas tecnologias. A qualificação de rastros comportamentais como dados pessoais, através da interpretação expansionista e pela Teoria do Mosaico (Madrid Conesa, 1984, apud Tena et al., 2020), submete tais práticas à LGPD, impondo, consequentemente, o respeito aos princípios de finalidade e necessidade. Reafirma-se que a proteção de dados é o instrumento de reequilíbrio na dataficação do trabalho, conciliando a eficiência empresarial com a dignidade e os direitos fundamentais do trabalhador.

2. O regime jurídico da proteção de dados nas relações de trabalho

A distinção fundamental nas relações de trabalho reside na natureza do controle exercido sobre a atividade laboral, de modo que a existência de controle direto configura a subordinação, enquanto o trabalho executado de maneira autônoma e sem diretrizes diretas é classificado como independente (Amaral; Moreira, 2024).

Com a vigência da Lei 12.551/2011, estabeleceu-se que os dispositivos telemáticos e informatizados utilizados para supervisão possuem equivalência jurídica aos métodos convencionais de controle, tornando viável a manutenção do vínculo de subordinação mesmo em regimes de trabalho à distância (Amaral; Moreira, 2024).

Essa prerrogativa patronal é definida como poder empregatício, o

qual assegura ao empregador a capacidade de exercer direção, comando e fiscalização sobre a prestação de serviços. Tal poder autoriza o monitoramento do trabalho, inclusive por intermédio de aparatos tecnológicos, embora sua aplicação deva ser conduzida com cautela e dentro de limites estabelecidos (Delgado, 2019).

Nesse sentido, ao exercer suas funções de organização, o empregador não é impedido de realizar o monitoramento das atividades, mas encontra uma vedação absoluta no que tange à invasão[4] da esfera de privacidade do empregado (Alkimin, 2007).

A tutela jurídica dos dados pessoais no contexto laboral não se limita à proteção da privacidade, mas busca resguardar o indivíduo contra práticas discriminatórias e formas excessivas de controle, visando garantir o pleno desenvolvimento da personalidade. Dentro da relação empregatícia, o trabalhador é qualificado como o titular de dados, uma vez que disponibiliza informações pessoais ao empregador, o qual assume a posição de controlador. Por conseguinte, qualquer dado relativo a uma pessoa natural identificada ou identificável, deve ser tratado sob o rigor da legislação vigente (Ferreira, 2023).

Contudo, a proteção de dados enfrenta a assimetria da relação laboral, na qual a vulnerabilidade e a subordinação do empregado, cuja subsistência depende do vínculo, comprometem sua real liberdade de vontade. Nesse cenário, o cumprimento rigoroso da LGPD é essencial. Tais preceitos abrangem tanto o trabalho presencial quanto o teletrabalho, modalidade consolidada que catalisou a reavaliação das

4. Para fins de compreensão neste trabalho, entende-se por "invasão da esfera privada" o momento no qual o empregador, utilizando-se de meios de monitoramento, passa a acompanhar/fiscalizar elementos que não fazem parte da relação laboral, existindo, portanto, uma observação da intimidade do funcionário/titular de dados, gerando incongruência com o interesse e a finalidade inicialmente proposta, que é exclusivamente na esfera da relação de emprego.

normas brasileiras sobre direitos.

Embora essa modalidade ofereça benefícios como a flexibilidade na gestão do tempo e a eliminação de deslocamentos, ela introduz o desafio crítico de estabelecer fronteiras nítidas entre a esfera laboral e a vida privada. A ausência dessa delimitação clara tem culminado tanto em jornadas de trabalho excessivas, elevando os índices de estresse e demandando uma regulamentação que assegure o bem-estar e o direito à desconexão do obreiro (Martos et al., 2025), quanto no que tange aos limites do tratamento de dados e da vigilância.

Nesse cenário, a privacidade deve ser compreendida sob a ótica do contrato de trabalho realizado de forma telepresencial, no qual as dimensões pessoais e profissionais do indivíduo acabam por se fundir no ambiente doméstico. O monitoramento remoto expõe dados que usualmente não seriam acessados no regime presencial, tais como a rotina familiar, a imagem do empregado em seu lar e a interação com outros residentes (Amaral, Moreira, 2024).

Dessa forma, o empregador deve exercer seu poder diretivo e, consequentemente, sua prerrogativa de controle tendo em mente, sempre, os direitos fundamentais do trabalhador que se exemplifica pelo direito à intimidade, à privacidade e à proteção de dados.

O avanço da digitalização e o advento de novas ferramentas de controle introduzem formas inéditas de vigilância – como os "bossware", explicados ao longo deste artigo – que, se não forem devidamente equilibradas, podem resultar em ambientes laborais opressores e na insatisfação crônica do titular de dados, de forma que a sensação de "disponibilidade constante" e o "monitoramento rigoroso", além de abalarem a relação empregatícia e ameaçar direitos básicos de privacidade, são vetores significativos para o esgotamento mental e o desenvolvimento da síndrome de burnout (Martos et al., 2025).

Isso demonstra, mais uma vez, que o uso de sistemas de fiscalização deve ser pautado por princípios intrínsecos à LGPD, como o da finalidade, da necessidade e da não discriminação, a fim de impedir que o controle da produtividade se transforme em uma vigilância indiscriminada.

Somado ao risco à privacidade, a dinâmica do trabalho remoto implica uma movimentação constante de dados que exige a adoção de melhores práticas de segurança por parte das organizações, de forma que a vulnerabilidade informacional é acentuada quando o empregado utiliza dispositivos pessoais desprovidos de sistemas de proteção adequados, o que pode comprometer a integridade e o armazenamento das informações tratadas (Amaral, Moreira, 2024).

Assim, a conformidade com a LGPD no contexto do teletrabalho exige normas claras que previnam abusos e garantam um ambiente de trabalho pautado no respeito mútuo e na segurança jurídica (Martos et al., 2025).

O conflito entre poder diretivo e privacidade exemplifica-se no caso do Banco Itaú, que demitiu teletrabalhadores com base em métricas de inatividade capturadas por softwares de monitoramento de tela, teclado e mouse. Conforme reportagem do Intercept Brasil, a controvérsia reside na falta de transparência sobre essa vigilância granular, ferindo o dever de informação (Pannunzio, 2025).

Quando o vínculo empregatício e a dignidade passam a depender de interações mecânicas com o hardware, urge investigar a essência do monitoramento: não simples logs técnicos, mas rastros que perfilam o comportamento digital do indivíduo. Assim, aferir a legalidade dessas práticas exige, preliminarmente, a devida qualificação jurídica dos dados coletados, conforme analisado a seguir.

3. Qualificação dos dados pessoais no monitoramento do trabalho

A compreensão da proteção de dados no ambiente laboral exige distinguir dado e informação. O dado corresponde a um elemento isolado que, por si só, não transmite uma mensagem estruturada, enquanto a informação resulta do tratamento e da análise desses elementos. No contexto organizacional, dados desconexos possuem pouco valor prático, passando a adquirir utilidade quando são organizados e interpretados, transformando-se em recurso estratégico para compreender o ambiente (Zavanella, 2023).

A legislação brasileira, acompanhando tendências internacionais como o Regulamento Geral sobre a Proteção de Dados (GDPR) da União Europeia, adotou uma definição expansionista de dado pessoal, fundamentada na figura da pessoa natural.

Enquanto a identificação direta refere-se à individualização inequívoca do sujeito por meio de documentos únicos, fotografias ou registros específicos, o conceito de "identificável" é mais abrangente, evocando a probabilidade de se atingir a identidade do titular por meio da combinação de informações aparentemente genéricas, como o bairro de residência somado a preferências pessoais. No ecossistema laboral telemático, essa distinção é importantíssima, pois quem busca manter sua privacidade o faz tanto em relação a conhecidos quanto a terceiros, fazendo com que qualquer rastro digital, que seja capaz de individualizar o trabalhador, se torne um objeto de proteção legal (Zavanella, 2023).

Nesse exato sentido, conseguimos perceber que a evolução tecnológica permitiu a disseminação de softwares de vigilância extremamente sofisticados, comumente chamados de "bossware" – que seria a junção das palavras "boss" e "softwares", evocando algo semelhante ao "sistema

do chefe" –, instalados nos dispositivos de trabalho para monitorar as atividades online dos empregados (Tung et al., 2025).

Esses sistemas não se restringem ao controle da jornada, mas realizam uma extração precisa de informações por meio de recursos como as teclas digitadas, capturas de tela, movimento do mouse e, até mesmo a análise de expressões faciais, capturando desde protocolos de transferência e aplicativos utilizados até preferências musicais e mensagens instantâneas (Zavanella, 2023).

O acompanhamento de métricas comportamentais, a exemplo dos registros de movimentação do mouse e da utilização do teclado, transforma os meios tecnológicos em instrumentos de vigilância permanente, possibilitando ao empregador registrar, de forma minuciosa, as características e atividades desempenhadas pelo trabalhador.

Embora o exercício do poder diretivo possa conferir uma legitimidade inicial à fiscalização com o intuito de melhorar a performance e resguardar o patrimônio empresarial, a intrusividade desses sistemas exige um balanceamento rigoroso (Zavanella, 2023).

O uso desproporcional dessas tecnologias, que resulta em uma vigilância massiva e indiscriminada, desvirtua a finalidade do controle de produtividade e cria ambientes laborais opressores, tornando-se um vetor significativo para a insatisfação e o esgotamento mental do titular (Martos et al., 2025).

Essa temática da vigilância e seus impactos na privacidade e proteção de dados ganhou contornos práticos significativos com o recente episódio ocorrido no Banco Itaú, no qual demissões foram fundamentadas em métricas de inatividade obtidas por meio de monitoramento tecnológico.

Conforme reportagem no Intercept Brasil, os funcionários que atuam em teletrabalho receberiam notebooks corporativos com softwares instalados, como o "xOne", que permitiriam acompanhar métricas granulares, incluindo o uso do teclado, a movimentação do mouse e o acesso a sites (Pannunzio, 2025).

Nesse sentido, esses programas trabalham com o rastreamento da atividade online, identificando se o colaborador cumpre a jornada ou possui períodos de inatividade, gerando painéis de produtividade acessíveis à empresa (G1, 2025). O problema central reside no fato de que o monitoramento do uso do mouse e teclado foi realizado, em muitos casos, sem o aviso prévio e adequado aos trabalhadores (Pannunzio, 2025).

Esse cenário ilustra o fenômeno da dataficação do trabalho, onde as experiências humanas e os aspectos profissionais são transformados em dados matemáticos com o intuito de serem avaliados de forma quantitativa e objetiva, de forma que essa metrificação utilizaria os dados como matéria-prima para gerar predefinições de comportamentos desejáveis no ambiente corporativo (Miziara, 2024).

Nesse sentido, o dado bruto, que isoladamente não transmite uma mensagem, transforma-se em uma informação muito estratégica após passar por processos de tratamento, análise e contextualização, de forma que o interesse empresarial na formação desse conhecimento objetiva a projeção de tendências futuras e o mapeamento detalhado do ambiente laboral por meio da análise em grandes volumes (Zavanella, 2023).

A utilização de técnicas como o data mining permite que o empregador reconheça correlações e padrões em bases de dados que seriam de difícil compreensão, possibilitando a classificação de pessoas a partir de informações brutas preexistentes (Zavanella, 2023).

Softwares de vigilância, como bossware, supracitado, rastreiam automaticamente atividades que vão desde o uso do teclado até expressões. Embora essas ferramentas possam ser apresentadas como úteis para garantir a produtividade, seu uso deve ser equilibrado para evitar a invasão da privacidade, uma vez que o controle constante pode levar ao esgotamento mental do trabalhador (Martos et al., 2025).

A profundidade dessa vigilância possibilita que dados inicialmen-

te comuns, como o ritmo de cliques, sejam transformados em dados pessoais anteriormente desconhecidos por meio do processamento algorítmico. O empregador, através de recursos de observação, promove uma extração de informações com alto nível de precisão, permitindo a coleta de dados sobre protocolos de transferência e aplicativos usados, o que pode revelar aspectos íntimos sobre o titular (Zavanella, 2023).

Portanto, a conformidade com a LGPD exige que o uso de tecnologias de monitoramento seja pautado por normas claras que previnam abusos e garantam um ambiente de trabalho pautado na segurança jurídica e no respeito aos direitos fundamentais (Martos et al., 2025).

A qualificação da telemetria como dado pessoal intrusivo fundamenta a avaliação de sua conformidade normativa. Embora o monitoramento seja histórico, a migração para sistemas digitais e IA ampliou sua escala e profundidade, submetendo o controle ubíquo à proteção de dados e exigindo a harmonização do poder diretivo com os direitos fundamentais. Os capítulos seguintes analisam a legalidade dessas práticas sob os princípios de finalidade, necessidade e não discriminação, além das bases legais e diretrizes de governança essenciais à preservação da dignidade humana no ambiente laboral digital.

4. Análise de legalidade do monitoramento de trabalho perante a LGPD

Antes de se adentrar o regime jurídico atual da proteção de dados pessoais em relação ao monitoramento do ambiente de trabalho, importa destacar que não se trata de uma atividade nova que surge em decorrência de uma tecnologia em específico, mas de uma prática que possui séculos de utilização e aperfeiçoamento.

Durante o século XX, por exemplo, uma das práticas mais comuns foi o controle por meio de registros manuais de produção. Em repartições públicas e escritórios, funcionários precisavam preencher livros de ponto, fichas de protocolo e relatórios diários, permitindo que a alta direção acompanhasse volume e ritmo de trabalho. Esse tipo de monitoramento não avaliava apenas a presença, mas também a produtividade individual e coletiva.

Em ambientes mais industrializados, o taylorismo e o fordismo influenciaram o trabalho documental (Meiksins, 1984). Cronometragem de tarefas administrativas, estudos de tempos e movimentos e padronização de procedimentos passaram a ser aplicados também em escritórios, com metas de produção de páginas datilografadas, número de processos analisados ou correspondências respondidas por dia.

Com a digitalização das atividades e a crescente demanda por eficiência organizacional, esses mecanismos foram gradualmente substituídos ou complementados por sistemas digitais capazes de coletar e analisar grandes volumes de informações. Independentemente do grau de sofisticação tecnológica, contudo, o monitoramento do trabalho envolve o tratamento de dados pessoais relacionados ao desempenho e às atividades do trabalhador. Por essa razão, tais práticas se submetem ao regime da LGPD, cujos princípios orientam a avaliação da licitude do monitoramento laboral.

4.1 Princípio da Finalidade

Especialmente em atividades de monitoramento, o princípio da finalidade cresce em relevância prática, visto que garante ao titular, sempre mediante informação prévia, os limites da legalidade do tratamento com base no propósito definido pelo agente de tratamento[5]. Assim, o

5. Os agentes de tratamento são os responsáveis pelo tratamento dos dados pessoais, sujeitos às regras da LGPD e à fiscalização da ANPD. O controla-

tratamento do dado pessoal deverá estar sempre vinculado à finalidade que motivou a sua aquisição, criando uma ligação entre a informação e sua origem (Doneda, 2015). Tal finalidade, inclusive, deverá ser avaliada até mesmo para tratamentos posteriores.

Para que o monitoramento do trabalhador seja válido, portanto, é preciso que a finalidade esteja claramente delimitada, seja legítima e compatível com a atividade exercida, além de ser previamente informada ao empregado. O controle não pode ser genérico, ilimitado ou desvinculado de uma necessidade concreta da relação de trabalho, sob pena de violar o princípio da finalidade e esvaziar a proteção conferida ao titular dos dados. Em outras palavras, o monitoramento só se justifica quando orientado por objetivos específicos, como segurança, organização do trabalho ou cumprimento de obrigações legais, devendo qualquer uso posterior das informações permanecer coerente com esses propósitos inicialmente declarados.

Quanto à definição de finalidade nas relações empregatícias, a própria CLT reconhece o poder de direção do empregador, que inclui a organização, fiscalização e controle da prestação de serviços. O artigo 2º atribui ao empregador a condução da atividade econômica e a supervisão do trabalho, enquanto o § 6º do artigo 6º admite o uso de meios tecnológicos para acompanhar a execução das atividades. Além disso, o artigo 74, §§ 1º e 2º, impõe às empresas com mais de vinte empregados o dever de registrar a jornada de trabalho (Brasil, 1943). Em conjunto, esses dispositivos fornecem fundamento jurídico para o monitoramento laboral, desde que exercido dentro de limites e sem abuso.

Nesse sentido, o Tribunal Superior do Trabalho firmou entendimento de que é lícito ao empregador instalar câmeras de vigilância em ambientes coletivos de trabalho, desde que posicionadas em áreas comuns e com prévia ciência dos empregados (TST, 2020). Ao mesmo tempo, a Corte estabeleceu limites claros, vedando a instalação de câmeras em locais destinados ao descanso ou em espaços com expectativa qualificada de privacidade, como banheiros e vestiários. O julgado evidencia a necessidade de compatibilizar o legítimo interesse do empregador em fiscalizar a prestação do trabalho com a proteção à privacidade e à dignidade do trabalhador.

O caso também ilustra que a responsabilidade pelo monitoramento recai principalmente sobre o empregador, que possui a prerrogativa de adotar medidas para proteger os ativos da empresa. Essa prerrogativa, contudo, não é ilimitada. O poder de fiscalização deve ser orientado por critérios objetivos e necessidades reais da organização, e não por interesses pessoais ou desconfianças subjetivas em relação a um trabalhador específico. Monitoramentos direcionados individualmente, sem justificativa legítima e proporcional, podem configurar abuso de direito e violação à privacidade do empregado.

Sanções nesse sentido, inclusive, já são realidade para o Brasil. Em 2023, o Tribunal Regional do Trabalho da 4ª Região afastou a validade de uma demissão por justa causa e condenou uma construtora ao pagamento de R$ 5 mil a título de indenização por danos morais, em razão de o empregador ter acessado e fiscalizado conversas mantidas pelo empregado por meio do WhatsApp (CPA, 2023). Na fundamentação da sentença, o juiz da 3ª Vara do Trabalho de São Leopoldo, destacou que o acesso e o tratamento, pelo empregador, de informações e dados pessoais provenientes de contas privadas dependem de autorização do titular e da existência de finalidade específica e legítima. Esse entendimento foi posteriormente acolhido e reproduzido pelo Tribunal Regional do Trabalho no acórdão, reforçando a compreensão de que

dor é quem toma as decisões referentes ao tratamento de dados pessoais e o operador, aquele que realiza o tratamento de dados pessoais em nome do controlador.

o monitoramento de comunicações privadas extrapola os limites do poder diretivo do empregador quando os requisitos do princípio da finalidade não são observados.

Em síntese, o monitoramento no contexto das relações de trabalho é juridicamente possível, mas condicionado ao respeito a limites claros. A finalidade do tratamento deve ser legítima, específica e previamente informada ao empregado, funcionando como critério central para aferir a licitude das práticas de controle. A CLT oferece base para o exercício do poder diretivo, mas a interpretação sistemática com a LGPD evidencia que esse poder não autoriza vigilância genérica, excessiva ou direcionada de forma arbitrária. Sempre que o monitoramento se afasta de uma necessidade concreta e proporcional, sobretudo quando atinge esferas de privacidade, tende a ser considerado abusivo e passível de sanção.

4.2 Princípios da adequação e da necessidade

Por meio dos princípios da adequação e da necessidade, define expressamente a LGPD que o tratamento de dados pessoais deve ser compatível com as finalidades informadas ao titular, sempre em observância ao contexto da atividade. Ademais, os dados pessoais tratados devem se limitar ao mínimo necessário para a execução pretendida, com uma coleta proporcional e não excessiva (Brasil, 2018).

Consequentemente, é preciso tomar especial atenção ao implementar mecanismos de monitoramento laboral para que os dados pessoais coletados sejam tão somente aqueles estritamente necessários para que a finalidade do monitoramento seja atingida. Tais princípios também impõe desafios sob o ponto de vista do armazenamento das informações, que só podem ser retidas pelo período necessário para o cumprimento da finalidade que justificou a sua coleta, devendo ser descartadas ou anonimizadas após o atingimento desse objetivo (com exceção de prazos extendidos por questões trabalhistas). Isso exige do empregador a adoção de práticas voltadas ao privacy by design, isto é, a adoção de medidas proativas que resguardem a privacidade do titular antes da implementação do monitoramento.

Conforme ensina Vainzof (2022), diante do princípio da necessidade, é preciso buscar as respostas para as seguintes perguntas: (i) a finalidade pretendida pode ser atingida de outro modo, sem a utilização dos dados pessoais?; (ii) quais dados pessoais e qual o volume adequado para o objetivo pretendido; e (iii) quais os potenciais riscos aos titulares.

Assim, em iniciativas como a de monitoramento do uso da Internet durante a jornada de trabalho, o empregador deve refletir sobre a real necessidade e a proporcionalidade das medidas adotadas. Não é razoável, por exemplo, registrar todo o conteúdo acessado ou tudo o que é digitado pelo empregado apenas para identificar eventuais quedas de produtividade, prática comum em algumas ferramentas de tecnologia de RH. Em muitos casos, é possível atingir o mesmo objetivo por meio da coleta de informações menos invasivas, como a identificação dos sites acessados e do tempo de navegação, sem acessar o teor das comunicações. Do contrário, o monitoramento pode acabar captando dados sensíveis de forma indevida: basta imaginar a hipótese de o trabalhador acessar seu e-mail pessoal e receber o resultado de um exame médico, o que implicaria a coleta de dados de saúde, que gozam de proteção reforçada.

Todavia, como bem ressalvam Pulga, da Silva e Pasqualeto (2025), com o avanço das tecnologias digitais, as fronteiras do que pode ou não ser considerado monitoramento legítimo no ambiente de trabalho tornam-se progressivamente mais tênues. A jurisprudência do Tribunal Superior do Trabalho tem admitido que o empregador acompanhe o uso de ferramentas e recursos disponibilizados pela própria empresa,

como computadores corporativos e e-mails institucionais, ao mesmo tempo em que veda a fiscalização de dispositivos e contas pessoais. Ainda assim, esse tipo de controle exige transparência, o trabalhador deve ser previamente informado sobre o monitoramento e orientado quanto aos seus riscos e implicações.

Os autores destacam também que o ato de questionar práticas de monitoramento que ultrapassam o que é considerado razoável pela legislação brasileira costuma ser especialmente difícil para os empregados. Em muitos casos, eles sequer conseguem compreender plenamente a extensão da vigilância a que estão submetidos, seja pela complexidade técnica das ferramentas utilizadas, seja pela falta de transparência e pelas constantes mudanças nos sistemas adotados pelas empresas. Além disso, mesmo quando percebem que o controle pode ter excedido os limites legais, formalizar uma reclamação contra o empregador gera receio, pois pode afetar negativamente a relação de trabalho e expor o empregado a possíveis retaliações.

Essa assimetria de poder evidencia a vulnerabilidade estrutural do trabalhador diante de práticas excessivas de monitoramento e reforça a necessidade de salvaguardas mais robustas. Nesse contexto, ganham relevância mecanismos institucionais de proteção, como a atuação dos sindicatos, dos órgãos de fiscalização trabalhista e da Agência Nacional de Proteção de Dados ("ANPD"), bem como a exigência de maior transparência e de políticas internas claras por parte das empresas. Sem esses instrumentos de equilíbrio, o risco é que a ampliação das tecnologias de controle transforme o ambiente de trabalho em um espaço de vigilância permanente, incompatível com a dignidade e os direitos fundamentais do empregado.

Diante desse panorama, conclui-se que o monitoramento laboral é legítimo apenas se estruturado sob critérios rigorosos de finalidade, necessidade e proporcionalidade, integrando a proteção de dados desde a concepção das ferramentas. Embora a tecnologia amplie o controle, ela intensifica riscos aos direitos fundamentais na relação assimétrica entre capital e trabalho. Portanto, a conformidade com a LGPD e a legislação trabalhista é parâmetro indispensável e mínimo para relações transparentes, equilibradas e compatíveis com a dinidade humana.

4.3 Princípio da não discriminação

Conforme mencionado nos pontos anteriores, para que o monitoramento laboral seja entendido como adequado, é imprescindível que a ação seja orientada por critérios objetivos e impessoais. Consequentemente, qualquer prática de monitoramento que selecione, direcione ou impacte determinados trabalhadores de forma desproporcional, sem justificativa legítima, tende a violar o princípio da não discriminação previsto na LGPD. Esse princípio veda o tratamento de dados pessoais para fins discriminatórios, ilícitos ou abusivos e impõe ao empregador o dever de estruturar seus mecanismos de controle de modo a evitar vieses, estigmatizações ou diferenciações arbitrárias entre empregados.

No contexto laboral, isso significa que o monitoramento não pode ser utilizado como instrumento de perseguição individual, de reforço de estereótipos ou de controle direcionado a grupos específicos, seja com base em características pessoais, seja em critérios subjetivos. É nesse ponto que a utilização da tecnologia deve ser especialmente supervisionada, pois sistemas automatizados de avaliação de desempenho e ferramentas de produtividade podem estar atreladas aos risco de geração de perfis discriminatórios (Pulga et al., 2025).

Nesse sentido, um relatório recente do Instituto for Public Policy Research no Reino Unido mostrou que trabalhadores negros são desproporcionalmente mais sujeitos a vigilância no ambiente de trabalho do que seus colegas não negros (Albon, Beeby, 2025). Segundo o estudo,

funcionários negros têm maior probabilidade de ocupar funções com menor autonomia, onde a vigilância é mais intensa, e, por isso, enfrentam monitoramento mais invasivo e frequente. O estudo também relembra que a utilização deste tipo de dados para o treinamento de sistema de IA possui o condão de criar novas tecnologias reprodutoras de vieses.

Outro caso ocorrido no estado da Geórgia, nos Estados Unidos, conhecido como "devious defecator", ilustra de forma clara como práticas de monitoramento podem assumir contornos discriminatórios. Diante de episódios de vandalismo em um armazém, em que fezes humanas eram deixadas no local de trabalho, o empregador decidiu recorrer a testes de DNA para identificar o responsável (Genetic Literacy Project, 2015). Em vez de adotar medidas gerais, como o uso de câmeras em áreas comuns, a empresa optou por submeter apenas dois empregados ao exame genético, ambos negros. O resultado demonstrou que nenhum deles era o autor da conduta, e os trabalhadores ajuizaram ação contra o empregador. A decisão judicial reconheceu que a escolha direcionada desses dois indivíduos poderia ser compreendida como prática discriminatória e caracterizadora de assédio, além de violar a Genetic Information Nondiscrimination Act (GINA). O caso é particularmente relevante porque evidencia não apenas a desproporcionalidade e inadequação do meio de monitoramento escolhido, mas também como uma legislação originalmente voltada a impedir discriminação genética em processos de contratação e manutenção do emprego passou a ser interpretada como instrumento de contenção de práticas abusivas de vigilância, reforçando a necessidade de critérios objetivos, proporcionais e não discriminatórios na fiscalização da conduta de trabalhadores.

Estudos e casos como os citados acima demonstram como o princípio da não discriminação é especialmente importante no contexto de monitoramento laboral. Trazendo ao contexto brasileiro, a observância do princípio exige que o empregador seja capaz de demonstrar a racionalidade e a legitimidade dos critérios utilizados no monitoramento, inclusive sob a perspectiva da prestação de contas (accountability). A ausência de transparência sobre como os dados são coletados, analisados e utilizados para tomada de decisões pode potencializar práticas discriminatórias e fragilizar ainda mais a posição do trabalhador na relação de emprego.

4.4 Princípio da transparênciae do livre acesso

Durante a análise dos princípios anteriores, buscou-se frisar que o tratamento de dados pessoais realizado no contexto do monitoramento laboral só será válido se o empregador fornecer ao empregado informações prévias suficientes, isto é, claras, acessíveis e compreensíveis sobre a existência do monitoramento, suas finalidades, os tipos de dados coletados, a forma de utilização dessas informações e o período de retenção.

O princípio da transparência exige que o titular não seja surpreendido por práticas ocultas de vigilância, mas que tenha condições reais de compreender como e por que seus dados estão sendo tratados no ambiente de trabalho. De modo complementar, o princípio do livre acesso assegura ao empregado o direito de consultar, a qualquer momento, as informações que o empregador mantém a seu respeito, bem como obter esclarecimentos sobre o tratamento realizado. Esses princípios reforçam a necessidade de políticas internas bem estruturadas, comunicações claras e canais efetivos de atendimento aos titulares, sob pena de o monitoramento tornar-se incompatível com a LGPD e com a própria lógica de proteção da autonomia e da dignidade do trabalhador.

Frise-se, novamente, que especialmente em questões de monitoramento, o titular carece de ampla informação sobre a forma como seus dados (desempenho, horários de entrada e saída, postura etc.) serão avaliados, reforçando as dificuldades de superar a barreira in-

formacional que pode atrapalhar o trabalhador no momento de tomar ações que o favoreçam.

Dessa forma, é fundamental que o empregador considere essa vulnerabilidade, e forneça medidas de transparência proporcionais (qualitativa e quantitativa) para que o empregador entenda exatamente às regras corporativas às quais está sujeito.

Nesse contexto, é preciso que o empregador se questione se o simples aviso genérico de que "o ambiente é monitorado" é suficiente para atender às exigências da LGPD. A resposta tende a ser negativa. A transparência efetiva pressupõe informações concretas e detalhadas, que permitam ao trabalhador compreender o alcance real do monitoramento e suas consequências práticas. Comunicados vagos ou cláusulas padronizadas inseridas em contratos de trabalho dificilmente cumprem esse papel, sobretudo quando não explicam quais ferramentas são utilizadas (incluindo IA), quais dados são coletados, como esses dados influenciam avaliações de desempenho/gerenciais e por quanto tempo permanecem armazenados.

Por essa razão, ganham especial relevância as políticas internas específicas sobre monitoramento e proteção de dados no contexto laboral. Esses documentos devem estabelecer, de forma acessível, as regras aplicáveis, os limites do controle exercido pelo empregador, os direitos dos empregados enquanto titulares de dados e os canais disponíveis para esclarecimento de dúvidas e exercício de direitos. Mais do que instrumentos formais, tais políticas devem ser efetivamente comunicadas e incorporadas à cultura da organização, com treinamentos e atualizações periódicas, especialmente diante da adoção de novas tecnologias de controle.

Além disso, em situações em que o monitoramento possa gerar riscos elevados aos direitos dos trabalhadores, como ocorre, com frequência, em ambientes de trabalho remoto e no uso de ferramentas de vigilância digital, mostra-se recomendável a elaboração de Relatórios de Impacto à Proteção de Dados (RIPD/DPIA) (Vainzof, 2022). Esses relatórios permitem mapear riscos, justificar escolhas tecnológicas e demonstrar que o empregador adotou medidas para mitigar eventuais impactos negativos, funcionando também como instrumento de accountability.

A lógica da responsabilização exige, ainda, a existência de canais acessíveis e efetivos de comunicação entre empregador e empregado, de modo que o titular possa exercer seus direitos previstos na LGPD sem receio e sem entraves desproporcionais. Isso inclui a possibilidade de solicitar acesso aos dados, correção de informações, esclarecimentos sobre critérios de avaliação e até a revisão de decisões tomadas com base em tratamento automatizado. Em ambientes de trabalho remoto, nos quais o monitoramento tende a ser mais intenso e menos visível, essas garantias tornam-se ainda mais relevantes, sob pena de se consolidar um modelo de vigilância permanente incompatível com os parâmetros legais de proteção de dados e com os direitos fundamentais do trabalhador (Pulga et al., 2025).

Destaca-se que a tomada de decisões em contextos de assimetria informacional pode, inclusive, abrir margem para repercussões negativas para a organização. Exemplo disso é o caso Itaú, supracitado neste trabalho, cuja repercussão ganhou grande apelo midiático.

Em matéria à BBC, um dos mil funcionários demitidos afirmou que as medidas exatas eram desconhecidas, "a gente suspeitava, porque tem um monte de monitoramentos nos nossos computadores. Mas não sabíamos que monitoravam cliques, alt tab, scroll, tempo em reunião, coisas assim" (BBC News, 2025).

Ressalta-se que o intuito da análise sobre o caso citado não é afirmar se a conduta adotada pelo banco está, ou não, juridicamente correta, mas destacar como a transparência é um pilar inegociável quando

se trata de monitoramento eficiente para ambos os lados. Ainda assim, a análise do ocorrido ajuda a observar que o respeito aos princípios da transparência e do livre acesso não apenas legitimam juridicamente o tratamento de dados no ambiente de trabalho, como também atuam como instrumento de proteção da dignidade do empregado e de prevenção de abusos. Ignorá-los expõe o empregador a riscos jurídicos, reputacionais e organizacionais, além de contribuir para a consolidação de práticas de vigilância incompatíveis com o modelo de proteção de dados pessoais adotado pelo ordenamento brasileiro.

Por fim, ainda que o presente trabalho não tenha abordado de forma exaustiva todos os princípios previstos na LGPD, é fundamental reconhecer que cada um deles desempenha papel relevante na conformação do tratamento de dados pessoais no contexto do monitoramento laboral. A observância parcial ou seletiva desses princípios compromete a licitude e a legitimidade das práticas adotadas. Assim, para que o monitoramento de empregados seja considerado adequado e compatível com a LGPD, impõe-se ao empregador o dever de respeitar o conjunto integral dos princípios legais, de maneira articulada e coerente, sob pena de violação ao regime de proteção de dados e aos direitos fundamentais dos trabalhadores.

5. Implementação de bases para o monitoramento

A adequação do monitoramento laboral à LGPD não depende apenas da observância de princípios e da adoção de boas práticas de governança, mas exige, de forma incontornável, a existência de uma base legal que legitime o tratamento de dados pessoais. A lei brasileira parte da premissa de que todo tratamento deve estar juridicamente justificado, de modo que a ausência de fundamento legal torna a atividade ilícita, ainda que haja transparência, políticas internas ou medidas de segurança. No contexto das relações de trabalho, essa exigência ganha contornos ainda mais relevantes, pois o tratamento costuma envolver dados sensíveis à esfera privada do empregado e ocorre em um ambiente marcado por assimetria de poder, o que demanda maior rigor na análise da licitude.

Especificamente em práticas de monitoramento de empregados, o leque de bases legais aplicáveis é mais restrito do que em outros contextos. Tendo em vista as especificidades desta forma de tratamento, apenas quatro hipóteses tendem a oferecer sustentação jurídica plausível: (i) o consentimento do titular; (ii) a execução do contrato de trabalho; (iii) o legítimo interesse do empregador; e (iv) o cumprimento de obrigação legal ou regulatória (BRASIL, 2018). Outras bases previstas na LGPD, embora existentes em abstrato, dificilmente se mostram compatíveis com a realidade do monitoramento laboral. Por essa razão, a análise cuidadosa dessas quatro bases torna-se etapa indispensável para avaliar se determinada prática de controle e vigilância pode ser considerada lícita e compatível com o regime de proteção de dados.

5.1 Consentimento

Define a LGPD que o consentimento cedido pelo titular, para ser considerado legítimo, deve ser obtido por meio de uma manifestação livre, informada e inequívoca do titular, sempre de acordo com uma finalidade pré-determinada (art. 5º, XIII). Em termos práticos, isso significa que o consentimento deve ser: (i) livre, ou seja, dado sem coação ou condicionamentos indevidos; (ii) informado, pois o titular precisa compreender claramente para que seus dados serão utilizados e quais as consequências desse tratamento; e (iii) inequívoco, porque deve resultar de uma manifestação positiva e clara de vontade, não podendo ser presumido nem obtido por meio de silêncio ou condutas ambíguas.

Quando observamos relações empregatícias, contudo, tais conceitos devem ser observados com um maior nível de cuidado. Isto porque a relação empregatícia é, acima de tudo, uma relação de poder, em que o direito de direção do empregador está atrelado à condução do negócio em questão. Tal poder, chamado de "Poder Diretivo" por Zavanella (2023), é materializado na prática a partir do momento em que o empregador dá ordens ao empregado para cumprir ações. Ademais, enquanto legalmente subordinado ao empregador, o empregado que não cumpre com tais ordens, está sob o risco de ter esse vínculo rescindido.

Em outras palavras, é o empregado quem se encontra na posição de sujeição à capacidade de ação do empregador e aos efeitos por ele pretendidos. O poder, nesse contexto, funciona como um instrumento que orienta e condiciona comportamentos. Esse poder não se limita à força física, à coerção ou à violência direta; ao contrário, manifesta-se de múltiplas formas, muitas vezes sutis, estruturais e simbólicas, que não se confundem com imposição violenta, mas que ainda assim influenciam de maneira significativa a conduta do trabalhador (Zavanella, 2023).

É justamente por conta deste poder, que o consentimento encontra obstáculos em relações de trabalho. Analisando novamente o definido pela LGPD, o consentimento do empregado deve ser dado voluntariamente e para uma finalidade específica, com o empregador fornecendo informações sobre quais dados são coletados, sua finalidade de uso, por quanto tempo serão armazenados, quem terá acesso a eles e se serão transmitidos a terceiros (Doneda, 2015). Mas e quando o tratamento de dados não for do interesse do empregado?

Diante da assimetria de poder, é natural que o empregado, temendo represálias por parte do empregado, consinta com o tratamento, mesmo quando suas vontades vão em sentido oposto. Assim, a exigência de que o consentimento seja verdadeiramente "livre" deve ser observada com especial rigor nas relações de trabalho, sob pena de esvaziar-se seu próprio sentido.

No contexto laboral, a dependência econômica e a posição hierárquica do empregado tendem a comprometer a espontaneidade da manifestação de vontade, transformando o consentimento, muitas vezes, em mera formalidade. Assim, a simples assinatura de cláusulas contratuais ou a aceitação genérica de políticas internas não são suficientes, por si só, para caracterizar um consentimento válido nos termos da LGPD, sobretudo quando não há alternativa real ao empregado senão concordar para preservar seu vínculo de emprego. Isso reforça a compreensão, já consolidada na doutrina e em orientações das autoridades de proteção de dados, de que o consentimento deve ser utilizado com cautela no âmbito das relações trabalhistas, privilegiando-se, sempre que possível, outras bases legais mais compatíveis com esse cenário de desequilíbrio estrutural (Zavanella, 2023).

Por tal motivo, o consentimento é uma base legal contraindicada para tratamentos "relevantes" dentro da relação trabalhista, sobretudo para monitoramentos, justamente porque dificilmente estará presente a liberdade necessária para que a manifestação de vontade seja considerada válida. Em situações nas quais o trabalhador percebe que a recusa pode gerar consequências negativas (ainda que implícitas), como desconfiança, prejuízo à avaliação de desempenho ou até risco ao próprio vínculo empregatício, o consentimento deixa de ser expressão de autonomia e passa a refletir apenas a assimetria da relação. Nesses casos, a utilização dessa base legal não apenas fragiliza a proteção do titular, como também expõe o empregador a riscos jurídicos, uma vez que o tratamento poderá ser considerado irregular.

Obviamente, existem possibilidades de coleta do consentimento sem influência da subordinação, como o caso da participação voluntária em programas corporativos de bem-estar. Para tais casos, a adesão é opcional, a recusa não gera qualquer consequência na relação de trabalho, e os benefícios oferecidos configuram vantagens adicionais, e não

condições para manutenção do emprego. Nesse cenário, a decisão do empregado de consentir é efetivamente livre, pois a negativa não afeta sua segurança no trabalho, sua progressão profissional ou sua rotina laboral, o que assegura a validade do consentimento e afasta a influência do desequilíbrio de poder característico da relação empregatícia (Zavanella, 2023).

Para fins de monitoramento laboral, o cenário é distinto. Como o tratamento de dados se vincula ao poder diretivo do empregador, a liberdade de escolha do empregado é reduzida. Por isso, a doutrina e as autoridades de proteção de dados entendem que o consentimento raramente atende ao requisito de liberdade da LGPD, devendo ser utilizado com cautela e, em regra, substituído por outras bases legais.

Ainda assim, caso se opte por essa base legal, o empregador deve adotar cautelas para preservar, na maior medida possível, a liberdade do titular. Isso inclui assegurar que o consentimento possa ser negado ou revogado sem prejuízo ao vínculo de emprego e delimitar claramente a finalidade do tratamento. Mesmo com essas medidas, permanece o entendimento de que se trata de uma base frágil para legitimar monitoramentos contínuos, o que reforça a preferência por fundamentos jurídicos mais adequados.

5.2 Legítimo interesse

Mundialmente, o legítimo interesse figura entre as bases legais mais invocadas para justificar práticas de monitoramento no contexto laboral, por oferecer maior flexibilidade ao controlador (GDPR Local, 2024). Nos termos da LGPD, contudo, aplica-se apenas ao tratamento de dados pessoais comuns e exige a avaliação da compatibilidade entre os interesses do empregador e os direitos fundamentais do trabalhador. Não se trata, portanto, de autorização genérica, mas de base condicionada a um teste estruturado de balanceamento.

O artigo 10 da LGPD estabelece parâmetros objetivos para essa avaliação, exigindo do controlador a realização de um verdadeiro teste de balanceamento. A primeira etapa consiste em identificar se há uma finalidade legítima que justifique o tratamento. Esse interesse deve ser lícito, concreto e compatível com o ordenamento jurídico, não podendo contrariar direitos fundamentais nem servir de pretexto para práticas abusivas (Vainzof, 2022). No contexto do monitoramento laboral, exemplos de finalidades potencialmente legítimas incluem a proteção do patrimônio da empresa, a segurança da informação, a prevenção de fraudes ou o cumprimento de deveres legais. Ainda assim, a finalidade precisa ser devidamente delimitada e justificada, não bastando alegações genéricas de "controle" ou "produtividade".

Superada essa fase, impõe-se a análise da necessidade do tratamento, conforme previsto no §1º do art. 10 da LGPD. Aqui, o empregador deve demonstrar que os dados efetivamente coletados são indispensáveis para alcançar a finalidade pretendida. Esse exercício envolve questionamentos práticos relevantes: o mesmo resultado poderia ser atingido com menor volume de dados? Seria possível utilizar técnicas menos invasivas? Existiria outra base legal mais adequada para sustentar o tratamento? Se houver alternativas menos intrusivas, a utilização do legítimo interesse perde força e pode ser considerada desproporcional.

A terceira etapa corresponde ao núcleo do teste: a ponderação entre os interesses do controlador e os direitos e expectativas legítimas do titular. É nesse momento que se avalia o impacto concreto do monitoramento sobre a esfera pessoal do trabalhador, sua autonomia, sua privacidade e sua dignidade. Deve-se considerar se o empregado pode razoavelmente esperar aquele tipo de tratamento no contexto da relação de trabalho, bem como os riscos associados, incluindo potenciais efeitos discriminatórios, constrangimentos indevidos ou restrições ex-

cessivas à liberdade individual. Tal avaliação deve ser realizada por meio do Legitimate Interest Assessment (LIA). Monitoramentos intensivos, opacos ou intrusivos tendem a desequilibrar essa balança em desfavor do controlador (Zavanella, 2023).

Por fim, mesmo quando o legítimo interesse é considerado aplicável, a LGPD exige que o tratamento seja acompanhado de salvaguardas adicionais. Isso inclui o dever de transparência, a adoção de medidas de mitigação de riscos e a implementação de práticas de accountability. No contexto laboral, isso se traduz na necessidade de políticas internas claras, informação adequada aos empregados, limitação de acessos aos dados coletados, definição de prazos de retenção e mecanismos de revisão periódica das práticas adotadas. Sem essas cautelas, o uso do legítimo interesse deixa de ser um instrumento de equilíbrio e passa a representar uma via de legitimação formal de práticas potencialmente abusivas.

Assim, embora o legítimo interesse possa, em muitos casos, ser a base legal mais adequada para justificar determinadas formas de monitoramento, sua aplicação exige rigor metodológico e responsabilidade. Trata-se menos de uma escolha discricionária do empregador e mais de um dever de demonstrar, de forma consistente e verificável, que o tratamento de dados é necessário, proporcional e respeitoso aos direitos fundamentais do trabalhador.

5.3 Cumprimento de obrigação legal

A base legal do cumprimento de obrigação legal ou regulatória ocupa posição própria no debate sobre monitoramento laboral, pois desloca o foco da vontade do empregador para uma imposição do ordenamento jurídico. Nesses casos, o tratamento de dados não resulta de escolha da empresa, mas da necessidade de cumprir deveres previstos em lei ou em normas regulatórias. Como aponta Zavanella (2023), essa base pode ser utilizada quando a legislação exige mecanismos de monitoramento, como em situações relacionadas à segurança no trabalho, à proteção de informações confidenciais ou ao cumprimento de regras específicas de determinadas atividades. Nessas hipóteses, não se exige consentimento do empregado, mas permanece o dever de transparência quanto aos dados coletados e às finalidades do tratamento.

A própria legislação trabalhista brasileira oferece um exemplo clássico dessa situação. O artigo 74 da CLT impõe às empresas com mais de vinte empregados o dever de manter registro fidedigno dos horários de entrada e saída, configurando uma obrigação legal de tratamento de dados pessoais. Ferramentas digitais de controle de ponto, como as utilizadas por empresas como a Mywork, que coletam dados de geolocalização e registro fotográfico para comprovar a jornada, podem se enquadrar nessa base legal quando utilizadas estritamente para cumprir essa exigência normativa (Pulga et al., 2025). Nesses casos, o tratamento não se fundamenta no consentimento ou no legítimo interesse, mas no cumprimento de obrigação legal, o que também delimita o tipo e a extensão dos dados coletados. Assim, quando esses mecanismos são utilizados estritamente para documentar o tempo de trabalho, atendendo à obrigação prevista na CLT, o enquadramento adequado é o da obrigação legal.

O ponto decisivo é a finalidade concreta: o uso de ferramentas para avaliação comportamental ou vigilância por webcam exige nova base jurídica e teste de proporcionalidade, por extrapolar o mero dever legal. Assim, a obrigação legal não legitima qualquer vigilância, devendo o tratamento vincular-se estritamente ao comando normativo e ao princípio da necessidade. Utilizar bases objetivas como pretexto para práticas invasivas desvirtua o monitoramento e afronta a lógica protetiva da LGPD.

5.4 Execução de contrato de trabalho

A base legal da execução de contrato também pode fundamentar determinadas práticas de tratamento de dados pessoais no contexto laboral, desde que exista um vínculo direto e necessário entre o tratamento e o cumprimento das obrigações assumidas pelas partes no contrato de trabalho. Nos termos da LGPD, essa hipótese autoriza o tratamento quando ele for indispensável para viabilizar a execução do contrato ou para atender a procedimentos preliminares solicitados pelo próprio titular. Assim, quando a coleta e o uso de dados são efetivamente essenciais para que o empregador e o empregado cumpram aquilo que foi pactuado, não se exige o consentimento do trabalhador.

No âmbito do monitoramento laboral, essa base pode ser invocada, por exemplo, em situações nas quais o próprio contrato prevê deveres objetivos relacionados à jornada, à produtividade ou à entrega de resultados. Se o contrato estabelece a obrigação de cumprir determinado horário, registrar presença, executar tarefas em plataformas específicas ou atingir metas mensuráveis, é legítimo que o empregador trate dados estritamente necessários para verificar o cumprimento dessas obrigações. Nesses casos, o monitoramento encontra fundamento direto no conteúdo contratual e não depende da invocação de outras bases.

Ainda assim, a execução do contrato não constitui autorização genérica para a vigilância, devendo o tratamento limitar-se ao proporcional e necessário para o cumprimento das cláusulas pactuadas. Mecanismos invasivos são injustificáveis quando meios menos intrusivos atingem o mesmo resultado, preservando a lógica da necessidade e adequação contra controles indevidos. Ademais, essa base legal não ampara o tratamento de dados sensíveis, como biometria ou de saúde, o processamento de tais informações exige fundamento jurídico específico na LGPD, sob pena de violação direta ao seu regime protetivo reforçado.

6. Conclusão

O presente trabalho buscou demonstrar que o monitoramento laboral, embora historicamente vinculado ao exercício legítimo do poder diretivo do empregador, assume contornos qualitativamente distintos na contemporaneidade em razão da digitalização do trabalho e da sofisticação das tecnologias de vigilância. A passagem de mecanismos pontuais de controle para sistemas contínuos, automatizados e altamente intrusivos transforma o monitoramento em um fenômeno que ultrapassa a mera organização da atividade produtiva e passa a impactar diretamente direitos fundamentais do trabalhador, em especial a privacidade, podendo aumentar, inclusive, a disparidade hierárquica entre empregado-empregador. Ainda, cria quase uma forma de disciplinamento do empregador que pouco pode ser questionada.

Ao longo do trabalho, evidenciou-se que os dados de telemetria e os rastros comportamentais coletados por ferramentas de monitoramento não são meros elementos técnicos, mas verdadeiros dados pessoais, capazes de individualizar, perfilar e influenciar decisões relevantes sobre a vida profissional do titular. Essa constatação é decisiva, pois submete tais práticas de forma inequívoca ao regime jurídico da LGPD.

A análise dos princípios da LGPD permitiu concluir que o monitoramento somente pode ser considerado legítimo quando orientado por finalidade específica e legítima, limitado ao mínimo necessário, estruturado de forma adequada ao contexto e conduzido sem gerar discriminações. A transparência e o livre acesso mostraram-se, nesse cenário, não apenas como requisitos formais, mas como condições para reduzir a assimetria informacional entre empregador e empregado e viabilizar o exercício efetivo dos direitos do titular. Práticas ocultas ou mal explicadas de vigilância tendem não apenas a violar a legislação de proteção de dados, mas também a comprometer a confiança e a própria legitimidade da gestão organizacional.

No campo das bases legais, o consentimento revela-se estruturalmente frágil nas relações de trabalho, diante da assimetria de poder e da dependência econômica do empregado, tornando-se, em regra, inadequado para justificar monitoramentos. O legítimo interesse e a execução do contrato podem constituir fundamentos mais plausíveis em determinadas hipóteses, desde que submetidos a rigoroso teste de proporcionalidade. Já o cumprimento de obrigação legal aplica-se apenas quando o tratamento decorre diretamente de imposição normativa, não podendo ser ampliado para legitimar formas expansivas de vigilância.

Conclui-se, portanto, que a vigilância tecnológica no ambiente de trabalho somente pode ser considerada legítima quando concebida dentro de um modelo de governança orientado por direitos fundamentais. Isso exige do empregador uma postura ativa de responsabilidade: mapear riscos, justificar escolhas tecnológicas, limitar a coleta ao estritamente necessário, garantir transparência real e assegurar mecanismos efetivos de controle por parte do titular. A LGPD, nesse contexto, não atua como entrave ao exercício do poder diretivo, mas como instrumento indispensável de reequilíbrio da relação empregatícia em um cenário de crescente dataficação do trabalho.

Referências

ALBON, Victoria; BEEBY, Sarah. Discrimination in surveillance? UK People Reward and Mobility Hub, 2025. Disponível em: https://www.ukemploymenthub.com/discrimination-in-surveillance/. Acesso em: 16 jan. 2026.

AJUNWA, Ifeoma. Ifeoma Ajunwa on the limitless boundaries of employee surveillance. **Digital Data Design Institute at Harvard**, 2021. Disponível em: https://d3.harvard.edu/ifeoma-ajunwa-on-the-limitless-boundaries-of-employee-surveillance. Acesso em: 16 jan. 2026

ALKIMIN, T. A. (2007). O controle do trabalho do empregado: do direito ao monitoramento à intimidade do trabalhador. São Paulo: **LTr.**

AMARAL, B. dos S.; MOREIRA, J. P. Garantia de proteção de dados frente ao trabalho remoto. **REVISTA DELOS**, [S. l.], v. 17, n. 61, p. e2915, 2024. DOI: 10.55905/rdelosv17.n61-170. Disponível em: https://ojs.revistadelos.com/ojs/index.php/delos/article/view/2915. Acesso em: 16 jan. 2026.

BBC NEWS BRASIL. Reportagem sobre demissões no Itaú e monitoramento de empregados. **BBC News Brasil**, 2025. Disponível em: https://www.bbc.com/portuguese/articles/c8xrqj2492wo. Acesso em: 16 jan. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais (LGPD).** Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 16 jan. 2025.

BRASIL. Decreto-Lei nº 5.452, de 1º de maio de 1943. **Consolidação das Leis do Trabalho (CLT).** Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del5452.htm. Acesso em: 16 jan. 2025.

CPA. Justiça do Trabalho condena empresas com base na LGPD. **Portal CPA**, 2023. Disponível em: https://netcpa.com.br/colunas/noticia-justica-do-trabalho-condena-empresas-com-base-na-lgpd/17060. Acesso em: 16 jan. 2026.

DA SILVA, A. P.; PULGA, C. B.; PASQUALETO, O. de Q. F. Data Protection and AI-Based Employee Monitoring: Legal Boundaries under Brazil's LGPD. **Beijing Law Review**, v. 16, p. 2442-2479, 2025. Acesso em: 16 jan. 2026.

DELGADO, M. (2019). Curso de direito do trabalho. São Paulo: **LTr**.

DONEDA, Danilo. Princípios de Proteção de Dados Pessoais. In: Lucca, Newton de; SIMÃO FILHO. Direito & Internet III: Marco Civil da Internet. **Quartir Latin**, 2015. T. I. p. 378.

FERREIRA, Victória. O tratamento de dados pessoais no contrato de trabalho. **Revista do Tribunal Superior do Trabalho**, [S. l.], v. 89, n. 3, p. 184–200, 2023. DOI: 10.70405/rtst.v89i3.10. Disponível em: https://revista.tst.jus.br/rtst/article/view/10. Acesso em: 16 jan. 2026.

G1. Como funcionam programas que monitoram funcionários no home office. **G1**, 2025. Disponível em: https://g1.globo.com/tecnologia/noticia/2025/09/12/monitoramento-de-funcionarios-coleta-dados-sobre-mouse-localizacao-e-mais.ghtml . Acesso em: 16 jan. 2026.

GDPR LOCAL. GDPR. Employee Monitoring: Compliance Considerations for Employers. Disponível em: https://gdprlocal.com/gdpr-employee-monitoring/. Acesso em: 16 mar. 2026.

MARTOS, José Antônio de Faria; SANTOS, Rafael Augusto dos; BARUFI, Renato Britto. Direito à desconexão e teletrabalho: entre a flexibilidade e os riscos da hiperconexão = The right to disconnect and teleworking: between flexibility and the risks of hyper connection. **Revista do Tribunal Superior do Trabalho**, Porto Alegre, v. 91, n. 1, p. 149-163, jan./mar. 2025. Acesso em: 16 jan. 2026.

MEIKSINS, Peter F. Scientific Management and Class Relations: A Dissenting View. **Theory and Society**, v. 13, n. 2, p. 177–209, 1984. Acesso em: 16 jan. 2026.

JACOB NETO, Elias. "Vigilância ou surveillance? Proposta para começar a compreender corretamente este fenômeno". Direito e Novas Tecnologias. **Florianópolis: CONPEDI**, 2014. p. 517 - 539. Disponível em: http://www.publicadireito.com.br/artigos/?cod=51c9d0385c088de7. Acesso em 16 de jan. 2026.

MIZIARA, Raphael. A proteção contra despedida algorítmica no contexto laboral: consequências pelo descumprimento do direito à explicabilidade previsto no art. 20 da LGPD. **Revista do Tribunal Superior do Trabalho**, [S. l.], v. 90, n. 1, p. 230–249, 2024. DOI: 10.70405/rtst.v90i1.44. Disponível em: https://revista.tst.jus.br/rtst/article/view/44. Acesso em: 16 jan. 2026.

PANNUNZIO, Pedro. Demissões no Itaú: programa que vigia funcionários foi usado sem aviso e monitorava teclado, cliques e navegação. **Intercept Brasil**, 2025. Disponível em: https://www.intercept.com.br/2025/09/14/demissoes-no-itau-programa-que-vigia-funcionarios-foi-usado-sem-aviso-e-monitorava-teclado-cliques-e-navegacao/. Acesso em: 16 jan. 2026.

TENA, Lucimara Plaza; SIQUEIRA, Dirceu Pereira; MORAIS, Fausto Santos de. Captação de dados pessoais pelo estado e o direito à privacidade em tempos de pandemia. **Revista Brasileira de Direitos Fundamentais & Justiça**, [S. l.], v. 14, n. 43, p. 407–425, 2021. DOI: 10.30899/dfj.v14i43.1022. Disponível em: https://dfj.emnuvens.com.br/dfj/article/view/1022. Acesso em: 16 jan. 2026.

Tribunal Superior do Trabalho. Recurso de Revista nº RR-21162-51.2015.5.04.0014, **DEJT** de 28 ago. 2020. Decisão que reconheceu a licitude do monitoramento de empregados por meio de câmeras em áreas coletivas de trabalho. Disponível em: https://www.tst.jus.br/web/guest/-/fiscaliza%C3%A7%C3%A3o-de-empregados-por-meio-de-c%C3%A2meras-em-locais-coletivos-%C3%A9-considerada-l%C3%ADcita. Acesso em: 16 jan. 2026.

TUNG, Irene; SONN, Paul K.; PINTO, Maya; DWORAK-FISCHER, Sally; BOXERMAN, Josh. When 'Bossware' Manages Workers: A Policy Agenda to Stop Digital Surveillance and Automated-Decision-System Abuses. **National Employment Law Project – NELP**, julho de 2025. Disponível em https://www.nelp.org/insights-research/when-bossware-manages-workers-digital-surveillance-automated-decision-system-abuses/.

VAINZOF, Rony. Capítulo I: Disposições Preliminares. In: MALDONADO, Viviane Nóbrega; BLUM, Renato Opice. **LGPD: Lei Geral de Proteção de Dados Pessoais Comentada**. 4. ed. São Paulo: Thomson Reuters Brasil, 2022.

ZAVANELLA, Fabiano. Aplicação da Lei Geral de Proteção de Dados nas relações trabalhistas: limites do consentimento do empregado e do legítimo interesse do empregador. 2023. Tese (Doutorado em Direito do Trabalho e da Seguridade Social) - **Faculdade de Direito, Universidade de São Paulo**, São Paulo, 2023. doi:10.11606/T.2.2023.tde-06032024-120308. Acesso em: 16 jan. 2026.

Capítulo II
Direitos & Biometria

O SUJEITO EM FRENTE ÀS CÂMERAS: DESAFIOS DO RECONHECIMENTO FACIAL DIANTE DA DIVERSIDADE HUMANA

Rafael Francisco França[1]

Resumo: a crescente implementação de tecnologias de reconhecimento facial em múltiplos setores da sociedade tem gerado intensos debates focados nos aparatos tecnológicos e nos seus enquadramentos normativos. Este artigo propõe uma inversão de perspectiva, concentrando a análise no titular dos dados biométricos. O objetivo central é investigar as vulnerabilidades e as falhas sistêmicas que emergem quando tais tecnologias são aplicadas a grupos específicos, cujas características desafiam os pressupostos de universalidade, unicidade e estabilidade dos algoritmos. O estudo examina as dificuldades no reconhecimento de crianças, em virtude das alterações craniofaciais do crescimento, e de idosos, devido aos efeitos do envelhecimento nos tecidos e na estrutura óssea. A análise demonstra que as falhas de reconhecimento são consequências inerentes a uma tecnologia que, ao operar com base em padrões, marginaliza a diversidade. Conclui-se que a avaliação da legitimidade do uso do reconhecimento facial demanda análise apro-

1. Mestre em Ciências Criminais pela PUCRS (2014), Doutor em Estudos Estratégicos Internacionais pela UFRGS (2018), doutorando em Direito pela NOVA School of Law (Lisboa/PT) e em Integração Contemporânea da América Latina (UNILA). Delegado de Polícia Federal.

fundada sobre quem são os sujeitos a serem reconhecidos, evidenciando a necessidade de uma abordagem jurídica mais crítica e restritiva.

Palavras-chave: Reconhecimento Facial; Biometria; Vulnerabilidade Biométrica; Envelhecimento Facial; Unicidade Biométrica; Dignidade da Pessoa Humana; Proteção de Dados Pessoais.

Abstract: The increasing implementation of facial recognition technologies in multiple sectors of society has generated intense debates focused on technological devices and their normative frameworks. This article proposes a shift in perspective, concentrating the analysis on the holder of biometric data. The central objective is to investigate the vulnerabilities and systemic failures that emerge when such technologies are applied to specific groups whose characteristics challenge the assumptions of universality, uniqueness, and stability of algorithms. The study examines the difficulties in recognizing children, due to craniofacial changes during growth, and the elderly, due to the effects of aging on tissues and bone structure. The analysis demonstrates that recognition failures are inherent consequences of a technology that, by operating based on patterns, marginalizes diversity. It concludes that evaluating the legitimacy of the use of facial recognition demands in-depth analysis of who the subjects to be recognized are, highlighting the need for a more critical and restrictive legal approach.

Keywords: Facial Recognition; Biometrics; Biometric Vulnerability; Facial Aging; Biometric Uniqueness; Human Dignity; Personal Data Protection.

1. Introdução

A tecnologia de reconhecimento facial (FRT, do termo anglicizado "Facial Recognition Technology") consolidou-se como um dos mais proeminentes e controversos instrumentos de vigilância e identificação da era digital. Sua proliferação em espaços públicos e privados, impulsionada por promessas de segurança e eficiência, suscita debates jurídicos e éticos de grande complexidade, que frequentemente se concentram na arquitetura dos sistemas, na capacidade de vigilância em massa e na adequação aos marcos regulatórios de proteção de dados.

Contudo, uma análise exaustiva da matéria não pode se limitar ao aparato tecnológico ou ao seu potencial abstrato de vigilância. Torna-se imperativo deslocar o foco analítico da câmera para o rosto, ou seja, do meio técnico para o titular dos dados biométricos, cuja singularidade e diversidade representam o desafio fundamental para a validade e a justiça de tais sistemas.

Este trabalho parte da premissa de que a eficácia e a neutralidade da FRT são questionadas quando confrontadas com a pluralidade das características humanas, considerando-se que a face, enquanto dado biométrico, não é uma constante imutável, sendo suscetível a transformações decorrentes de fatores biológicos, ambientais e sociais. A pretensa universalidade dos algoritmos de reconhecimento colide com a realidade de grupos populacionais cujas feições desviam das normas majoritariamente utilizadas para o treinamento dos sistemas, resultando em taxas de erro desproporcionais e, consequentemente, em formas de exclusão e discriminação.

A metodologia empregada neste artigo fundamenta-se em abordagem jurídico-crítica e teórica, utilizando o método dedutivo para analisar as implicações da FRT. De tal modo, a pesquisa é de natureza qualitativa e bibliográfica, estruturada a partir da revisão de literatura

especializada das ciências computacionais, textos normativos, além de estudos de caso e dados experimentais de performance algorítmica. Assim, o percurso analítico propõe uma inversão de perspectiva ao deslocar o foco do aparato tecnológico (o sensor) para o titular dos dados biométricos (o rosto), permitindo análise das vulnerabilidades humanas diante da automação.

Em tal contexto, o problema central reside em falhas sistêmicas e no potencial discriminatório inerentes à tecnologia de reconhecimento facial, que opera sob pressupostos de universalidade e estabilidade que a diversidade humana constantemente desafia. Diante disso, o artigo busca responder à seguinte pergunta: em que medida as características biológicas de grupos específicos comprometem a legitimidade, a eficácia e a justiça da implementação dessa tecnologia em larga escala?

A contribuição deste trabalho reside em demonstrar que as falhas de reconhecimento não são meras anomalias técnicas passageiras, mas consequências de um modelo que marginaliza o que desvia do padrão. Ao evidenciar que o reconhecimento facial pode falhar sistematicamente com sujeitos vulneráveis ou em processo de transformação, propõe-se substrato teórico para o necessário debate jurídico, defendendo a necessidade de uma abordagem mais restritiva e cautelosa. Como contribuição principal, esse estudo reforça que a proteção da dignidade humana e da não discriminação deve prevalecer sobre promessas abstratas de eficiência e segurança pública.

2. O tempo e o resto: o fator idade no reconhecimento facial

Ao se comentar sobre tecnologias de reconhecimento facial, é preciso ter em conta que não se trata somente de que tipo de sensor ou de equipamento está sendo empregado. É preciso considerar, antes e

porém, peculiaridades dos titulares dos dados biométricos que serão captados, o que sem dúvida exercerá forte influência na tomada de qualquer decisão a partir da análise dessas informações. Atente-se para o fato de que, para que se tenha condições ideais de captação dos traços faciais, faz-se necessária a captação de frente, ou seja, que o rosto do indivíduo esteja voltado para a câmera que vai coletar os dados. Nesse contexto, estudiosos das ciências computacionais admitem que, dentre as fases do processo de reconhecimento facial, a parte da captação dos dados é a mais desafiadora (Yu et al., 2024).

De todo modo, também é importante apontar que a tecnologia de reconhecimento facial é considerada como modo passivo de coleta de dados biométricos (Pato; Millett, 2010, p. 65). Dessa forma, não é preciso que haja interação direta e proposital com os sensores na grande maioria dos casos. Todavia, é requisito básico de funcionamento dos sistemas que os titulares das informações virem o rosto para o local em que está a câmera, dizendo-se, por isso, que para que haja identificação ou verificação é, sim, necessário certo modo cooperativo por parte das pessoas. Observa-se, quanto ao que foi acima apontado, que a voluntariedade e o modo pelo qual as imagens faciais foram captadas para a formação dos modelos a serem comparados exercem importante influência no resultado e na acurácia dos sistemas contendo reconhecimento facial (Introna; Nissenbaum, 2009).

Por tal razão, é preciso considerar as nuances das legislações sobre o assunto, mormente levando em conta a publicação do EU AI Act e o RGPD na União Europeia, textos que vêm servindo de base aos projetos de lei brasileiros acerca do tema. Analisando tal demanda, é possível atestar que, de todo modo, os fornecedores de tais sistemas devem atentar e garantir que, se projetados e destinados a interagir com pessoas, estes têm que possibilitar que os usuários consigam tomar conhecimento de tal situação, ou seja, que estão tendo seus dados bio-

métricos coletados e que, assim, estão interagindo com a inteligência artificial, isso sob a ótica de um indivíduo medianamente bem-informado, considerando-se para tanto o contexto de utilização dos sistemas e as circunstâncias do uso. Ainda, é necessário que todas estas informações estejam à disposição dos usuários, de forma clara e inteligível, até a ocorrência da primeira interação com o sistema de IA (Karassawa; Bigas; Serraglio, 2024, p. 153; Krastins; Almeida, 2024, p. 169).

No entanto, observa-se que há características peculiares de certos titulares de dados pessoais (traços faciais) e até mesmo consequências do processamento das informações que podem acarretar erros no sistema de reconhecimento. Assevera-se que as imagens das faces humanas podem variar considerando vários fatores, como, por exemplo, postura da cabeça na captação, estilo usado em relação aos cabelos pelo titular dos dados, uso de barba, aplicação de cosméticos, isso sem contar as expressões naturais (riso, olhos cerrados, boca aberta etc.), o volume do rosto, incidência de idade e muitas outras variantes (Guo; Kennedy, 2023, p. 402). Assim, considera-se a existência de indivíduos que possuem expressão facial anômala, seja em decorrência de acidentes ou traumas, seja em virtude de eventos congênitos. Da mesma forma, e conforme será abordado no subitem 3.1, há pessoas que possuem bastante semelhança em consideração aos detalhes da face, o que pode chegar ao extremo no caso de gêmeos univitelinos. Ainda, e passando para o processamento das informações biométricas captadas, é possível também ocorrer perda de qualidade nos dados a serem analisados durante a conversão das imagens para representação binária, principalmente no que tange a detalhes importantes para identificação/autenticação dos titulares[2] (Pato & Millett, 2010, p. 30).

2. Quanto a este ponto, importa ressaltar que há sistemas de reconhecimento que, por características de sua utilização, somente serão empregados em relação a número restrito de titulares de dados pessoais, como por exemplo, em controles de acesso de um prédio ou de uma empresa privada. Pato, J. N., & Millett, L. I. (2010). Biometric recognition: challenges and opportunities. p. 31. No entanto, esse não é caso dos estudos aqui desenvolvidos, em que, a princípio, os sensores estão instalados em locais públicos e/ou semipúblicos.

Em relação aos titulares de dados pessoais que têm seus traços faciais captados e analisados por FRT, os mais comuns são os clientes-alvos dos sensores, ou seja, quem realmente pode ser reconhecido durante o processo em tela, pessoas normais cujos traços podem ou não ser identificados ou autenticados após os processos tecnológicos. Há, ainda, aqueles que passam pelos sistemas de reconhecimento facial ou qualquer outro método de identificação biométrica e que tentam burlar ou enganar a tecnologia (Raposo, 2024, p. 1859). Dentre estes, destacam-se duas classes de fraudadores. Os primeiros são os impostores, quais sejam, os que não devem ser reconhecidos, mas tentam fazê-lo de todas as maneiras. Os segundos, os fraudadores ou dissimuladores de identidade, os quais, pelas vias normais do emprego do método, deveriam ser reconhecidos durante os procedimentos, mas tentam de muitas formas não o ser[3] (Pato; Millett, 2010, p. 48–49).

Assim, aponta-se que, por exemplo, as modificações nos traços faciais de crianças são diferentes das encontradas no rosto de adultos em

3. Esse é, por exemplo, o caso de Eduardo Miranda de Souza, conhecido como "Cara Quadrada", o qual foi detido em São Paulo/SP ao tentar passar por mais uma cirurgia plástica em 2024, já tendo realizado alterações nos olhos, nas orelhas e nas bochechas para tentar despistar os sensores de reconhecimento facial espalhados pelo Estado de São Paulo. Terra. (2024, abril 2). Quem é Cara Quadrada preso ao tentar fazer cirurgia plástica para despistar a polícia. Portal Terra.https://www.terra.com.br/noticias/brasil/cidades/quem-e-cara-quadrada-preso-ao-tentar-fazer-cirurgia-plastica-para-despistar-a-policia,2c12a17189b6fa8ebb2b8d18a51404d2b82254tk.html, acesso em 07/09/2025.

decorrência da ação do tempo (Ramanathan; Chellappa, 2006). Por isso, de antemão, impõem-se a avaliação correta sobre quem pode ser alvo da tecnologia, detalhando ao máximo características especiais de diversos grupos (Serec; Dall'Agnese, 2024, p. 284).

Dessa maneira, e a seguir, o texto traz o fator "tempo" para os titulares dos dados pessoais. Aqui, os rostos são o objetivo principal.

2.1 Crianças

A aplicação da tecnologia de reconhecimento facial em crianças é bastante controversa para alguns de seus fins, ainda mais se for considerado que estão muito mais expostas a terem suas informações biométricas captadas e analisadas por que frequentam mais espaços públicos (Lynch; Gordon; Campbell, 2023). Assim, o uso de métodos para identificação biométrica de bebês, recém-nascidos ou não, e crianças é já tema de estudo desde há muito tempo. Começando com a coleta de impressões digitais, o interesse parte sempre da necessidade maior de se saber quem é, por onde passou, a quem pertence e, principalmente, onde está tal pessoa.

O modo de realizar o reconhecimento facial em crianças e em futuros adultos é fortemente influenciado por vários fatores, dentre os quais o crescimento do crânio e da modificação dos tecidos do rosto, o que é incrementado pelas modificações crânioencefálicas, o que inclusive trazem em seu bojo a chegada da dentição permanente (Ricanek Jr et. al., 2015, p. 17). São vários os exemplos em que tais fatores foram levados em consideração para a aplicação da tecnologia.

Desse modo, e somente como um dos vários exemplos, cita-se um dos casos emblemáticos. Em 2009, um menino chamado Gui Hao de apenas três anos de idade foi sequestrado na China, na Província de Sichuan, enquanto brincava em frente à loja de seus pais. Pelo que consta, a criança foi vendida pelos sequestradores para outra família que residia na Província de Guangdong, distante aproximadamente 1.500 quilômetros do local onde morava. Mesmo que a polícia chinesa tenha conseguido prender o líder da associação criminosa responsável por este e por outros sequestros de crianças em 2014, os policiais afirmaram ter sido impossível localizar todas as vítimas, eis que a aparência dos menores teria mudado muito com o passar dos anos, tanto que nem mesmo os pais seriam capazes de reconhecê-los.

Aqui, o uso de inteligência artificial em reconhecimento facial merece destaque: em 2019, um sistema contendo FRT conseguiu localizar Gui Hao, já com treze anos, ao processar a imagem contida no banco de dados quando ele ainda era bebê. O sistema conseguiu projetar seus traços e formar imagem que representou como estaria com o passar dos anos. Ao final, o menino voltou para casa. Interessante notar que não se trata de caso isolado, haja vista que, em 2023, outras duas crianças retornaram para casa após terem sido abduzidas passados mais de vinte anos (Harari, 2024).

O uso da tecnologia de reconhecimento facial para aplicação no contexto de crianças e adolescentes apresenta um cenário complexo, permeado por desafios técnicos e considerações éticas (Best-Rowden, et. al., 2016). Embora prometa otimizar processos e aumentar a segurança, a identificação facial de menores é intrinsecamente dificultada pelas transformações físicas características de seu período de crescimento (Barrett, 2020). Ainda segundo Barret (2020), os desafios inerentes ao uso de tecnologias de reconhecimento facial em crianças devem fomentar a discussão no contexto mais amplo dos perigos que essa tecnologia representa para toda a população. Argumenta-se, para tanto, que, embora a vulnerabilidade particular das crianças possa levá-las a parecerem candidatas naturais a proteções reforçadas (Kamble; Dale, 2021), os danos que o reconhecimento facial causa a esse grupo

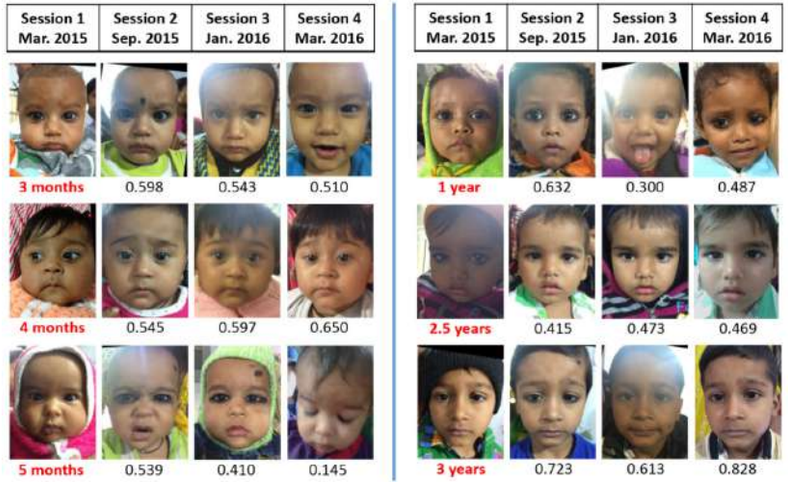
etário compartilham similaridades com os prejuízos sofridos por outros grupos demográficos e pela população em geral. Desse modo, Barret enfatiza que as inexatidões dos sistemas de reconhecimento facial em relação a crianças, assim como a outros grupos como pessoas com pele mais escura, mulheres e idosos, são um dos principais desafios, com o potencial de levar a discriminação em qualquer contexto de implantação (2020, p. 275).

De qualquer forma, indica-se que as tecnologias de reconhecimento facial têm demonstrado um desempenho menos preciso para crianças, o que as coloca em risco quando esses sistemas são utilizados por autoridades policiais ou sistemas de segurança escolar (Barrett, 2020, p. 226). Essa imprecisão, similar à observada em outros grupos demográficos, pode levar a identificações errôneas e potenciais acusações injustas. Além disso, Barret (2020) aponta para o fato de que as modificações físicas inerentes ao crescimento infantil representam um obstáculo significativo para a acurácia desses sistemas ao longo do tempo. Em tal cenário, a rápida evolução das feições de crianças e adolescentes dificulta a criação de sistemas de reconhecimento facial consistentemente eficazes, levantando sérias preocupações sobre a confiabilidade de tais tecnologias quando aplicadas a menores.

Uma das principais barreiras para o reconhecimento facial eficaz nesse grupo etário reside nas modificações ósseas significativas que ocorrem durante a fase de desenvolvimento (Best-Rowden et. al., 2016). Diferentemente do envelhecimento facial em adultos, que se manifesta predominantemente através de alterações nos tecidos moles, a infância e a adolescência são marcadas por um intenso crescimento craniofacial, alterando a forma e as proporções da face de maneira substancial (Ricanek Jr. et al., 2015, p. 16). A taxa de crescimento craniofacial é mais acentuada nos primeiros anos de vida, especialmente entre o nascimento e os cinco anos de idade. Essa rápida evolução

morfológica implica em mudanças constantes nas características faciais, tornando o reconhecimento facial de crianças pequenas ao longo do tempo um problema particularmente desafiador para as tecnologias existentes (Ricanek Jr. et al., 2015, p. 17). Ademais, o crescimento facial não segue uma trajetória linear, sendo influenciado por fatores como a puberdade e o desenvolvimento da dentição permanente, que podem desencadear saltos nas taxas de crescimento. Essas variações no ritmo do desenvolvimento ósseo introduzem uma camada adicional de complexidade para os sistemas de reconhecimento facial que buscam manter a acurácia ao longo do tempo (Ricanek Jr. et al., 2015, p. 17). Abaixo, somente como exemplo, variações de faces em tal seara:

Figura 1 Crianças (NITL)



Imagens faciais longitudinais de seis crianças do banco de dados de imagens faciais de Recém-Nascidos, Bebês e Crianças Pequenas (NITL), coletadas durante quatro sessões diferentes (março de 2015, setembro de 2015, janeiro de 2016 e março de 2016) no

Hospital Saran Ashram, Dayalbagh, Índia. A idade de cada criança na primeira aquisição é apresentada em vermelho, e as pontuações de comparação de um software comercial de última geração entre a primeira imagem e todas as imagens subsequentes são apresentadas em preto. Fonte: Best-Rowden et. al., 2016

Apesar dessas mudanças estruturais, o padrão de crescimento facial em crianças geralmente preserva certas semelhanças com as características originais, em um fenômeno conhecido como crescimento gnomônico (Ricanek Jr. et al., 2015, p. 17). No entanto, as alterações nas proporções entre os diferentes elementos da face, impulsionadas pelo crescimento ósseo subjacente, representam um obstáculo significativo para a identificação consistente.

Outro desafio reside nas limitações de desenvolvimento das crianças em relação à compreensão dos riscos de privacidade e das implicações da vigilância. Indica-se que os jovens têm menos controle sobre onde vão e o que fazem, e avaliações imprecisas de suas faces podem ter um impacto particularmente forte em suas vidas, especialmente em contextos como o uso pela lei (Barrett, 2020, p. 226). Além disso, o efeito inibidor dessas tecnologias sobre a liberdade de expressão pode restringir o desenvolvimento emocional e intelectual das crianças. Neste mesmo contexto, a falta de plena compreensão dos riscos por parte dos jovens os torna ainda mais vulneráveis à coleta e uso indevido de seus dados faciais, mesmo que se considere a utilização da tecnologia para fins diversos da segurança pública. Indica-se que, quando o sistema de reconhecimento facial não dispuser de mecanismos para atendimento a crianças, deve-se abrir exceção ou solução alternativa para sua passagem ou acesso (Andrejevic; Selwyn, 2020), evitando-se qualquer constrangimento gerado nessa situação.

Em relação ao sistema legal brasileiro, e embora haja o já mencionado desejo de mimetizar a legislação da União Europeia no que diz respeito ao uso do reconhecimento facial em espaços acessíveis ao público, é necessário verificar que o ordenamento jurídico do Brasil traz agora novos contornos para a captação de dados biométricos de crianças e adolescentes (Almeida, 2022, p. 271). Isso porque o Estatuto da Criança e do Adolescente (Lei nº 8.069/1990) e a própria LGPD vedam a coleta e o processamento de dados biométricos desse grupo em especial (Tajra, 2024).

A proposta acima indicada foi fortemente influenciada pelo recente advento da Lei nº 15.211/2025, a qual trouxe novas nuances ao Estatuto Digital da Criança e do Adolescente, sendo batizado de ECA Digital. Observa-se que, em seu artigo 9º, tal legislação estabelece que fornecedores de serviços digitais devem adotar mecanismos confiáveis de verificação de idade, vedando a mera autodeclaração (Brasil, 2026). Para cumprir essa obrigação, muitas plataformas estão recorrendo a tecnologias de reconhecimento facial para estimativa de idade ou verificação documental por vídeo selfie (Sellan, Melo, 2026). Essa exigência, portanto, cria dilema ético e técnico: para proteger a criança de conteúdos nocivos, o sistema exige a entrega de um dado biométrico sensível. Por conseguinte, o desafio para a nova ANPD e para os agentes de tratamento é garantir que essa verificação ocorra sem a criação de bancos de dados biométricos permanentes. O ECA Digital proíbe expressamente o uso de dados coletados para verificação etária para qualquer outra finalidade, como monetização ou perfilamento comportamental.

Considerando o acima exposto, os sistemas de reconhecimento facial também têm sido considerados como ferramentas potenciais para auxiliar na localização de crianças desaparecidas, levando em conta que, dentre os dados pessoais biométricos, a face é o mais propício a este desiderato (Kamble; Dale, 2021, 2018). A capacidade de identificar

rapidamente indivíduos em grandes multidões poderia, teoricamente, agilizar o processo de busca (Ricanek Jr. et al., 2015, p. 15). Contudo, a eficácia dessa aplicação é diretamente impactada pelas dificuldades inerentes ao reconhecimento de faces em constante mudança devido ao crescimento ósseo.

As tecnologias de reconhecimento facial contemporâneas, maioritariamente desenvolvidas e treinadas com dados de faces adultas, frequentemente demonstram uma acurácia inferior quando aplicadas à identificação de crianças. Esse viés negativo em relação ao reconhecimento infantil é um reflexo da inadequação dos modelos em lidar com as dinâmicas específicas do crescimento facial (Srinivas et al., 2019). Considera-se que a precisão da identificação facial é fortemente dependente da idade do sujeito, com indivíduos mais velhos sendo mais facilmente reconhecidos e distinguidos (Srinivas et al., 2019).

Em contrapartida, a identificação de crianças, cujas feições estão em um estado de transformação contínua, apresenta um nível de dificuldade significativamente maior. Para mitigar esses desafios, esforços têm sido direcionados para a criação de bancos de dados faciais específicos para crianças, visando fornecer um conjunto de dados mais representativo para o treinamento e a avaliação de algoritmos de reconhecimento. A disponibilidade de dados longitudinais, que acompanham o desenvolvimento facial dos indivíduos ao longo do tempo, é elemento-chave para o aprimoramento da performance dos sistemas.

De tal forma, o emprego de algoritmos baseados em redes neurais profundas (DNN) supera limitações de iluminação, ângulos e envelhecimento facial, perfazendo-se adaptação contextual para situações diversas. Eles convertem rostos em vetores matemáticos, comparando-os com bancos de dados mesmo em condições adversas, sendo essa a principal vantagem sobre os sistemas comuns. O Amazon Rekognition, sistema desenvolvido pela empresa Amazon, por exemplo, promete utilizar essa técnica para buscar crianças desaparecidas, com precisão de 99% em correspondências, adaptando-se a mudanças físicas progressivas (AWS, 2024). Nesse mesmo sentido, há quem aponte para a necessidade de elaboração de modelos em 3D para que seja possível simular as modificações que podem ocorrer nas faces de crianças com o passar do tempo. Em tal situação, o uso de tais modelos em detrimento dos que somente se baseiam em 2D é recomendado, considerando-se em princípio que as alterações faciais nas pessoas ocorrem em três, e não duas, dimensões. Ainda, importante mencionar que o emprego de modelos em 3D pode compensar alterações ambientais, tais como mudanças de expressões e intensidade de luz (Park; Tong; Jain, 2008, 2010).

Em suma, os desafios do reconhecimento facial de crianças são multifacetados, abrangendo desde as dificuldades técnicas relacionadas à acurácia e ao desenvolvimento físico até as vulnerabilidades inerentes à sua imaturidade e limitada autonomia. Por isso, assevera-se que esses desafios, juntamente com os possíveis danos quase universais e compartilhados da tecnologia de reconhecimento facial, reforçam a necessidade de uma proibição abrangente do seu uso, em vez de soluções limitadas ou específicas para crianças. A severidade dos danos para as crianças, portanto, serve como um argumento adicional e contundente para a implementação de uma proibição que proteja a todos dos riscos dessa tecnologia.

Desse modo, a necessidade de sistemas capazes de manter a acurácia ao longo do tempo, especialmente em aplicações sensíveis como a localização de crianças desaparecidas, demanda um investimento contínuo em pesquisa e desenvolvimento, com uma compreensão aprofundada das particularidades do desenvolvimento facial infantil.

2.2 Idosos

A aplicação da tecnologia de reconhecimento facial a esta faixa etária específica apresenta desafios singulares, primariamente devido às alterações físicas e fisiológicas inerentes ao processo de envelhecimento (Modi; Patel, 2021). É notório que a face humana possui importantes características que podem ser aproveitadas por sistemas de identificação/autenticação, apresentando traços que permitem distinguir uma pessoa de outra. No entanto, é também a biometria que mais sofre com o processo de envelhecimento, afetando com rigor a desenvoltura dos sistemas que contêm FRT (Boussaad; Boucetta, 2022, p. 2.975). A seguir, novamente como exemplo apenas, a evolução do processo de envelhecimento no rosto de um homem ao longo dos anos:

Figura 2 Exemplo de modificações na face masculina com o envelhecimento. Fonte: Boussaad; Boucetta, 2022, p.2.977

Como exposto acima, um dos principais desafios reside nas modificações da textura da pele, como o surgimento de rugas e linhas de expressão, que se intensificam com o passar dos anos. Essas alterações podem levar a inconsistências entre as características faciais armazenadas nos modelos de reconhecimento e as capturadas em momentos posteriores, comprometendo a precisão da identificação (Albert et. al., 2011). Além das mudanças na textura, estudos demonstram que o envelhecimento também acarreta alterações na forma facial, incluindo ganho ou perda de peso, e modificações na estrutura óssea craniofacial. Essas transformações sutis, mas progressivas, podem alterar o contorno geral do rosto, dificultando a correspondência com os dados biométricos originais (Albert et. al., 2011, p. 89–90; Ling et al., 2007).

Os estudos citados no parágrafo anterior apontam diversos efeitos do natural envelhecimento que são desafios aos sistemas de FRT. Assim, a redução da elasticidade da pele é fator significativo que influencia a eficácia do reconhecimento facial em idosos. A perda de firmeza da pele pode levar à flacidez e ao deslocamento de tecidos moles, alterando a aparência de feições importantes para a identificação (Bruce; Young, 2016, p. 24–25). Além disso, a perda de tônus muscular e volume nos tecidos faciais também contribui para as mudanças na aparência com a idade. Essa atrofia muscular pode afetar a definição de características faciais e a expressividade, tornando o reconhecimento mais complexo (Albert et. al., 2011, p. 92). As alterações ósseas, resultantes do processo contínuo de remodelação óssea, podem igualmente impactar a morfologia facial ao longo do tempo. A seguir, mais um exemplo:

Figura 3 Quadro demonstrativo da evolução do processo de envelhecimento facial. Fonte: Lanitis, 2010, p. 144.

Essas mudanças estruturais, embora pequenas, podem influenciar a forma geral da cabeça e do rosto, afetando a precisão do reconhecimento facial. No terço superior da face, a queda das sobrancelhas e a flacidez das pálpebras são alterações comuns associadas ao envelhecimento. A perda de elasticidade da pele na região superior dos olhos pode criar dobras e excesso de pele, modificando significativamente a aparência da área dos olhos, crucial para o reconhecimento facial. Na região média da face, o desenvolvimento do sulco nasolabial e a diminuição do depósito de gordura nas bochechas são características do envelhecimento (Patterson et al., 2007). Da mesma forma, a profundidade do sulco nasolabial aumenta com a idade, enquanto a perda de gordura nas bochechas pode levar a uma aparência mais encovada, afetando os algoritmos de reconhecimento facial. No terço inferior da face, a formação de rugas verticais acima do lábio superior e o afinamento dos lábios são mudanças típicas do envelhecimento (Patterson et al., 2007).

Nesse cenário, o alongamento e o adelgaçamento dos lábios, juntamente com o desenvolvimento de sulcos nos cantos da boca, podem alterar significativamente a aparência da região inferior do rosto. O surgimento da papada (jowls) e a deposição de gordura abaixo da linha da mandíbula, levando a um queixo mais flácido, são também desafios para o reconhecimento facial em idosos (Greenbaum, 2019). Essas alterações na região inferior da face, resultantes da perda de volume, do deslocamento de gordura e de mudanças na mandíbula, podem dificultar a identificação, embora haja quem aponte que, com o uso de IA e de novas técnicas de identificação por ela proporcionadas, os sistemas acabam por utilizar formatos 2D de partes do rosto que não se modificam tanto com o passar dos anos (Greenbaum, 2019). Mesmo assim, por exemplo, as mudanças dentárias e na arcada dentária, que podem ocorrer com a idade, também podem influenciar sutilmente o formato facial. Embora muitas dessas alterações sejam pequenas, podem contribuir para a dificuldade no reconhecimento facial ao longo do tempo (Albert et. al., 2011).

Ademais, as variações diurnas na aparência facial, como a intensificação de rugas ao longo do dia devido à diminuição do inchaço facial matinal, podem introduzir mais um nível de complexidade para os sistemas de reconhecimento facial aplicados a idosos (Lanitis, 2010). Além das alterações físicas diretas, condições médicas comuns em idosos, como cataratas, glaucoma e artrite, podem afetar as características biométricas faciais, da íris e das impressões digitais, respetivamente, impactando indiretamente a precisão dos sistemas de reconhecimento (Rebera; Guihen, 2012). Além disso, e pela evolução natural das pessoas em vida societária, a sequência de coleta de dados biométricos das faces de idosos cai bastante, haja vista que, para diversos fins, como profissionais e escolares, por exemplo, estes procedimentos costumam ocorrer em outras fases do percurso. Isso faz com que as bases de dados para comparação sejam diminuídas em número e em qualidade ao longo da terceira idade (Rebera; Guihen, 2012), dificultando ainda mais a formação de modelos para comparação em termos de reconhecimento facial para idosos.

Para superar esses desafios, diversas técnicas avançadas de reconhecimento facial têm sido desenvolvidas (Modi; Patel, 2021). Uma abordagem concentra-se na extração de características faciais que sejam mais resistentes às mudanças relacionadas à idade, buscando assinaturas faciais que permaneçam relativamente estáveis ao longo do tempo. Outra técnica promissora envolve o uso de modelos generativos de envelhecimento, capazes de simular as transformações faciais que ocorrem com o passar dos anos (Mortezaie; Hassanpour, 2019). Esses modelos podem ser utilizados para normalizar imagens faciais, reduzindo as variações devido à idade antes do processo de reconhecimento.

Assim, métodos discriminativos também são empregados, com foco na identificação de padrões que discriminam indivíduos, independentemente das mudanças na aparência causadas pelo envelhecimento. Esses métodos buscam aprender representações faciais que sejam invariantes à idade (Akhtar et al., 2013). O rápido avanço do aprendizado profundo e das redes neurais convolucionais (CNNs) tem permitido o desenvolvimento de sistemas de reconhecimento facial mais robustos ao envelhecimento (Li; Hu; Yip, 2018). O treinamento dessas redes com grandes conjuntos de dados de rostos de diferentes idades ajuda a aprender características discriminativas menos suscetíveis às alterações relacionadas ao tempo. De tal maneira, abordagens de aprendizado conjunto que aprendem simultaneamente representações de características e funções de métrica de distância também têm se mostrado eficazes para o reconhecimento facial invariante à idade. Essas técnicas buscam otimizar a extração de características e a medida de similaridade para melhorar a precisão do reconhecimento em diferentes faixas etárias. Por outro lado, técnicas de envelhecimento e desenvolvimento faciais sintéticos, baseadas em modelos de aparência ativa (AAMs), também têm sido exploradas para simular os efeitos do envelhecimento e auxiliar na criação de sistemas de reconhecimento facial mais robustos à passagem do tempo (Mittal; Patel, 2023). A ideia é usar imagens sintéticas para aumentar os conjuntos de treinamento e melhorar a capacidade dos algoritmos de reconhecer rostos de idosos.

A análise de componentes faciais, como a forma dos olhos, nariz e boca, e o rastreamento de suas mudanças ao longo do tempo, também são abordagens utilizadas (Ramanathan; Chellappa; Biswas, 2009). Ao focar em partes específicas do rosto que podem envelhecer de maneira mais previsível ou menos variável, os sistemas de reconhecimento podem se tornar mais precisos. Por isso, é preciso reconhecer que o uso de reconhecimento facial para fins de identificação e autenticação pode se tornar desafiador se for considerado o tempo decorrido entre a formação do "template" e a submissão aos sensores para nova avaliação (Boussaad; Boucetta, 2022, p. 2.975). A principal preocupação em tais casos é exatamente a mudança provocada pela natural ação do tempo sobre os traços faciais das pessoas, ou seja, as modificações nos rostos com o passar dos anos (Pato & Millett, 2010, p. 32).

De acordo Lanitis (2010), o envelhecimento dos traços faciais não pode ser controlado pelos sistemas, pelo que não é possível eliminar as variações por ele natural ou artificialmente causadas durante a captação das fotos para formação de bancos de dados ou para identificação/verificação. Ainda, foi dito que fatores biológicos e também fatores externos podem exercer influências nos rostos das pessoas, o que pode levar a modificações relevantes para fins de reconhecimento facial. Os primeiros, biológicos, têm relação com o gênero, com ascendência, genética e até mesmo doenças enfrentadas pelas pessoas. Os segundos, externos, têm como exemplos o uso de cigarros, consumo de álcool, exposição ao sol ou a intempéries, fatores emocionais como o estresse e, ainda, ganho ou perda repentina de peso. Todos estes eventos podem exercer influências e acelerar o processo de envelhecimento da face (Lanitis, 2010).

De todo modo, o uso de reconhecimento facial em idosos é um campo em evolução que enfrenta desafios significativos devido às complexas e variadas mudanças faciais associadas ao envelhecimento. No entanto, o desenvolvimento de técnicas avançadas, como a extração de características robustas à idade, modelos de envelhecimento sintéticos, aprendizado profundo e análise de componentes faciais, oferece um caminho promissor para a criação de sistemas eficazes e confiáveis para esta importante parcela da população.

3. Situações especiais. A diversidade e a natureza do rosto humano em contexto

A implementação de tecnologias de reconhecimento facial por entes públicos e privados tem sido acompanhada por uma crescente preocupação teórica e jurídica acerca da sua eficácia frente à diversidade morfológica e identitária humana. Embora apresentados como instrumentos de otimização da segurança e gestão de fluxos, esses sistemas operam sob uma lógica de padronização que frequentemente negligencia variações fenotípicas atípicas.

Nota-se, portanto, que a "facialização" da sociedade, ao institucionalizar biometrias que falham em reconhecer a pluralidade da face humana, pode exacerbar desigualdades históricas, transformando uma ferramenta de segurança em um vetor de marginalização e controle discriminatório.

3.1 Gêmeos idênticos

Conforme já citado neste trabalho, um sistema de reconhecimento com o uso de biometria deve conter uma série de requisitos que o torne viável. Assim, dentre os mais fundamentais, encontram-se a universalidade, significando que todas as pessoas devem ter a característica que servirá para distingui-las, a estabilidade ou permanência, pela qual tal característica tem de ser perene, não podendo desaparecer com o tempo, e unicidade, indicando que não pode haver duas pessoas que contenham ou que portem a mesma característica.

No entanto, considerando o que foi apontado pelo parágrafo acima, e quando se tratar de gêmeos monozigóticos, o quesito unicidade já não pode ser facilmente preenchido em se falando de reconhecimento facial (Afaneh; Noroozi; Toygar, 2017). Quando se trata de identificar ou autenticar padrões faciais de gêmeos idênticos, os sistemas contendo FRT encaram um de seus maiores desafios[4] (Li et al., 2024; Pato; Millett, 2010, p. 29; Phillips et al., 2011). Embora haja pessoas que sejam naturalmente parecidas entre si, a reprodução dos traços biométricos entre irmãos univitelinos ou monozigóticos pode possibilitar que um se passe pelo outro, mesmo involuntariamente (Afaneh et al., 2017; Raposo, 2024, p. 1859), até mesmo para, por exemplo, desbloquear o acesso a aparelhos celulares que usam o reconhecimento facial como modo de segurança (Dion, 2019). Abaixo, parte do desafio aqui proposto:

Figura 4 gêmeas idênticas em fotos batidas em ambientes controlados. Fonte: Phillips et al., 2011.

Em um estudo levado a contento entre agosto de 2009 e agosto de 2010 (intervalo de coleta das amostras) (Phillips et al., 2011), diversos pares de gêmeos idênticos foram submetidos à coleta e verificação de

4. Quanto a este cenário da tecnologia de reconhecimento facial, interessante indicar que há desafios também em relação às semelhanças entre pais e filhos em alguns casos, nada comparável aos problemas gerados por gêmeos idênticos. Li, S. Z., & Jain, A. K. (2011). Introduction. Em S. Z. Li & A. K. Jain (Eds.), (2.ª ed., pp. 01–18). Springer. https://doi.org/10.1007/978-0-85729-932-1, acesso em 16/08/2025.

identidades a partir de seus traços faciais. As fotos, tiradas em ambientes controlados e não controlados, sem expressão facial ou sorrindo, e com o intervalo de um ano entre as coletas, revelaram que o sistema empregado alcançou sua melhor performance quando as fotografias comparadas foram tiradas no mesmo dia, em ambiente controlado, sendo que os gêmeos mantiveram suas expressões naturais, sem estar sorrindo, portanto. Da mesma forma, foi também concluído que questões de gênero não exerceram influências sobre os resultados, mas o tempo decorrido entre as fotos, sim. Por fim, o estudo indicou que é mais fácil distinguir gêmeos idênticos acima de 40 anos de idade que abaixo dessa faixa etária.

A performance dos sistemas contendo FRT na distinção de gêmeos é influenciada por diversos fatores. Alguns estudos sugerem que a precisão tende a melhorar em pares de gêmeos mais velhos, um fenômeno atribuído a mudanças epigenéticas que se manifestam em características faciais sutis ao longo da vida (Mccauley et al., 2021). Inversamente, a precisão diminui drasticamente em condições de imagem não ideais, como iluminação ambiente variável, expressões faciais não neutras e diferentes ângulos de câmera, que são a norma em espaços públicos (Mccauley et al., 2021). A dependência de condições ideais de captura de imagem torna a aplicação da FRT em contextos do mundo real, como a segurança pública, inerentemente arriscada e questionável do ponto de vista da sua fiabilidade probatória.

Curiosamente, a capacidade humana de distinguir gêmeos, muitas vezes recorrendo à identificação de marcas faciais distintivas como sinais, cicatrizes ou pequenas assimetrias, frequentemente supera os algoritmos atuais (Biswas; Bowyer; Flynn, 2011). Isto sugere que os sistemas de TRF ainda não conseguem capturar a totalidade das microcaracterísticas que compõem a unicidade de uma face, mesmo entre indivíduos geneticamente idênticos.

No entanto, estudos demonstram que mesmo os melhores algoritmos comerciais apresentam taxas de erro até dez vezes maiores quando expostos a gêmeos idênticos em comparação a não-gêmeos (Paone et al., 2014; Sun et al., 2010). Somente como exemplo, cita-se a coleta de dados biométricos de faces pela base Twins Days, com 17,486 imagens capturadas no mesmo dia, sendo que tal atividade revelou taxas de erros entre 4.1% e 17.4% em relação a gêmeos, contra 0.2% a 2.4% em pessoas comuns (Paone et al., 2014). Diga-se de passagem, que, no cenário brasileiro, onde projetos-piloto de smart cities começam a monitorar fluxos em estádios, semelhantes margens podem implicar centenas de falsos positivos em grandes eventos. Ademais, é necessário considerar que nenhum sistema de reconhecimento facial atual atinge acurácia forense suficiente para diferenciar gêmeos em condições de iluminação de rua, manutenção de expressões variadas e resoluções típicas de CFTV (circuito fechado de televisão)[5] (Mccauley et al., 2021).

3.2 Pessoas com parte do rosto coberta

A implementação de sistemas de reconhecimento facial como ferramenta de controle de fluxos em espaços acessíveis ao público representa um avanço tecnológico com potencial para otimizar a segurança e a gestão de acesso (Tapsoba, 2023; Zhang et. al., 2020). No entanto, a eficácia e a equidade destas tecnologias são intrinsecamente desafiadas

5. Estudos multibiométricos sugerem que a fusão com iris ou impressão digital derruba a taxa de falsos positivos a praticamente zero, mas tais modalidades exigem cooperação do titular de direitos, ação inviável em monitoramento à distância. Vide Sun, Z., Paulino, A. A., Feng, J., Chai, Z., Tan, T., & Jain, A. K. (2010). A Study of Multibiometric Traits of Identical Twins. Biometric technology for human identification, 7667(VII), 283–294. https://doi.org/https://doi.org/10.1117/12.851369, acesso em 24/07/2025.

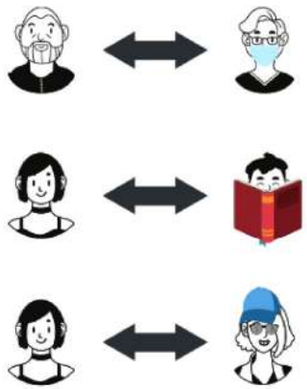


Figura 05 Ilustração de algumas oclusões que podem aumentar muito a dificuldade da tarefa de verificação facial (Fonte: Neto et al., 2022, p. 86223)

pela presença de pessoas que, por diversas razões culturais, de saúde ou religiosas, utilizam coberturas faciais que obscurecem parcial ou totalmente os seus rostos (Lynch, 2024, p. 06; Pato; Millett, 2010, p. 89; Ragashe; Goswami; Raghuwanshi, 2015). Esta realidade impõe uma análise aprofundada das limitações e das implicações da aplicação do reconhecimento facial neste contexto específico. O desafio central reside na própria mecânica do reconhecimento facial, que tradicionalmente se baseia na análise de características faciais distintivas, como o nariz, a boca e os olhos.

Quando estas características são obscurecidas por máscaras de proteção individual, véus religiosos como o hijab, o niqab ou a burca, ou outros tipos de coberturas faciais utilizadas por razões culturais ou de saúde, a capacidade dos algoritmos de reconhecimento facial de identificar e verificar indivíduos é significativamente comprometida (Maghari, 2021; Neto et al., 2022). Esta diminuição na precisão pode levar a falsos negativos, impedindo o acesso de indivíduos legítimos, ou a falsos positivos, comprometendo a segurança do sistema (Zhang et al., 2020, p. 40).

As razões para a utilização de coberturas faciais são multifacetadas. A pandemia de COVID-19 tornou o uso de máscaras de proteção uma prática comum em muitos espaços públicos, visando a proteção individual e coletiva contra a propagação do vírus. Do ponto de vista religioso e cultural, certas tradições prescrevem o uso de véus que cobrem diferentes porções do rosto, como o hijab, que geralmente cobre o cabelo e o pescoço, o niqab, que cobre o rosto deixando apenas os olhos visíveis, e a burca, que cobre todo o corpo, incluindo o rosto com uma malha para os olhos. Estas práticas são expressões de identidade religiosa e cultural e são protegidas por leis de liberdade religiosa em muitos ordenamentos jurídicos (Garcez; Moura, 2023).

Em se tratando de uso da tecnologia para identificação/autenticação de pessoas que, por fins diversos, cobrem partes da face ou dos traços detectáveis da cabeça, tais como cabelos, orelhas, olhos ou boca, considera-se interessante o uso de modelos adaptados para tanto. Quer isso dizer que, para programação e treinamento dos sistemas contendo FRT, é de bom alvitre já usar desde o início fotos cortadas, deixando de fora do espectro partes da face que costumam ser obliteradas por alguns (WEF, 2020, p. 11). Como exemplo, e considerando mulheres que utilizam hijab em seus afazeres diários, melhor que se considere o uso de sistemas que somente analisem o rosto, deixando de fora das comparações traços das orelhas e dos cabelos junto à face delas.

A literatura científica tem dedicado atenção significativa ao problema do reconhecimento facial sob oclusão, explorando diversas abordagens para mitigar o impacto das coberturas faciais no desempenho dos sistemas (Zou; You; Akashi, 2021). Algumas destas abordagens incluem a reconstrução das áreas ocluídas através da utilização de informações de áreas não ocluídas do rosto, explorando a simetria facial, ou através do treino de modelos com grandes conjuntos de dados contendo

imagens de rostos ocluídos, permitindo que a rede neural aprenda a extrair características discriminativas mesmo na presença de oclusões (Neto et al., 2022). Outras técnicas visam a detecção e a exclusão das regiões ocluídas durante o processo de reconhecimento, focando-se nas características faciais visíveis (Zeng; Veldhuis; Spreeuwers, 2021).

No âmbito do reconhecimento facial cross-spectral, a utilização de imagens térmicas em conjunto com imagens no espectro visível tem sido investigada como potencial solução para certos tipos de oclusões (Anghelone, 2023). Uma vez que a imagem térmica captura a emissão de calor dos objetos, ela pode, teoricamente, fornecer informações sobre a estrutura facial mesmo sob certas coberturas. No entanto, esta abordagem também enfrenta desafios, como a variação significativa na aparência facial entre diferentes espectros e a necessidade de bases de dados multiespectrais para o treino eficaz dos algoritmos.

Apesar dos avanços nas técnicas de reconhecimento facial sob oclusão, o desempenho destes sistemas geralmente ainda não atinge a precisão dos sistemas treinados e testados com rostos totalmente visíveis, especialmente em cenários não controlados (Neto et al., 2022). A variabilidade nas poses faciais, nas expressões, na iluminação e na qualidade da imagem, inerente a espaços públicos, adiciona camadas de complexidade ao problema. Além disso, a diversidade étnica e as diferentes formas e tamanhos de oclusões representam desafios adicionais para a generalização e a robustez dos modelos de reconhecimento facial (Zhang et. al., 2020).

A utilização do reconhecimento facial em espaços públicos, especialmente considerando indivíduos com rostos cobertos, levanta questões éticas e sociais importantes (Garcez & Moura, 2023). O direito à privacidade, a liberdade religiosa e a não discriminação são princípios fundamentais que devem ser considerados cuidadosamente ao implementar tais tecnologias. A imposição de sistemas de controle de fluxos

que penalizam ou dificultam a participação de indivíduos que utilizam coberturas faciais por razões legítimas pode ser vista como uma violação destes direitos (Fontes, 2017).

Em alguns contextos europeus, o debate sobre o uso do véu islâmico em espaços públicos e locais de trabalho ilustra a tensão entre a segurança pública, a laicidade e a liberdade religiosa. As decisões de tribunais europeus sobre a proibição do uso do véu islâmico em certas situações demonstram a complexidade de equilibrar diferentes direitos e interesses em sociedades multiculturais. A implementação de sistemas de reconhecimento facial em espaços públicos deve levar em consideração estas sensibilidades culturais e religiosas para evitar a exclusão e a discriminação.

Além das questões religiosas, as coberturas faciais utilizadas por razões de saúde, como máscaras, também devem ser consideradas. A implementação de sistemas de reconhecimento facial que exigem a remoção destas máscaras para o acesso a espaços públicos pode ser impraticável e contraproducente em contextos de saúde pública. É crucial que os sistemas de controle de fluxos sejam adaptáveis e respeitem as necessidades de saúde dos indivíduos.

As limitações inerentes aos sistemas de reconhecimento facial na presença de oclusões levantam questões sobre a sua adequação como ferramenta primária de controle de fluxos em espaços públicos onde a utilização de coberturas faciais é comum ou necessária. A dependência excessiva desta tecnologia pode levar a ineficiências, erros e potenciais injustiças. Portanto, é imperativo considerar abordagens complementares ou alternativas para a identificação e verificação de indivíduos em tais cenários. Alternativas como a utilização de outras modalidades biométricas (reconhecimento da íris, da voz, da marcha), a combinação de múltiplas modalidades (fusão de características faciais visíveis com outras informações biométricas) (Yaman; Eyiokur; Ekenel, 2022), ou

métodos de identificação não biométricos (cartões de acesso, códigos QR, reconhecimento manual por pessoal treinado) podem oferecer soluções mais robustas e inclusivas para o controlo de fluxos em espaços públicos (Z. Zhang et. al., 2020). A escolha da tecnologia mais apropriada deve ser informada por uma análise cuidadosa do contexto específico, dos requisitos de segurança, das considerações éticas e da diversidade da população que utiliza o espaço.

4. Considerações finais

A jornada analítica empreendida neste artigo buscou deslocar o epicentro do debate sobre o reconhecimento facial da tecnologia para o ser humano. Ao examinar as especificidades dos titulares dos dados, evidencia-se que a FRT é um sistema técnico cujas premissas são constantemente desafiadas pela multifacetada realidade da condição humana. As dificuldades encontradas no reconhecimento de diferentes grupos não representam meras exceções, mas indicativos de falhas estruturais que comprometem sua fiabilidade.

A análise demonstrou que as transformações naturais do ciclo da vida, como o desenvolvimento craniofacial em crianças e o envelhecimento em idosos, impõem barreiras significativas à precisão e à estabilidade temporal da identificação. O caso dos gêmeos idênticos ataca o pilar da unicidade, revelando a incapacidade dos sistemas atuais de atingir acurácia forense em cenários não controlados.

Em síntese, esse estudo aponta que o reconhecimento facial falha de maneira sistemática com aqueles que se afastam da norma, priorizando a padronização em detrimento da diversidade humana. Diante deste quadro, a questão que se impõe ao Direito é se a utilização da FRT em larga escala é compatível com os princípios da dignidade humana e da não discriminação.

Conclui-se que as evidências sugerem que os riscos de erro e exclusão são inerentes, demandando uma postura de máxima cautela e a consideração de moratórias ou proibições estritas para a salvaguarda da sociedade. Por fim, as discussões sobre a política da face e a acessibilidade em biometria reforçam a necessidade de um design inclusivo e de uma análise jurídica que proteja os titulares contra a exacerbação de exclusões históricas.

Referências

AFANEH, Ayman; NOROOZI, Fatemeh; TOYGAR, Önsen. Recognition of identical twins using fusion of various facial feature extractors. **Eurasip Journal on Image and Video Processing**, v. 2017, n. 1, 1 dez. 2017.

AKHTAR, Zahid et al. Face Recognition under Ageing Effect: A Comparative Analysis. **ICIAP**, p. 309–318, 2013.

ALBERT, Midori; SETHURAM, Amrutha; RICANEK, Karl. Implications of Adult Facial Aging on Biometrics. In: **Biometrics - Unique and Diverse Applications in Nature, Science, and Technology.** [S. l.]: InTech, 2011. p. 89–106.

ALMEIDA, Eloisa Machado de. Reconhecimento facial, vigilância e transparência. In: BRITO CRUZ, Francisco; SIMÃO, Bárbara (orgs.). **Direitos Fundamentais e Processo Penal na Era Digital:** Doutrina e Prática em Debate. São Paulo: InternetLab, 2022. v. V, p. 262–272.

ANDREJEVIC, Mark; SELWYN, Neil. Facial recognition technology in schools: critical questions and concerns. **Learning, Media and Technology**, v. 45, n. 2, p. 115–128, 2 abr. 2020.

ANGHELONE, David. Vision par ordinateur et apprentissage profond appliqués à la reconnaissance faciale dans le spectre invisible. 2023. Thèse (Doctorat en Informatique) — Université Côte d'Azur, 2 out. 2023.

AWS. Os fatos sobre a tecnologia de reconhecimento facial com inteligência artificial. 2024. Disponível em: https://aws.amazon.com/pt/rekognition/the-facts-on-facial-recognition-with-artificial-intelligence/. Acesso em: 14 abr. 2025.

BARRETT, Lindsey. Ban Facial Recognition Technologies for Children - And for Everyone Else. **Boston University Journal of Science and Technology Law**, v. 26, p. 223–285, 2020.

BEST-ROWDEN, Lacey; HOOLE, Yovahn; JAIN, Anil K. Automatic Face Recognition of Newborns, Infants, and Toddlers: A Longitudinal Evaluation. In: **IEEE**, set. 2016. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7736912. Acesso em: 8 jan. 2026.

BISWAS, Soma; BOWYER, Kevin W.; FLYNN, Patrick J. A Study of Face Recognition of Identical Twins by Humans. **IEEE International Workshop on Information Forensics and Security**, p. 1–6, 2011.

BOUSSAAD, Leila; BOUCETTA, Aldjia. Deep-learning based descriptors in application to aging problem in face recognition. **Journal of King Saud University - Computer and Information Sciences**, v. 34, n. 6, p. 2975–2981, 1 jun. 2022.

BRASIL. Ministério da Justiça e da Segurança Pública. S**ancionado projeto que fortalece a ANPD e fixa início da vigência do ECA Digital.** Governo Federal, 14 mar. 2026. Disponível em: https://www.gov.br/mj/pt-br/assuntos/noticias/sancionado-projeto-que-fortalece-a-anpd-e-fixa-inicio-da-vigencia-do-eca-digital. Acesso em: 15 mar. 2026.

BRUCE, Vicki; YOUNG, Andrew W. Face perception. New York: Routledge; Psychology Press, 2016.

DION, Astrid Priscilla. Social Implications of the Facial Recognition System. [S. l.: s. n.]. Disponível em: https://www.academia.edu/38867110/Social_Implications_of_the_Facial_Recognition_System. Acesso em: 7 set. 2025.

FONTES, Paulo Gustavo Guedes. Laicidade e proibição do véu islâmico na França. Disponível em: https://blog.grupogen.com.br/juridico/postagens/artigos/laicidade-e-proibicao-do-veu-islamico-na-franca/. Acesso em: 19 maio 2024.

GARCEZ, Marina Thaís Rodrigues; MOURA, Thobias Prado. A influência do Soft Law em decisões do Tribunal Europeu de Direitos Humanos e do Tribunal de Justiça da União Europeia sobre o véu islâmico. **FGV Rio de Janeiro Law School**, p. 87–97, 2023.

GREENBAUM, Dov. Where Everybody Knows Your Name: Facial Recognition Technology and Its Applications in Society. [S. l.: s. n.]. Disponível em: https://www.calcalistech.com/ctech/articles/0,7340,L-3758811,00.html. Acesso em: 24 ago. 2025.

GUO, Zhilong; KENNEDY, Lewis. Policing based on automatic facial recognition. **Artificial Intelligence and Law**, v. 31, n. 2, p. 397–443, 1 jun. 2023.

HARARI, Yuval Noah. Nexus: uma breve história das redes de informação, da Idade da Pedra à inteligência artificial. São Paulo: Companhia das Letras, 2024.

INTRONA, Lucas D.; NISSENBAUM, Helen. Facial Recognition Technology: A Survey of Policy and Implementation Issues. [S. l.: s. n.]. Disponível em: https://nissenbaum.tech.cornell.edu/papers/facial_recognition_report.pdf. Acesso em: 14 maio 2025.

KAMBLE, Vaishali; DALE, Manisha. Face recognition of children using AI classification approaches. In: **Institute of Electrical and Electronics Engineers Inc.**, 5 mar. 2021.

KAMBLE, Vaishali H.; DALE, Manisha P. Infants and Toddlers Biometric Recognition: A Review. **Asian Journal For Convergence In Technology (AJCT)**, 15 abr. 2018.

KARASSAWA, Gisele; BIGAS, Bruna; SERRAGLIO, Lorena Pretti. Sistemas de Inteligência Artificial de risco inaceitável. In: VAINZOF, Rony et al. (orgs.). **Comentários ao EU AI Act.** São Paulo: Thomson Reuters, 2024. p. 137–157.

KRASTINS, Alexandra; ALMEIDA, Eduarda Costa. IA de alto risco e medidas de governança. In: VAINZOF, Rony et al. (orgs.). **Comentários ao EU AI Act.** São Paulo: Thomson Reuters, 2024. p. 159–174.

LANITIS, Andreas. Facial Biometric Templates and Aging: Problems and Challenges for Artificial Intelligence. **AIAI Workshops**, p. 142–149, 2010.

LI, Haoxi; HU, Haifeng; YIP, Chitung. Age-Related Factor Guided Joint Task Modeling Convolutional Neural Network for Cross-Age Face Recognition. **IEEE Transactions on Information Forensics and Security**, v. 13, n. 9, p. 2383–2392, 1 set. 2018.

LI, Stan Z. et al. Handbook of Face Recognition. 3. ed. [S. l.]: Springer, 2024.

LING, Haibin et al. A Study of Face Recognition as People Age. In: **IEEE Computer Society**, 21 out. 2007. Disponível em: https://ieeexplore.ieee.org/document/4409069. Acesso em: 8 jan. 2026.

LYNCH, Nessa. Facial Recognition Technology in Policing and Security—Case Studies in Regulation. **Laws**, v. 13, n. 3, p. 1–14, 1 jun. 2024.

LYNCH, Nessa; GORDON, Faith; CAMPBELL, Liz. Facial recognition technology: The particular impacts on children. In: **Privacy, Technology, and The Criminal Process.** London: Routledge, 2023. p. 136–155.

MAGHARI, Ashraf Y. A. Recognition of partially occluded faces using regularized ICA. **Inverse Problems in Science and Engineering**, v. 29, n. 8, p. 1158–1177, 2021.

MCCAULEY, John et al. Identical Twins as a Facial Similarity Benchmark for Human Facial Recognition. In: BRÖMME, A. et al. (orgs.). Bonn: **BIOSIG**, 2021.

MITTAL, Sonia; PATEL, Sanskruti. Age Invariant Face Recognition Techniques: A Survey on the Recent Developments. **International Journal of Engineering Trends and Technology**, 1 maio 2023.

MODI, Prashant; PATEL, Sanjay. A State-of-the-Art Survey on Face Recognition Methods. **International Journal of Computer Vision and Image Processing**, v. 12, n. 1, p. 1–19, 5 nov. 2021.

NETO, Pedro C. et al. Beyond Masks: On the Generalization of Masked Face Recognition Models to Occluded Face Recognition. **IEEE Access**, v. 10, p. 86222–86233, ago. 2022.

PARK, Unsang; TONG, Yiying; JAIN, Anil K. Face rec ognition with temporal invariance: A 3D aging model. In: **IEEE**, 2008. Disponível em: https://www.researchgate.net/publication/224401085_Face_recognition_with_temporal_invariance_A_3D_aging_model. Acesso em: 25 set. 2024.

PARK, Unsang; TONG, Yiying; JAIN, Anil K. Age-invariant face recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 5, p. 947–954, 2010.

PATO, Joseph N.; MILLETT, Lynette I. Biometric recognition: challenges and opportunities. Washington D.C.: [s. n.]. Disponível em: https://nap.nationalacademies.org/catalog/12720/biometric-recognition-challenges-and-opportunities. Acesso em: 11 set. 2024.

PHILLIPS, P. Jonathon et al. Distinguishing Identical Twins by Face Recognition. In: Santa Barbara: **IEEE**, 2011. Disponível em: https://www3.nd.edu/~kwb/Phillips_EtAl_FG_2011.pdf. Acesso em: 18 set. 2024.

RAGASHE, Monu U.; GOSWAMI, M. M.; RAGHUWANSHI, M. M. Approach Towards Real Time Face Recognition In Streaming Video Under Partial Occlusion. In: **IEEE**, 2015. Disponível em: https://ieeexplore.ieee.org/document/7282394/. Acesso em: 29 mar. 2025.

RAMANATHAN, Narayanan; CHELLAPPA, Rama. Modeling Age Progression in Young Faces. **IEEE - Computer Society Conference on**, v. 1, p. 387–394, 2006.

RAPOSO, Vera Lúcia. When facial recognition does not 'recognise': erroneous identifications and resulting liabilities. **AI and Society**, 2023.

REBERA, Andrew P.; GUIHEN, Barry. Biometrics for an Ageing Society Societal and Ethical Factors in Biometrics and Ageing. In: Darmstadt: **IEEE**, 2012.

RICANEK JR., Karl; BHARDWAJ, Shivani; SODOMSKY, Michael. A Review of Face Recognition against Longitudinal Child Faces. **BIOSIG**, p. 15–76, 2015.

SELLAN, Piero; MELO, Leonardo. O reconhecimento facial para aferição de idade. **Migalhas**, 5 mar. 2026. Disponível em: https://www.migalhas.com.br/depeso/451124/estrategia-ou-dever-o-reconhecimento-facial-para-afericao-de-idade. Acesso em: 15 mar. 2026.

SEREC, Letícia De Conti; DALL'AGNESE, Luisa Abreu. Diretrizes para análise jurídica de Produtos de IA: Avaliação de Risco e Governança. In: FACHINETTI, Aline Fuke (org.). **Inteligência Artificial Responsável.** São Paulo: Thomson Reuters Brasil, 2024. p. 275–294.

SRINIVAS, Nisha et al. Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults. In: Long Beach: **IEEE** CVF, 2019.

SUN, Zhenan et al. A Study of Multibiometric Traits of Identical Twins. **Biometric technology for human identification**, v. 7667, n. VII, p. 283–294, 14 abr. 2010.

TAJRA, Alex. Reconhecimento facial de crianças em estádios fere LGPD e ECA. Disponível em: https://www.conjur.com.br/2024-ago-27/reconhecimento-facial-de-criancas-em-estadios-fere-lgpd-e-eca-diz-relatorio. Acesso em: 15 jan. 2025.

TAPSOBA, Aboubacar Salif. Conception d'un Système de Reconnaissance Faciale Masquée. 2023. Mémoire (Maîtrise en Mathématiques et Informatique

Appliquées) — Université du Québec à Trois-Rivières, jul. 2023.

WEF. White Paper A Framework for Responsible Limits on Facial Recognition Use Case: Flow Management. [S. l.: s. n.]. Disponível em: https://www.weforum.org. Acesso em: 15 mar. 2026.

YAMAN, Dogucan; EYIOKUR, Fevziye Irem; EKENEL, Hazım Kemal. Multimodal soft biometrics: combining ear and face biometrics for age and gender classification. **Multimedia Tools and Applications**, v. 81, n. 16, p. 22695–22713, 1 jul. 2022.

YU, Shiqi et al. Face Detection. In: LI, Stan Z.; JAIN, Anil K.; DENG, Jiankang (orgs.). **Handbook of Face Recognition.** 3. ed. [S. l.]: Springer, 2024. p. 103–135.

ZENG, Dan; VELDHUIS, Raymond; SPREEUWERS, Luuk. A survey of face recognition techniques under occlusion. **IET Biometrics**, v. 10, n. 6, p. 581–606, 1 nov. 2021.

ZHANG, Zhifeng et al. A survey on occluded face recognition. In: **Association for Computing Machinery**, 18 dez. 2020.

ZOU, Min; YOU, Mengbo; AKASHI, Takuya. Reconstruction of Partially Occluded Facial Image for Classification. **IEEJ Transactions on Electrical and Electronic Engineering**, v. 16, n. 4, p. 600–608, 1 abr. 2021.

Capítulo III
Tecnologia & Regulação

Trusted Execution Environments em ecossistemas de criptografia ponta a ponta: análise de ameaças e desafios regulatórios no Brasil

Thobias Prado Moura[1]
Nathalia Sautchuk Patricio[2]

Resumo: A adoção de Ambientes de Execução Confiável em infraestruturas de nuvem reconfigura a segurança da informação ao estender a tutela criptográfica para o momento do processamento. Este artigo examina as implicações dessa arquitetura e demonstra que a ortogonalidade entre ambientes seguros e criptografia de ponta a ponta transfere o eixo da confiança da posse matemática das chaves para a governança da Raiz de Confiança (Root of Trust), deslocando o foco da tensão regulatória da interceptação de tráfego para o controle sobre a infraestrutura de verificação e atualização de software. Argumenta-se que o processamento de mensagens em representações intermediárias dentro dos TEEs gera desafios para a qualificação jurídica do dado, uma vez que a informação existe em texto claro apenas de forma transitória em memória isolada, situando-se em uma zona de incerteza entre a

1. É Doutorando em Direito com Especialidade em Direito e Segurança pela Universidade Nova de Lisboa, em Portugal, com L.L.M pela FMP e pela Universidade de Lisboa e graduado em Direito pela UFU. thobias@pradomoura.com.br.
2. É Doutoranda em ciência da computação na Universidade de Ciências Aplicadas de Karlsruhe, na Alemanha, formada em Engenharia da Computação pela USP.

pseudonimização e a anonimização. Conclui-se que, no Brasil, a combinação entre ambiguidade na aplicação de princípios gerais a tecnologias específicas e um histórico de conflitos jurisdicionais em matéria de criptografia cria o risco de captura institucional da tecnologia, na medida em que o agente que controla as atualizações de software dos ambientes seguros pode, por meio de modificações no código executado, transformar silenciosamente a tecnologia de proteção da privacidade em instrumento de vigilância e coleta de provas pelo Estado.

Palavras-chave: Ambientes de Execução Confiável; Computação Confidencial; Governança de Algoritmos; Criptografia; Vigilância.

Abstract: The adoption of Trusted Execution Environments (TEEs) in cloud infrastructures reconfigures information security by extending cryptographic protection to the critical moment of processing. This article examines the implications of this architecture and demonstrates that the orthogonality between secure environments and end-to-end encryption shifts the axis of trust from the mathematical possession of keys to the governance of the Root of Trust, moving the focus of regulatory tension from traffic interception to the control over the software verification and update infrastructure. It is argued that the processing of messages into intermediate representations within TEEs poses challenges for the legal qualification of data, since information exists in plaintext only transiently in isolated memory, placing it in a zone of uncertainty between pseudonymization and anonymization. It is concluded that, in Brazil, the combination of ambiguity in applying general principles to specific technologies and a history of jurisdictional conflicts over encryption creates the risk of institutional capture of the technology, since the agent controlling the software updates of secure environments can, through modifications to the executed code, silently transform privacy-protecting technology into an instrument of state surveillance and evidence collection.

Keywords: Trusted Execution Environments; Confidential Computing; Algorithmic Governance; Encryption; Surveillance.

1. Introdução

A adoção de Ambientes de Execução Confiável (Trusted Execution Environments – TEEs) tem se consolidado como uma tendência nas plataformas modernas, especialmente com a popularização dos smartphones como principal meio de interação digital e acesso a serviços remotos (MUÑOZ et al., 2023, p. 1-2). Originalmente, os TEEs foram propostos como uma solução para proteger a confidencialidade do código em cenários nos quais a computação é terceirizada para operadores potencialmente não confiáveis. Com o avanço da terceirização do armazenamento e do processamento de dados para a nuvem, surgiram preocupações não apenas em relação à confidencialidade dos dados, mas também do próprio código executado nesses ambientes. Para responder a esses desafios, a computação confidencial, baseada em TEEs, passou a ser amplamente adotada nas infraestruturas de nuvem comerciais contemporâneas, oferecendo garantias adicionais de segurança e isolamento durante a execução (PUDDU et al., 2022, p. 1).

Partindo da necessidade de se preservar e reforçar as propriedades de confidencialidade e integridade de sistemas computacionais complexos, através de ambientes de processamento isolados, os TEEs existentes permitem uma grande variedade de funcionalidades e aplicações, tais como:

geração segura de credenciais, armazenamento seguro de chaves,

inicialização segura, verificação da integridade do kernel, periféricos e sensores confiáveis, pagamentos móveis usando emulação de elementos seguros, sistemas de proteção de conteúdo digital, serviços para gerenciar e emitir bilhetes online, mecanismos de autenticação de acesso ao armazenamento em nuvem, segurança de dispositivos IoT, dentre outros (MUÑOZ et al., 2023, p.2)[3].

Dessa forma, o artigo busca apresentar os TEEs, descrevendo inicialmente seu funcionamento técnico, bem como suas principais vantagens e limitações. Em seguida, discute os diferentes tipos de ataques direcionados a TEEs. A análise é aprofundada por meio de cenários hipotéticos, como o uso de TEEs em aplicações amplamente difundidas a fim de ilustrar riscos e impactos reais. O trabalho também aborda questões legais e sociais, examinando como a ortogonalidade entre TEEs e criptografia de ponta a ponta, quando inserida em um ordenamento jurídico que prioriza a "verdade real" na investigação criminal, pode dar ensejo a uma "vigilância paradoxal", especialmente com o esforço cada vez mais de Estados, especificamente o do Brasil de legislar e aplicar normas, mesmo quando isso impacte propriedades relacionadas à Internet aberta e tecnologias globais.

3. Dentre as diferentes definições optou-se pela descrita por MUÑOZ, A.; RÍOS, R.; ROMÁN, R.; LÓPEZ (2023), uma vez que não há consenso sobre uma definição sobre TEEs. Knodel et al. (2024) conceitua o termo a partir de uma analogia de um cofre onde o código que está a ser executado não poderá sofrer qualquer tipo de intrusão por terceiros, enquanto Sabt et al. 2015 vai além, atrelando também a aplicação do modelo de separação de kernel. O Confidential Computing Consortium (2022) segue linha similar a MUÑOZ, A, et al, enquanto que a aplicação de TEEs leva a crer que não há um único modelo de TEEs, fazendo com que sua definição também dependa da solução aplicada por cada fabricante. MUÑOZ, Antonio et al. A survey on the (in) security of trusted execution environments. Computers & Security, v. 129, p. 103180, maio 2023.

Nela, a infraestrutura desenhada para o isolamento técnico corre o risco de ser convertida, via atualizações forçadas e backdoors de governança, em um mecanismo de produção de provas, exigindo um reexame dos limites da intervenção estatal na engenharia de segurança.

Toma-se nota que o artigo não é exaustivo e busca, de forma exploratória, contribuir para a compreensão dos impactos técnicos e jurídicos da adoção de ambientes de execução confiável em ecossistemas de criptografia, identificando pontos de fricção ainda pouco discutidos na literatura brasileira.

Do ponto de vista metodológico, o trabalho é de natureza exploratória, combinando revisão de literatura técnica e jurídica com análise documental. As fontes técnicas compreendem artigos científicos sobre segurança de TEEs, documentação oficial de provedores de computação confidencial (Meta, Google) e relatórios de auditoria independente (NCC Group, Trail of Bits). As fontes jurídicas incluem legislação brasileira e estrangeira, jurisprudência do Supremo Tribunal Federal e doutrina especializada em proteção de dados e criptografia. A articulação entre as dimensões técnica e jurídica é realizada por meio de análise categorial, na qual os conceitos da engenharia de segurança são confrontados com as categorias normativas do direito da proteção de dados e do processo penal, identificando pontos de convergência e tensão.

Ao articular referências da engenharia de segurança com categorias do direito da proteção de dados e do processo penal, o trabalho pretende oferecer um enquadramento analítico inicial sobre como a proteção do dado em uso, mediada por ambientes seguros e mecanismos de atestação, desloca o debate regulatório da interceptação do tráfego para a governança da Raiz de Confiança, sem pretender esgotar o tema ou antecipar soluções normativas fechadas.

2. Ambientes de execução confiável sob uma perspectiva técnica

Dentre as variadas definições existentes na literatura, considera-se que os TEEs são áreas isoladas por hardware em microprocessadores que permitem a execução segura de códigos, garantindo assim a confidencialidade e a integridade dos dados e do código (MUÑOZ et al., 2023, p. 1). O principal objetivo desse isolamento é fornecer segurança para todo o sistema (MUÑOZ et al., 2023, p. 1).

O Confidential Computing Consortium (2022) destaca que os TEEs são ambientes que oferecem um alto nível de garantia em três propriedades principais. Primeiramente, garantem a confidencialidade dos dados, de modo que entidades não autorizadas não conseguem visualizar as informações enquanto estão em uso dentro do TEE. Em segundo lugar, asseguram a integridade dos dados, impedindo que usuários ou sistemas não autorizados adicionem, removam ou alterem os dados durante sua execução. Por fim, os TEEs também protegem a integridade do código, evitando que terceiros não autorizados modifiquem, excluam ou adicionem instruções ao código que está sendo executado em ambiente seguro. Juntos, esses atributos garantem não apenas a confidencialidade dos dados, mas também que os cálculos realizados sejam realmente corretos, permitindo que se confie nos resultados desses cálculos.

De acordo com Puddu et al. (2022), dependendo das características de um TEE específico, ele pode oferecer funcionalidades adicionais além das garantias básicas de segurança. Alguns TEEs, por exemplo, podem proteger também o código durante o uso, impedindo que entidades não autorizadas o visualizem, o que é útil para proteger algoritmos considerados propriedade intelectual confidencial.

Outros podem exigir inicialização autenticada, realizando verificações de autorização antes de iniciar um processo e recusando aqueles que não forem autorizados ou autenticados. Quanto à programabilidade, alguns TEEs permitem a execução de código arbitrário, enquanto outros suportam apenas operações limitadas ou podem ser compostos totalmente por código fixado na fabricação.

Muitos TEEs oferecem ainda atestabilidade, fornecendo evidências sobre sua origem e estado atual, de modo que uma parte externa possa verificar se deve confiar no código em execução; normalmente essas evidências são assinadas por hardware confiável para garantir que não foram geradas por malware ou terceiros não autorizados. A atestação é o processo que garante confiança na Computação Confidencial, verificando digitalmente que dados sensíveis são processados apenas em TEEs de hardware avaliados (CONFIDENTIAL COMPUTING CONSORTIUM, 2024).

Um aspecto adicional relevante para a segurança dos TEEs é a garantia de "frescor" (freshness) dos binários executados. Esse conceito refere-se à capacidade do sistema de assegurar que apenas versões atualizadas e autorizadas do código sejam carregadas no ambiente seguro, impedindo a reversão para versões anteriores potencialmente vulneráveis (rollback). A verificação de frescor é tipicamente realizada por meio de mecanismos de atestação que comparam os valores de medição do software em execução com referências conhecidas e atualizadas, garantindo que o ambiente não tenha sido comprometido por binários obsoletos ou maliciosos. Esse requisito é fundamental porque, mesmo que um TEE ofereça isolamento robusto, a execução de código desatualizado pode reintroduzir vulnerabilidades já corrigidas, anulando as garantias de segurança do ambiente.

Por fim, alguns TEEs incluem mecanismos de recuperabilidade, permitindo restaurar o funcionamento em caso de estados não conformes ou potencialmente comprometidos, como quando um componente de

firmware ou software falha na autenticação de inicialização; nesses casos, é possível atualizar o componente e tentar novamente, mantendo certos elementos do TEE como raízes confiáveis para suportar a recuperação.

Muñoz et al. (2023) destacam que existem muitas implementações diferentes de TEEs. Existem implementações em que o TEE é implementado com software, enquanto também existem várias implementações de TEE por hardware.

Outro parâmetro usado para classificar as diferentes implementações é o nível de privilégio com que são executadas, ou seja, se estamos lidando com um TEE privilegiado ou não privilegiado.

Os TEEs não privilegiados suportam múltiplas implantações, permitindo incluir uma nova funcionalidade simplesmente adicionando novas instâncias sem estender a base de computação confiável do sistema, o que aumenta a superfície de ataque do sistema.

A maioria desses TEEs utiliza um monitor seguro desde a fase de projeto (que geralmente é baseado em software) ou aproveitando diretamente os ambientes seguros suportados por hardware. Mas há também os TEEs privilegiados que, na maioria dos casos, têm acesso a todos os recursos do sistema.

2.1 Ataques e Contramedidas em Ambientes de Execução Confiável

Embora o TEE tenha sido projetado para fornecer meios avançados de execução segura de código, ele ainda pode ser atacado. De acordo com Muñoz et. al. (2023), há quatro tipos de ataques: ataques baseados em software, ataques arquitetônicos, ataques de side-channel e ataques microarquitetônicos.

Ataques baseados em software exploram falhas no sistema operacional ou nos aplicativos em execução. Erros de programação podem causar problemas nos mecanismos de proteção de memória, de segurança ou na configuração de dispositivos, abrindo espaço para ataques (MUÑOZ et al., 2023, p. 4).

Esses erros podem surgir durante a execução do sistema e serem explorados, por exemplo, por validação incorreta de parâmetros ou estouros de buffer, permitindo desde o vazamento de informações sensíveis até o comprometimento do kernel.

Ataques arquitetônicos exploram falhas no próprio projeto do hardware. Eles podem ocorrer devido a mecanismos inadequados de isolamento entre o espaço seguro e o não seguro ou por fragilidades nos sistemas de proteção de memória, resultando em exposição de dados ou vazamento de informações sensíveis (MUÑOZ et al., 2023, p. 4, 12).

Ataques por side-channel utilizam informações indiretas, como tempo de execução ou consumo de energia, para inferir dados confidenciais, especialmente chaves criptográficas. Um caso específico é a injeção de falhas, na qual o atacante provoca condições anormais, como variações de tensão, temperatura ou pulsos eletromagnéticos, para induzir erros e revelar informações secretas (MUÑOZ et al., 2023, p. 4, 13).

Ataques microarquitetônicos exploram componentes internos do processador, como caches e mecanismos de predição de execução. Ao observar padrões de acesso à memória ou tempos de resposta, um atacante pode inferir dados do ambiente seguro (MUÑOZ et al., 2023, p. 4, 14-15, 17-18).

Ataques de temporização de cache, por exemplo, permitem deduzir chaves criptográficas ao correlacionar tempos de execução com acessos à memória compartilhada. Entre os ataques microarquitetônicos, há os ataques especulativos que exploram técnicas modernas de otimização dos processadores, como execução especulativa e fora de ordem. Embora resultados incorretos sejam descartados, alterações internas, como mudanças no cache, podem permanecer e ser exploradas

para vazar informações confidenciais.

A segurança dos TEEs também pode ser comprometida por erros no gerenciamento de chaves, programação insegura e falhas na implementação. Por exemplo, chaves de criptografia armazenadas na memória podem ser acessadas por invasores, mas isso pode ser evitado usando elementos de segurança ligados ao hardware.

Muitos TEEs são programados em linguagens que não protegem bem a memória, o que pode causar falhas; por isso, usar ambientes de execução seguros e linguagens de programação mais protegidas ajuda a reduzir riscos.

Também existem métodos que verificam automaticamente se o sistema está correto e seguro. Como muitos ataques chegam na forma de malware, existem ferramentas que analisam o código, acompanham o fluxo de dados e usam a nuvem para detectar ameaças. Mais recentemente, técnicas de Inteligência Artificial (IA) têm sido usadas para identificar e classificar malwares, e o isolamento de aplicativos com hipervisores leves ajuda a tornar os sistemas ainda mais seguros.

Pesquisas sugerem várias formas de proteger TEEs contra ataques. Muitas dessas soluções focam em isolar melhor os componentes, proteger a memória e garantir que os dados não sejam alterados (MUÑOZ et al., 2023, p. 31-33).

Entre as estratégias estão separar melhor os ambientes, controlar o acesso à memória, usar cópias seguras em vez de memória compartilhada e gerenciar o ciclo de vida do sistema de forma mais rigorosa. Outras proteções incluem criptografia, proteção de dados individuais, prevenção de ataques que exploram o cache e reforço do software e do hardware. Além disso, mecanismos como inicialização segura, armazenamento protegido e verificação remota ajudam a aumentar a confiança e a integridade do sistema. Apesar disso, ainda existem limitações frente a ataques sofisticados e explorando falhas do hardware.

Ambientes de execução confiáveis e sistemas operacionais ainda têm fraquezas na proteção da memória e na defesa contra ataques sofisticados. Por exemplo, técnicas que embaralham a posição dos dados na memória nem sempre funcionam bem, permitindo que invasores descubram onde as informações estão armazenadas. Outras proteções tradicionais, como páginas de guarda e mecanismos contra estouros de memória, nem sempre são suficientes (MUÑOZ et al., 2023, p. 33).

Ataques que exploram como o processador prevê e executa instruções também criam riscos adicionais, exigindo barreiras de memória, isolamento e outras estratégias para reduzir o problema. Embora essas medidas ajudem, muitas ainda são parciais, podem criar novos riscos ou precisar de hardware extra, mostrando que ainda é um desafio proteger completamente esses sistemas.

Diante desse panorama de ameaças e contramedidas, torna-se relevante examinar como empresas de tecnologia têm implementado TEEs em produtos de ampla adoção, a fim de compreender tanto as garantias oferecidas quanto os desafios práticos que emergem da interação entre esses ambientes seguros e ecossistemas reais de comunicação e processamento de dados.

3. Exemplo de aplicações usando TEE

3.1 Aplicação de TEE pelo Whatsapp

Em abril de 2025, a Meta apresentou uma nova solução chamada Processamento Privado (em inglês, Private Processing) para o WhatsApp, desenvolvida com base em um TEE (META, 2025a; WHATSAPP, 2025). Essa iniciativa visa habilitar recursos alimentados por IA, preservando o nível de privacidade que os usuários esperam da plataforma. Por meio do Processamento Privado, funcionalidades de IA (como a de resumo de mensagens) podem ser fornecidas sem conceder à Meta ou

ao WhatsApp acesso ao conteúdo das mensagens (META, 2025a; NCC GROUP, 2025, p. 6).

A Meta argumenta que a utilização dos recursos de IA, incluindo aqueles que dependem do Processamento Privado, continua sendo totalmente opcional, muito embora a funcionalidade de inteligência artificial seja uma funcionalidade ativada por padrão em boa parte dos países (META, 2025a; WHATSAPP, 2025; NCC GROUP, 2025, p. 9).

Além disso, a empresa sustenta que os usuários mantêm o controle sobre suas conversas mais confidenciais, pois o recurso Privacidade Avançada de Chat do WhatsApp permite que eles impeçam que as mensagens sejam usadas para funcionalidades relacionadas à IA (META, 2025a; WHATSAPP, 2025).

O Processamento Privado estabelece um ambiente seguro baseado em nuvem no qual modelos de IA podem analisar e processar dados do usuário sem expô-los a terceiros não autorizados. O processo começa com a obtenção de credenciais anônimas para verificar se as solicitações recebidas são originárias de clientes legítimos do WhatsApp. Assim, o Processamento Privado pode autenticar os usuários em seu sistema, mas continua sem conseguir identificá-los. Para oferecer suporte ao Oblivious HTTP (OHTTP)[4], o sistema também recupera chaves

4. Oblivious HTTP (OHTTP) é um protocolo proposto pelo IETF com o objetivo de aumentar a privacidade das comunicações HTTP, desacoplando a identidade do cliente do conteúdo das requisições. Diferentemente do HTTPS tradicional, no qual o servidor final tem acesso tanto ao endereço IP do cliente quanto ao conteúdo da requisição, o OHTTP introduz um retransmissor intermediário (*relay*). Nesse modelo, o cliente criptografa a requisição utilizando a chave pública do servidor final, de modo que o *relay* possa encaminhá-la sem acesso ao seu conteúdo. Assim, o *relay* conhece a identidade do cliente, mas não o conteúdo da comunicação, enquanto o servidor final tem acesso ao

de criptografia públicas HPKE[5] de uma rede de entrega de conteúdo (CDN) de terceiros (META, 2025a; NCC GROUP, 2025, p. 6-7).

Em seguida, é estabelecida uma conexão OHTTP entre o dispositivo do usuário e um gateway Meta por meio de um relay de terceiros, garantindo que o endereço IP do solicitante permaneça oculto tanto para a Meta quanto para o WhatsApp (META, 2025a; NCC GROUP, 2025, p. 6-7).

Posteriormente, é criada uma sessão de Atestado Remoto e Segurança da Camada de Transporte (ou inglês, Remote Attestation and Transport Layer Security, RA-TLS) entre o dispositivo do usuário e um TEE (META, 2025a; NCC GROUP, 2025, p. 6). Durante essa etapa, as medições de atestado são verificadas em relação a um livro-razão de terceiros, garantindo que o cliente se conecte apenas a códigos que satisfaçam os requisitos de transparência verificáveis (META, 2025a; NCC GROUP, 2025, p. 3, 32).

Uma vez estabelecida a sessão segura, o dispositivo do usuário envia uma solicitação para o Processamento Privado (por exemplo, para resumo de mensagens). Essa solicitação é criptografada de ponta a ponta usando uma chave efêmera acessível apenas ao dispositivo e aos TEEs selecionados, garantindo que nem a Meta nem o WhatsApp possam des-

conteúdo da requisição, mas não à identidade do cliente. Essa separação reduz significativamente a possibilidade de rastreamento e correlação de dados, contribuindo para a preservação da privacidade do usuário.

5. *Hybrid Public Key Encryption* (HPKE) é um padrão criptográfico definido pelo IETF que especifica mecanismos para a combinação de criptografia assimétrica e simétrica, permitindo o estabelecimento seguro de chaves e a proteção eficiente de mensagens. O HPKE é utilizado em protocolos orientados à privacidade, como o Oblivious HTTP, nos quais o remetente criptografa dados utilizando a chave pública do destinatário, garantindo que apenas a entidade detentora da chave privada correspondente possa acessar o conteúdo da comunicação.

criptografar o conteúdo (META, 2025a; NCC GROUP, 2025, p. 7).

Os modelos de IA então processam os dados dentro de uma máquina virtual confidencial (em inglês, confidential virtual machine ou CVM), uma forma de TEE, sem armazenar nenhum conteúdo da mensagem. Quando necessário, as CVMs podem se comunicar com outras CVMs por meio das mesmas conexões RA-TLS usadas pelos dispositivos clientes para concluir o processamento (META, 2025a; NCC GROUP, 2025, p. 7).

Por fim, os resultados processados são devolvidos ao dispositivo do usuário, criptografados com uma chave conhecida exclusivamente pelo dispositivo e pelo servidor de Processamento Privado designado. Após a conclusão da sessão, o Processamento Privado não mantém acesso a nenhum dado da mensagem (META, 2025a; NCC GROUP, 2025, p. 10).

Embora a arquitetura do Processamento Privado represente avanço significativo na proteção de dados em uso, alguns pontos de atenção merecem destaque.

Primeiramente, a segurança do sistema depende fundamentalmente da confiabilidade do processo de atestação remota e da integridade do livro-razão de terceiros utilizado para validação; caso esses componentes sejam comprometidos, toda a cadeia de confiança é afetada.

Adicionalmente, embora a Meta afirme que os recursos de IA são opcionais, a ativação por padrão em diversos países levanta questionamentos sobre o consentimento informado dos usuários, especialmente quanto à compreensão efetiva dos riscos e das limitações técnicas envolvidas.

Além disso, a dependência de uma infraestrutura centralizada sob controle de uma única empresa concentra a governança da Raiz de Confiança, criando um ponto único de falha que poderia ser explorado tanto por agentes maliciosos quanto por pressões regulatórias para modificar o comportamento do sistema por meio de atualizações de software.

Por fim, a ausência de mecanismos independentes de auditoria contínua, para além das avaliações pontuais realizadas pela NCC Group e pela Trail of Bits, constitui limitação relevante, uma vez que a verificação da integridade do ambiente seguro precisa ser um processo permanente, e não uma validação estática.

3.2 Aplicação de TEE pelo Google

O Google Cloud (2025) oferece diversos produtos de Confidential Computing, incluíndo Confidential VM, Confidential Space, Google Cloud Attestation e Split-trust encryption tool (STET).

As instâncias Confidential VM utilizam criptografia de memória baseada em hardware para proteger dados e aplicações durante o uso, configurando um TEE (GOOGLE, 2026a). As instâncias Confidential VM oferecem isolamento, garantindo que as chaves de criptografia sejam geradas e armazenadas apenas em um hardware dedicado, ficando inacessíveis ao hipervisor (MICROSOFT, 2025)[6].

Além disso, fornecem atestação, permitindo verificar a identidade e o estado da VM, assegurando que seus componentes críticos não tenham sido modificados. Dependendo do tipo de máquina e da CPU, essas VMs podem utilizar tecnologias que fornecem criptografia de memória e proteção contra ataques de hipervisor, criam domínios de confiança isolados que protegem contra ataques físicos à memória, e estendem o TEE para permitir a execução segura de cargas de trabalho de IA e machine learning.

Já o Confidential Space é um ambiente seguro e isolado que per-

6. Um hipervisor é um software que permite criar, executar e gerenciar máquinas virtuais (VMs), possibilitando que vários sistemas operacionais diferentes rodem no mesmo hardware físico, de forma isolada e segura.

mite processar dados sensíveis de múltiplas partes, como informações pessoais, segredos criptográficos ou modelos de machine learning, mantendo a confidencialidade para seus proprietários (GOOGLE, 2026b). Ele utiliza um TEE para garantir que apenas cargas de trabalho autorizadas possam acessar os dados, com suporte de atestação e sistemas operacionais reforçados para proteger tanto a execução quanto os dados em relação ao operador. É composto por três elementos principais: a carga de trabalho, que contém o código que processa os recursos protegidos e roda em uma Confidential VM; o serviço de atestação, que fornece evidências da execução segura por meio de tokens; e os recursos protegidos, que podem estar na nuvem ou em outros ambientes.

O Google Cloud Attestation oferece uma solução unificada para verificar remotamente a confiabilidade desses ambientes, gerando provas criptográficas baseadas em valores de referência e políticas próprias, que podem ser validadas por chave pública ou certificado raiz. Essas verificações são convertidas em requisições verificáveis que seguem o padrão IETF Remote ATtestation ProcedureS (RATS) Entity Attestation Token (EAT) (NCC GROUP, 2025, p. 16; IETF, 2023).

A arquitetura RATS envolve três entidades principais: o attester (ambiente confidencial), o verifier (nesse caso, o Google Cloud Attestation) e a relying party (serviços ou aplicativos que confiam nos resultados), com papéis adicionais de configuração e endosso para garantir a integridade das avaliações. O Google Cloud Attestation usa o modelo passport, onde o attester solicita a atestação, o verifier avalia e emite o resultado, e o attester apresenta à relying party, enquanto mecanismos criptográficos garantem a atualidade e validade dos resultados.

A STET permite transferir dados sensíveis para dentro e fora do Google Cloud de forma segura, verificável e criptograficamente protegida, mesmo contra acessos não autorizados, inclusive de pessoas com privilégios internos no Google Cloud (GOOGLE, 2026c). Isso é possível porque a STET utiliza dois sistemas de gerenciamento de chaves (KMS): um interno ao Google Cloud e outro externo.

Dessa forma, mesmo que um dos sistemas seja comprometido, o outro continua ativo, garantindo a privacidade e a segurança dos dados. Em um cenário com um único KMS, esse sistema teria controle total sobre as chaves de criptografia.

Se o operador do KMS tivesse acesso aos dados cifrados, poderia descriptografá-los por conta própria. Embora isso possa ser aceitável quando o KMS é operado por uma entidade totalmente confiável, existem situações em que é necessário remover esse controle unilateral.

Com a STET, a confiança é dividida entre dois KMS, de modo que nenhum deles possui informações suficientes para descriptografar os dados sozinho. Para que os dados fossem acessados em texto simples, seria necessário que os dois operadores agissem de forma conjunta e ainda tivessem acesso ao texto cifrado. Dessa maneira, a STET garante que as únicas entidades com acesso aos dados em texto simples sejam o originador dos dados (como um sistema local) e o consumidor autorizado (como uma carga de trabalho executando em uma instância de VM confidencial).

Embora as soluções de Confidential Computing do Google ofereçam mecanismos robustos para proteção de dados em uso, ainda existem pontos de atenção. Primeiramente, a confiança na infraestrutura subjacente permanece parcialmente dependente do provedor, especialmente no que se refere à correta implementação dos mecanismos de atestação e ao gerenciamento das chaves criptográficas.

Além disso, a complexidade da arquitetura, que envolve múltiplos componentes como TEEs, serviços de atestação e sistemas de gerenciamento de chaves, pode dificultar a correta configuração e auditoria por parte dos usuários, aumentando o risco de erros operacionais.

Outro aspecto relevante diz respeito às limitações de desempenho

e compatibilidade, uma vez que nem todas as cargas de trabalho são plenamente adaptadas para execução em ambientes confidenciais.

Por fim, embora mecanismos como o split-trust reduzam riscos de acesso indevido, eles não eliminam completamente cenários de conluio entre partes confiáveis. Dessa forma, apesar dos avanços significativos, a adoção dessas tecnologias deve ser acompanhada de uma avaliação criteriosa de riscos, requisitos de confiança e trade-offs operacionais.

4. Implicações sócio-legais de ambientes de execução confiáveis

As arquiteturas de processamento privado analisadas nas seções precedentes demonstram que a adoção de TEEs transcende a mera atualização incremental de padrões de segurança. A transição técnica da tutela do dado em repouso ou em trânsito para a proteção do dado em uso (data-in-use) impõe uma reconfiguração fundamental na fenomenologia jurídica dos objetos digitais.

Esta seção examina como essa transformação técnica interage com ordenamentos jurídicos que buscam conciliar a proteção da privacidade individual com demandas estatais de acesso a dados para fins investigativos. A análise não pressupõe que tais tensões sejam inerentemente irreconciliáveis, mas identifica pontos de fricção que merecem atenção tanto da engenharia quanto da regulação.

4.1 A mensagem como conjunto de estados e seus reflexos

A introdução de TEEs na infraestrutura de comunicação altera a natureza do objeto protegido. A mensagem deixa de operar juridicamente como artefato unitário e estático para funcionar tecnicamente como um conjunto dinâmico de estados que variam conforme a finalidade computacional. Embora a mensagem original permaneça como referência semântica atribuída ao usuário, o sistema passa a operar com representações intermediárias cujos formatos e persistências são definidos por decisões de produto e validados por declarações de proveniência (KOCAOĞULLAR et al., 2024, p. 2-3)[7].

Para fins analíticos, propõe-se uma tríade conceitual na qual a categoria "mensagem" refere-se à unidade socialmente reconhecida de comunicação, situada em interação com expectativas de confidencialidade, historicamente protegida pelo regime jurídico do sigilo (OTTO; TEN HOMPEL; WROBEL, 2022, p. 21, 36, 111).

Distinta desta figura é a "representação computacional", conceito que diz respeito à tradução técnica do conteúdo para processamento interno no TEE, que em sistemas de IA envolve conversão em vetores matemáticos (embeddings) capazes de capturar relações semânticas na memória volátil (MIKOLOV et al., 2013, p. 1). Por fim, a categoria "derivado" designa o conjunto de artefatos que emergem como resultado do processamento, abrangendo inferências, relatórios de auditoria e evidências de atestação que comprovam a integridade da execução (CONFIDENTIAL COMPUTING CONSORTIUM, 2022, p. 11).

7. Além disso, conforme a análise de Kocaoğullar et al., a segurança em Computação Confidencial exige a distinção entre a intenção semântica original (o código-fonte) e o objeto técnico operacional (o binário). Este último atua como uma representação preparada cujas propriedades são determinadas por 'declarações de proveniência' — que nada mais são do que o registro das decisões de produto (configurações de build, ferramentas e flags) que transformam a mensagem original em um derivado executável pela infraestrutura. KOCAOĞULLAR, C. et al. *A Confidential Computing Transparency Framework for a Comprehensive Trust Chain*. In: Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, 2024. p.5-6.

Sob a ótica da Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018), essa distinção de estados é relevante para determinar a natureza jurídica do dado em cada etapa do processamento.

A representação computacional decifrada dentro do ambiente seguro configura tratamento de dados pessoais conforme o inciso X do Artigo 5º, uma vez que a informação torna a pessoa natural identificável perante o algoritmo.

Contudo, a arquitetura do TEE introduz uma dimensão temporal crítica, pois o dado existe em texto claro apenas transitoriamente na memória isolada e tecnicamente inacessível ao controlador da infraestrutura de nuvem. Doneda e Machado (2019, p. 142) observam que tecnologias criptográficas atuam em zona conceitual entre pseudonimização (reversível) e anonimização (irreversível). No caso dos TEEs, a qualificação jurídica depende da verificabilidade técnica da ausência de estado persistente (statelessness).

Essa dinâmica torna-se empiricamente visível na análise de arquiteturas recentes de mensageria assistida por inteligência artificial, a exemplo das implementações de Processamento Privado documentadas pela Meta em 2025 para o WhatsApp.

Nesses cenários, o TEE é empregado para converter a mensagem cifrada em representação legível pela máquina dentro da memória volátil. Auditorias independentes da NCC Group (2025, p. 10) e Trail of Bits (2025, p. 6) indicam que o conteúdo original é automaticamente descartado após geração da resposta. O processo resulta na produção de um derivado – a inferência ou o resumo gerado – enquanto a representação vetorial existe apenas transitoriamente. A tutela jurídica, neste caso, incide sobre a garantia de ausência de estado persistente do sistema, uma vez que a privacidade depende da destruição automática da representação imediatamente após o uso.

Uma dinâmica análoga, porém com função sociotécnica distinta, ocorre em ecossistemas de análise de dados e publicidade, como o Confidential Matching suportado pelo Google Confidential Space. Nesse domínio, o TEE opera não para processar linguagem natural visando o usuário final, mas para sustentar uma economia de integridade entre anunciantes e plataformas.

O objeto protegido, composto pelos dados de clientes (first-party data), é processado em ambientes seguros isolados para gerar correspondências de audiência.

O ponto crítico jurídico aqui reside nos derivados gerados, que se manifestam como relatórios de atribuição e garantias de atestação. Diferentemente da IA generativa, onde o risco se concentra na representação do conteúdo, em sistemas de computação confidencial voltados para matching o risco se concentra na veracidade e na persistência desses derivados. Tais metadados de conformidade constituem evidências técnicas que, embora não revelem o conteúdo bruto, criam um rastro de auditoria que valida a operação e desloca o eixo da confiança da reputação institucional para a verificação criptográfica da infraestrutura.

Se o derivado não permitir a reidentificação do titular original, como ocorre em relatórios agregados de publicidade no Google Confidential Space, ele poderia transitar para o regime de dado anonimizado conforme o Artigo 12 da LGPD e escapar das restrições de tratamento. Contudo, se o derivado for uma resposta personalizada entregue ao usuário, o ciclo inteiro permanece sob o escopo da proteção de dados pessoais. Consequentemente, a segurança do ambiente seguro não pode ser juridicamente compreendida apenas como a tutela do dado original em repouso ou em trânsito, mas sim como a proteção de um fluxo de execução verificável.

Cabe observar que a LGPD oferece instrumentos parcialmente aplicáveis a essa discussão, embora insuficientes para enquadrar o problema em sua totalidade.

O Art. 46 da lei impõe ao controlador o dever de adotar medidas técnicas e administrativas de segurança aptas a proteger dados pessoais contra acessos não autorizados (BRASIL, 2018). Nessa perspectiva, a adoção de TEEs poderia ser interpretada como cumprimento desse dever, na medida em que eleva a proteção do dado em uso. Inversamente, uma atualização de software que enfraqueça o isolamento do ambiente seguro poderia configurar violação do mesmo dispositivo. O Art. 48, por sua vez, estabelece o dever de comunicação de incidentes de segurança à ANPD e aos titulares, o que levanta a pergunta sobre se a modificação do código de um TEE para permitir acesso de terceiros aos dados em processamento constituiria incidente notificável.

Contudo, é necessário reconhecer que a LGPD possui limitação estrutural relevante para o cenário aqui analisado. O Art. 4º, III, da lei exclui expressamente de seu âmbito de aplicação o tratamento de dados realizado para fins de segurança pública, defesa nacional e investigação de infrações penais. Isso significa que, precisamente nos casos em que o risco de captura institucional é mais agudo, ou seja, quando a autoridade estatal determina a modificação do código executado no TEE para fins investigativos, a LGPD deixa de incidir.

O enquadramento normativo desse cenário exige, portanto, recurso a outras camadas do ordenamento, notadamente a proteção constitucional do sigilo das comunicações e o regime da interceptação telemática, que serão examinados na subseção seguinte.

Essa insuficiência normativa da LGPD torna-se ainda mais preocupante quando se consideram as fragilidades técnicas dos próprios ambientes seguros. Cerdeira et al. (2020, p. 1416) alertam que a confidencialidade em arquiteturas baseadas em hardware, como TrustZone e SEV-SNP, depende intrinsecamente de como a computação é materializada e da robustez da arquitetura privilegiada contra classes recorrentes de vulnerabilidades.

Além de tudo, ao aprofundarem a análise sobre a interação entre software e hardware, Van Bulck et al. (2019, p. 1741) demonstram que o isolamento pode degradar-se significativamente quando a execução se apoia em runtimes complexos, criando superfícies de ataque na interface entre o ambiente seguro e o sistema operacional não confiável.

Desse modo, a regulação desses ambientes exige uma compreensão detalhada de como as representações são formadas e como os derivados são extraídos, sob pena de se tutelar o objeto inadequado, o arquivo cifrado estático, enquanto violações ocorrem nos estados dinâmicos de processamento.

4.2 Governança de atualizações e pontos de tensão

A reconfiguração da mensagem como conjunto de estados processáveis, conforme delineado anteriormente, levanta questões sobre como atores estatais podem buscar acesso a informações protegidas por TEEs. Esta seção examina tensões potenciais entre arquiteturas de segurança e demandas estatais de hacking governamentais.

Os TEEs não constituem arquitetura monolítica, mas espectro de implementações heterogêneas, variando de extensões de processador como Intel SGX e AMD SEV a arquiteturas de isolamento como ARM TrustZone, cuja eficácia depende das decisões de design de cada fabricante (MUÑOZ et al., 2023, p. 1).

Enquanto tecnologias ortogonais à criptografia de ponta a ponta, os TEEs representam um avanço na engenharia de segurança projetado para proteger o "dado em uso" contra o comprometimento da infraestrutura hospedeira, oferecendo assim garantias de confidencialidade e integridade que a criptografia de transporte, isoladamente, não é capaz de fornecer (CONFIDENTIAL COMPUTING CONSORTIUM, 2022, p. 5; KNODEL et al., 2025, p. 12).

Nesse novo arranjo, a segurança do ambiente seguro não opera como um atributo estático e permanente, mas como um estado revogável que depende de um regime contínuo de verificação de conformidade (attestation) e da garantia de "frescor" (freshness) dos binários executados, que assegura a execução exclusiva de versões atualizadas e autorizadas do código (TRAIL OF BITS, 2025, p. 13). Desse modo, a integridade do TEE está diretamente condicionada à governança do ciclo de atualizações de software, uma vez que cada nova versão implantada redefine as condições de confiança da plataforma.

Contudo, ao deslocar a garantia de segurança da posse exclusiva das chaves pelo usuário para a integridade da plataforma de hardware, a arquitetura altera o ponto de equilíbrio político da confiança.

A vulnerabilidade sistêmica, neste ponto, aumenta a superfície de ataque da interceptação do tráfego, para também a governança das atualizações dos sistemas de TEEs, ou seja, se a infraestrutura falhar em garantir a integridade do ciclo de atualização ou permitir a reversão para versões vulneráveis (rollback), a proteção do ambiente seguro é anulada.

Consequentemente, inaugura-se um novo ponto de tensão regulatória na autoridade sobre a "Raiz de Confiança" (Root of Trust). Segundo Sabt et al. (2015, p. 60), a segurança do sistema passa a depender inteiramente da chancela emitida pelo provedor da plataforma sobre quais versões de código são autorizadas a rodar.

Isso cria um ponto de centralização de poder onde um agente malicioso pode explorar a política de atualização para subverter a proteção do ambiente seguro sem violar formalmente a criptografia de transporte.

Nesse sentido, relatório da Trail of Bits (2025, p. 26) classifica explicitamente engenheiros internos e fornecedores da cadeia de suprimentos como vetores de ameaça capazes de implantar imagens de máquinas virtuais confidenciais maliciosas ou manipular artefatos de construção. Se um agente hostil conseguir forçar um ataque de downgrade, carregando binários antigos e bugados para tomar o controle do sistema, conforme descrito por Muñoz et al. (2023, p. 87), ou comprometer as chaves de assinatura que validam a Raiz de Confiança, ele poderá transformar a opacidade do ambiente seguro em um vetor de exfiltração de dados invisível ao usuário.

Especificamente considerando o contexto criptográfico brasileiro, Doneda e Machado (2019, p. 139) identificam "pouca densidade normativa sobre uso da criptografia" no ordenamento nacional, enfatizando que o Marco Civil da Internet (Lei 12.965/2014) protege fluxo comunicacional mas não detalha tecnologias específicas, enquanto o Decreto 8.771/2016 (Art. 13º) estabelece deveres de gestão de vulnerabilidades que poderiam ser reinterpretados, e que há ausência de lei específica sobre controle de exportação ou regimes de custódia de chaves.

Contudo, essa caracterização merece nuance. O ordenamento brasileiro possui proteção constitucional ao sigilo das comunicações (Art. 5º, XII, CF/88), LGPD (Lei 13.709/2018) com princípios robustos de segurança e prevenção, jurisprudência do STF estabelecendo limites à interceptação estatal, tipificação penal de invasão de dispositivo informático (Art. 154-A, CP), e Lei de Abuso de Autoridade (Lei 13.869/2019) que sanciona excessos.

Portanto, práticas de hacking governamental por agentes estatais não resultam de lacuna normativa, mas de uma alegada indeterminação na incidência de princípios e garantias já consolidados diante de contextos tecnológicos específicos, especialmente quanto à extensão de sua aplicabilidade a técnicas intrusivas sofisticadas e de difícil enquadramento nas categorias jurídicas tradicionais.

A Lei nº 9.296/1996, ao regulamentar o inciso XII do Art. 5º da Constituição, estende em seu Art. 1º, parágrafo único, a possibilidade de interceptação ao "fluxo de comunicações em sistemas de informática e telemática" (BRASIL, 1996).

Ocorre que o histórico legislativo do dispositivo constitucional revela intenção deliberada do constituinte de restringir a interceptação à modalidade telefônica de comunicação. Conforme demonstrou Queiroz (2019, p. 42-44), os constituintes inverteram a ordem das modalidades comunicacionais listadas no inciso XII ao longo da tramitação, deslocando "telefônicas" para a última posição precisamente para fazer dela, e apenas dela, o "último caso" sujeito à exceção.

O critério subjacente a essa restrição, reside na distinção entre comunicações que deixam vestígios materializáveis nas pontas emissora e receptora, passíveis de busca e apreensão, e aquelas que não os deixam, para as quais a interceptação seria o único meio de acesso ao conteúdo (QUEIROZ, 2019, p. 45-47). Nessa perspectiva, a extensão pretendida pelo parágrafo único do Art. 1º da Lei 9.296/1996 às comunicações telemáticas escritas seria inconstitucional, porque tais comunicações deixam registros acessíveis por meios menos invasivos. Aplicado aos TEEs, esse enquadramento revela uma lacuna normativa genuína.

A mensagem original trocada entre os usuários é texto escrito, que permanece armazenado nos dispositivos e pode ser acessado por busca e apreensão, nos termos do Art. 240 do Código de Processo Penal.

O que não deixa vestígio algum é o estado intermediário de processamento dentro do enclave, no qual a mensagem existe em texto claro apenas transitoriamente na memória volátil. Esse estado não é fluxo comunicacional passível de interceptação, nem registro armazenado passível de apreensão. Constitui uma terceira categoria, o dado em uso, que não foi contemplada pelo Constituição de 1988 nem pela Lei 9.296/1996. Forçar o acesso a esse estado exigiria não a interceptação do canal, que permanece protegido pela criptografia de ponta a ponta, mas a modificação do próprio código executado no ambiente seguro, configurando forma de intervenção estatal que pode ser adequadamente descrita como hacking governamental sobre infraestrutura de computação confidencial.

A jurisprudência do Supremo Tribunal Federal oferece parâmetros relevantes para essa discussão. No julgamento da ADPF 403 e da ADI 5527, que trataram dos bloqueios judiciais do WhatsApp ocorridos entre 2015 e 2016, o Ministro Edson Fachin sustentou que a criptografia constitui meio legítimo de exercício do direito à privacidade e que ordens judiciais não podem exigir o impossível tecnicamente (BRASIL, 2020a).

A Ministra Rosa Weber, em voto na mesma ADI 5527, aprofundou a análise ao argumentar que a imposição de vulnerabilidades em sistemas de segurança compromete a proteção de todos os usuários, e não apenas dos investigados, configurando medida desproporcional (BRASIL, 2020b).

Embora esses precedentes tenham sido proferidos em contexto de criptografia de ponta a ponta, e não especificamente sobre TEEs, o raciocínio subjacente é diretamente aplicável. Se a imposição de backdoors em canais cifrados foi considerada desproporcional, a modificação forçada do código executado dentro de ambientes seguros suscita objeções análogas, na medida em que compromete a integridade da Raiz de Confiança de toda a plataforma e afeta indiscriminadamente o conjunto de usuários.

A convergência entre a tese de Queiroz sobre os limites constitucionais da interceptação e os votos proferidos na ADPF 403 e ADI 5527 sugere que o ordenamento brasileiro contém fundamentos constitucionais robustos para resistir à instrumentalização dessas infraestruturas pelo aparato investigativo.

Em outra linha, a análise de cenários comparados revela diferentes abordagens estatais ao redor do mundo. Na Austrália, o Assistance and Access Act (2018) confere às autoridades poder de emitir notificações exigindo modificações técnicas em sistemas. Jarvis (2021, p. 372-373)

adverte que tais distinções são inexequíveis em TEEs, onde qualquer backdoor compromete a integridade total do ambiente. Na Índia, as Intermediary Guidelines (2021) implementaram regras de rastreabilidade pressionando pela reengenharia de arquiteturas de segurança (SOUZA; MANGETH, 2019, p. 117).

No Reino Unido, o Online Safety Act (2023) introduziu poderes para que a Ofcom exija a implementação de tecnologias de detecção de conteúdo ilegal em serviços de mensageria, incluindo aqueles protegidos por criptografia de ponta a ponta (UNITED KINGDOM, 2023, s. 122). Essa previsão é funcionalmente análoga ao cenário hipotético descrito neste artigo, isto é, a imposição de filtros de conteúdo dentro de ambientes seguros, demonstrando que a pressão regulatória para acessar dados processados em TEEs não é meramente especulativa, mas reflete tendência legislativa concreta em jurisdições democráticas.

Nos Estados Unidos, o CLOUD Act (Clarifying Lawful Overseas Use of Data Act, 2018) estabeleceu mecanismo pelo qual autoridades americanas podem exigir de provedores de serviço a entrega de dados armazenados em servidores no exterior, mediante ordem judicial (UNITED STATES, 2018, § 2713).

No contexto dos TEEs, esse instrumento adquire relevância particular; se os dados são processados em ambientes seguros hospedados em infraestrutura de nuvem global, a jurisdição sobre a governança da atualização de software pode ser reivindicada (como já é em diversos casos relacionados à metadados e fluxo internacional de dados, dentre outros) pelo país sede do provedor, independentemente da localização física dos dados ou da nacionalidade dos usuários.

Assim, se uma autoridade estatal determinasse que provedor implementasse filtros de conteúdo dentro do TEE para detectar material de abuso sexual infantil, gerasse relatórios de auditoria sobre mensagens processadas, ou exportasse derivados probatórios de execução, tal demanda não violaria formalmente a criptografia de ponta a ponta (canal permanece cifrado), mas subverteria funcionalmente a promessa de isolamento do TEE. A questão central que emerge é se tal medida constituiria exercício legítimo de poder investigativo ou vulnerabilidade sistêmica inadmissível.

Nessa linha, Knodel et al. (2025, p. 40) descrevem as "Crypto Wars 3.0" como fase em que Estados abandonam demandas tecnicamente inviáveis de quebra de criptografia de ponta a ponta para demandar "moderabilidade auditável" na infraestrutura de processamento. Se exigências de logging detalhado para auditoria forense forem incorporadas ao código do ambiente seguro, podem introduzir inadvertidamente vazamentos via canais laterais (side-channel leakage), conforme Muñoz et al. (2023, p. 13) demonstram em análise de ataques de temporização. Dessa forma, a funcionalidade de monitoramento inserida não distinguiria entre requisição judicial legítima e exploração por atacante externo que manipule mesmos canais.

Emerge assim um dilema de design intrínseco. Para que o derivado tenha valor probatório, precisa ser verificável. Se o sistema opera como caixa-preta absoluta, perde utilidade forense. Isso gera pressão funcional para incorporar mecanismos de extração e exportação de provas, o que pode degradar garantias de privacidade.

O ordenamento brasileiro dispõe de salvaguardas que poderiam contestar atualizações forçadas indevidas. O Art. 154-A do Código Penal tipifica como crime a violação indevida de mecanismo de segurança. A Lei de Abuso de Autoridade (Lei 13.869/2019) sanciona excessos no exercício de funções públicas. O princípio da proporcionalidade exige que restrições a direitos fundamentais sejam adequadas, necessárias e proporcionais em sentido estrito.

Esses instrumentos funcionam como barreiras normativas para

contestar modificações arquiteturais que ampliem indevidamente a superfície de coleta, mas o processo discursivo de construção de uma "ameaça criptográfica" que legitima a adoção de medidas excepcionais compromete esses princípios e direitos já positivados no ordenamento jurídico brasileiro.

4.3 Vigilância paradoxal como risco contingente

A análise desenvolvida identifica três camadas de complexidade que interagem de forma sistêmica. Na camada técnica, os TEEs alteram a ontologia do objeto protegido, criando estados intermediários (representação computacional, derivados) cuja governança é crítica para a privacidade. Na camada sócio-jurídica, ordenamentos jurídicos precisam adaptar categorias tradicionais (mensagem, interceptação, prova) para acomodar processamento em ambientes seguros sem enfraquecer garantias fundamentais. Na camada político-institucional, existe tensão potencial, não inevitável, entre demandas de investigação criminal e integridade de arquiteturas de segurança.

Essa tensão não se resolve por superioridade abstrata de um valor sobre outro, mas exige desenho institucional que preserve proporcionalidade, transparência sobre capacidades técnicas e suas limitações, e debate democrático sobre limites aceitáveis de vigilância.

O risco identificado não é que os TEEs sejam inerentemente inseguros, mas que sua governança seja capturada por lógicas que priorizem controle sobre privacidade. Contudo, esse risco é contingente, dependendo da qualidade da implementação técnica, robustez das salvaguardas legais e em que medida operações de persecução criminal configuram acesso juridicamente relevante ao conteúdo protegido e quais parâmetros devem orientar sua admissibilidade, controle e valoração, sem desestruturar as garantias constitucionais que incidem sobre o sigilo das comunicações.

A "vigilância paradoxal", onde infraestrutura de isolamento é convertida em mecanismo de produção de provas, constitui possibilidade que requer mitigação ativa, não destino inevitável.

A eficácia da segurança prometida pelos TEEs depende também do compromisso da sociedade e do Estado com a preservação da integridade criptográfica. Sem regulação sobre medidas de hacking governamental que reconheçam a inviolabilidade do processamento privado como pilar da confiança digital, a superioridade técnica dos ambientes seguros pode ser capturada por lógica utilitarista de controle, reiterando a lição de que a segurança da informação é, indissociavelmente, uma construção técnica, jurídica e política.

5. Conclusão

A trajetória analítica percorrida neste artigo demonstra que a integração de TEEs às infraestruturas de comunicação e processamento de dados constitui avanço técnico indispensável para a segurança da economia digital contemporânea. Ao estender a tutela criptográfica do repouso e do trânsito para o momento crítico do processamento, os TEEs respondem a uma demanda latente por confidencialidade em nuvem e mitigam vetores de ataque que a criptografia de ponta a ponta não seria capaz de endereçar isoladamente.

Resta evidente que a capacidade de isolar a execução de código em ambientes protegidos por hardware eleva substancialmente a barreira de entrada para atacantes externos e malwares que visam comprometer o sistema operacional hospedeiro, configurando estratégia robusta de defesa em profundidade. Contudo, o reconhecimento desses benefícios de endurecimento (hardening) da infraestrutura não deve obscurecer a compreensão das novas vulnerabilidades sistêmicas introduzidas por

esse paradigma.

A reconfiguração ontológica do objeto protegido, que deixa de ser apenas a mensagem cifrada para se tornar conjunto dinâmico de representações computacionais e derivados, desloca o eixo da confiança da matemática imutável das chaves para a governança fluida dos ciclos de atualização de software.

Conforme discutido nas seções anteriores, esse deslocamento cria dependência crítica em relação à integridade da Raiz de Confiança e à resistência das políticas de atestação contra manipulações, sejam elas oriundas de agentes maliciosos na cadeia de suprimentos ou de possíveis coerções estatais revestidas de pretensa legalidade.

A análise do cenário brasileiro revela que essa fragilidade de governança poderia ser exacerbada por um ambiente jurídico marcado pela ambiguidade na aplicação de princípios gerais a tecnologias específicas. Contudo, o ordenamento dispõe de salvaguardas legais que protegem a infraestrutura contra ordens de modificação forçada, incluindo proteção constitucional ao sigilo, LGPD, tipificação penal de invasão de dispositivo e Lei de Abuso de Autoridade. A eficácia dessas proteções depende de sua aplicação rigorosa e de monitoramento constante da sociedade civil.

Este trabalho não pretendeu realizar crítica exaustiva ou invalidar a adoção de tecnologias de computação confidencial, mas sim evidenciar que mesmo as arquiteturas de segurança mais sofisticadas permanecem sujeitas a riscos de desvio de finalidade quando dissociadas de arranjo institucional protetivo.

Esta análise baseou-se em revisão de literatura técnica e jurídica, documentação pública de provedores e auditorias de segurança disponíveis. Não foram realizadas entrevistas com autoridades judiciais, provedores de serviço ou fabricantes de hardware, o que constitui limitação a ser endereçada em pesquisas futuras. A análise do cenário brasileiro é preliminar e demanda investigação empírica mais robusta sobre práticas investigativas e decisões judiciais envolvendo tecnologias criptográficas e ambientes de execução confiável.

Pesquisas futuras poderiam investigar empiricamente como autoridades judiciais brasileiras interpretam requisitos de acesso a dados em sistemas com TEE, analisar comparativamente marcos regulatórios de diferentes jurisdições, desenvolver frameworks técnico-jurídicos para auditabilidade verificável sem comprometimento da privacidade, e examinar mecanismos de transparência e accountability em provedores de computação confidencial.

Referências

BRASIL. Lei nº 9.296, de 24 de julho de 1996. Regulamenta o inciso XII, parte final, do art. 5° da Constituição Federal. **Diário Oficial da União**, Brasília, DF, 25 jul. 1996.

BRASIL. Supremo Tribunal Federal. **Arguição de Descumprimento de Preceito Fundamental nº 403**/SE. Relator: Min. Edson Fachin. Brasília, DF, 27 maio 2020a.

BRASIL. Supremo Tribunal Federal. **Ação Direta de Inconstitucionalidade nº 5527**/DF. Relatora: Min. Rosa Weber. Brasília, DF, 27 maio 2020b.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União, Brasília**, DF, 15 ago. 2018.

CERDEIRA, David et al. SoK: Understanding the Prevailing Security Vulnerabilities in TrustZone-assisted TEE Systems. In: **IEEE SYMPOSIUM ON SECURITY AND PRIVACY** (SP). San Francisco: IEEE, 2020. p. 1416-1432.

CONFIDENTIAL COMPUTING CONSORTIUM. A Technical Analysis of Confidential Computing. v. 1.3. [S.l.]: **The Confidential Computing Consortium**, nov. 2022.

DONEDA, Danilo; MACHADO, Diego. Proteção de dados pessoais e criptografia: tecnologias criptográficas entre anonimização e pseudonimização de dados. In: DONEDA, Danilo; MACHADO, Diego (Coord.). A Criptografia no Direito Brasileiro. 1. ed. São Paulo: **Revista dos Tribunais**, 2019. p. 137-162.

GOOGLE. 2026a. Confidential Computing overview. **Google Cloud Documentation**, [S.l.], atualizado em 05 fev. 2026. Disponível em: https://cloud.google.com/confidential-computing. Acesso em: 10 fev. 2026.

GOOGLE. 2026b. Confidential Space overview. **Google Cloud Documentation.** [S. l.], 9 fev. 2026. Disponível em: https://docs.cloud.google.com/confidential-computing/confidential-space/docs/confidential-space-overview. Acesso em: 10 fev. 2026.

GOOGLE. Split-trust Encryption Tool. **Google Cloud Documentation.** [S. l.], 2026. Disponível em: https://docs.cloud.google.com/confidential-computing/docs/split-trust-encryption-tool. Acesso em: 10 fev. 2026.

GOOGLE CLOUD. Visão geral da computação confidencial. [S.l.]: **Google Cloud Documentation**, 10 dez. 2025.

IETF. A Taxonomy of operational security considerations for manufacturer installed keys and Trust Anchors. **IETF Internet-Draft**, [S.l.], 2023. Disponível em: https://datatracker.ietf.org/doc/rfc9711/. Acesso em: 10 fev. 2026.

JARVIS, Craig. Crypto Wars: The Fight for Privacy in the Digital Age. 1. ed. Boca Raton: **CRC Press**, 2021.

KNODEL, Mallory et al. How To Think About End-To-End Encryption and AI: Training, Processing, Disclosure, and Consent. [S.l.]: **New York University; Cornell University**, mar. 2025.

KOCAOĞULLAR, Ceren et al. A Confidential Computing Transparency Framework for a Comprehensive Trust Chain. In: **PROCEEDINGS OF THE 2024 ACM SIGSAC CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY.** [S.l.]: ACM, 2024.

META. 2025a. Building Private Processing for AI tools on WhatsApp. **Engineering at Meta**, 29 abr. 2025. Disponível em: https://engineering.fb.com/2025/04/29/security/whatsapp-private-processing-ai-tools/. Acesso em: 09. fev. 2026.

META. 2025b. Private Processing for WhatsApp Overview: Technical White Paper and Security Guide. v. 1. [S.l.]: **Meta**, 10 jun. 2025.

MICROSOFT. Trusted Execution Environment (TEE). **Microsoft Learn**, [S.l.], atualizado em 07 maio 2025. Disponível em: https://learn.microsoft.com/en-us/azure/confidential-computing/overview. Acesso em: 10 fev. 2026.

MIKOLOV, Tomas et al. Distributed Representations of Words and Phrases and their Compositionality. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS**, 26., 2013.[S.l.: s.n.], 2013. p. 1-9.

MUÑOZ, Antonio et al. A survey on the (in)security of trusted execution environments. Computers & Security, v. 129, p. 103180, maio 2023.

NCC GROUP. Security and Privacy Assessment: WhatsApp Message Summarization Service. [S.l.]: **NCC Group**, 26 ago. 2025.

OTTO, Boris; TEN HOMPEL, Michael; WROBEL, Stefan (Ed.). Designing Data Spaces: The Ecosystem Approach to Competitive Advantage. Cham: **Springer**, 2022.

PUDDU, Ivan et al. On (the Lack of) Code Confidentiality in Trusted Execution Environments. **arXiv** preprint arXiv:2212.07899, 2022.

QUEIROZ, Rafael Mafei Rabelo. Privacidade, criptografia e dever de cumprimento de ordens judiciais por aplicativos de troca de mensagens. In: DONEDA, Danilo; MACHADO, Diego (Coord.). A Criptografia no Direito Brasileiro. São Paulo: **Thomson Reuters Brasil**, 2019. p. 35-48.

SABT, Mohamed; ACHEMLAL, Mohammed; BOUABDALLAH, Abdelmadjid. Trusted Execution Environment: What It Is, and What It Is Not. In: IEEE INTERNATIONAL CONFERENCE ON TRUST, SECURITY AND PRIVACY IN COMPUTING AND COMMUNICATIONS (TRUSTCOM), 14., 2015, Helsinki. Finland [S.l.]: **IEEE**, 2015. p. 57-64.

SOUZA, Carlos Affonso; MANGETH, Ana Lara. A Criptografia entre Flexibilização e Bloqueio de Aplicações: lições internacionais e a experiência brasileira. In: DONEDA, Danilo; MACHADO, Diego (Coord.). A Criptografia no Direito Brasileiro. 1. ed. São Paulo: **Revista dos Tribunais**, 2019. p. 69-117.

TRAIL OF BITS. Meta WhatsApp Private Processing: Security Assessment with Fix Review. New York: **Trail of Bits**, 26 ago. 2025.

UNITED KINGDOM. Online Safety Act 2023. London: **The Stationery Office**, 26 oct. 2023. Disponível em: https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted. Acesso em: 28 fev. 2026.

UNITED STATES. Clarifying Lawful Overseas Use of Data Act (CLOUD Act). Pub. L. No. 115-141, 132 Stat. 348. Washington, DC, 23 mar. 2018.

VAN BULCK, Jo et al. A Tale of Two Worlds: Assessing the Vulnerability of Enclave Shielding Runtimes. In: **ACM SIGSAC CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY** (CCS '19), 2019, London. UK. New York: ACM, 2019. p. 1741–1758.

WHATSAPP. Catch up on conversations with Private Message Summaries. **WhatsApp Blog**, 25 jun. 2025. Disponível em: https://blog.whatsapp.com/catch-up-on-conversations-with-private-message-summaries. Acesso em: 09 fev. 2026.

ZUBOFF, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: **PublicAffairs**, 2019.

Capítulo IV
Segurança e Nuvem

Análise de Segurança da Computação Confidencial

Jéferson Campos Nobre [1]
Laura Soares [2]

1. Jéferson Campos Nobre é Professor do Instituto de Informática (INF) e do Programa de Pós-Graduação em Computação (PPGC) da Universidade Federal do Rio Grande do Sul (UFRGS), tendo Doutorado em Computação (UFRGS) com período sanduíche na Cisco Systems (EUA). Atualmente, Jéferson é co-coordenador da Internet Engineering Task Force Latin America (IETF-LAC) Task Force, co-secretário do Network Management Research Group (NMRG) da Internet Research Task Force (IRTF) é membro do Comitê Permanente do Workshop Pré-IETF (WPIETF). Ele também fez parte do comitê de programa do ACM, IRTF & ISOC Applied Networking Research Workshop (ANRW). Jéferson é autor de RFCs, colaborou em diversos Internet-Drafts e publicou no IRTF & ISOC Workshop on Research and Applications of Internet Measurements (RAIM), o qual foi precursor do ANRW. Participou de diversos IETF meetings e colabora em diferentes Working Groups (WGs) e Research Groups (RGs). Ele recebeu em diversas oportunidades o Internet Society Fellowship to the Internet Engineering Task Force e o Applied Networking Research Workshop Travel Grant. Além da experiência relacionada ao IETF, Jéferson já colaborou em diversas posições de coordenação de eventos acadêmicos e participa do comitê de programa de diferentes eventos relacionados a redes de computadores e cibersegurança em adição ao corpo editorial de revistas nos mesmos tópicos. Anteriormente, Jéferson atuou como engenheiro de telecomunicações em operadoras de atuação internacional. Seus principais tópicos de interesse incluem Gerenciamento e Segurança de Redes, Mecanismos de Medição de Rede, Gerenciamento e Segurança em Redes Verdes, e Internet Quântica.
2. Laura Rodrigues Soares é doutoranda do Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande do Sul (UFRGS), com bolsa integral da CAPES. Seu projeto de pesquisa de doutorado foca nas áreas de cibersegurança e redes sustentáveis (green networking). Também atua no projeto do CGI.br para fomentar a participação de pesquisadores brasileiros nos encontros do Internet Engineering Task Force (IETF). Mestre em Ciência da Computação (2024) e bacharel em Ciência da

Leandro Bertholdo [3]
Roberta Robert [4]

Computação (2021), ambos pela UFRGS. Durante a graduação foi bolsista de iniciação científi ca pela CAPES, além de estagiar em organizações na área de cibersegurança. Durante o mestrado, publicou em conferências internacionais e nacionais de relevância como o IEEE Symposium on Computers and Communications (ISCC) e o Simpósio Brasileiro de Segurança da Informação (SBSeg). Ministrou também um minicurso no Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS) no tópico de planejamento de cibersegurança com ênfase no contexto médico. Alguns de seus tópicos de interesse são Gerenciamento e Segurança Redes, Efi ciência Energética e Impactos Ambientais de Sistemas Distribuídos, incluindo Redes de Comunicação, e Internet das Coisas no contexto de Computação Aplicada à Saúde.

3. Leandro Bertholdo é professor e pesquisador na Universidade Federal do Rio Grande do Sul (UFRGS), atualmente Pós-Doutorando no Instituto de Informática desta universidade. Tem doutorado em Ciência da Computação pela University of Twente, na Holanda, e mais de 25 anos de experiência em redes de computadores, segurança cibernética e resiliência de infraestrutura da Internet. Coordenou operações críticas como o Ponto de Presença da Rede Nacional de Ensino e Pesquisa no Rio Grande do Sul (PoP-RS da RNP), o Centro de Tratamento e Resposta a Incidentes de Segurança do Rio Grande do Sul (CERT-RS) e o Ponto de Troca de Tráfego Internet Exchange de Porto Alegre (IX.br de Porto Alegre), além de liderar projetos nacionais em monitoramento de tráfego e segurança de redes. Sua produção científi ca inclui publicações em conferências de alto impacto como USENIX Security Symposium, Internet Measurement Conference (IMC), Conference on Network and Service Management (CNSM) e Traffi c Measurement and Analysis Conference (TMA), com foco em roteamento interdomínio, anycast, pontos de troca de tráfego e medição da Internet. Atua como revisor em periódicos internacionais e já foi agraciado com o prêmio "Construtores da Internet.br". Atualmente, desenvolve pesquisas sobre automação de roteamento, redes verdes e Internet Exchanges.

4. Roberta Robert é Especialista de Segurança em IA no Hospital Israelita Albert Einstein, com mais de 15 anos de experiência em Segurança Ofensiva, Segurança de Aplicações e Arquitetura Segura. É bacharel em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (UFRGS), com parte da formação realizada em Cibersegurança na Polytech Nice Sophia (França), através do programa BRAFITEC – CAPES. Atualmente, também atua como pesquisadora e mestranda no grupo de Redes e Segurança do Instituto de Informática da UFRGS, sob orientação de Jéferson Campos Nobre (UFRGS) e co-orientação de Bruno Castro da Silva (UMass – Amherst), com foco em Evasão de Malwares e Segurança de Inteligências Artifi ciais. É uma das pioneiras na comunidade brasileira de hacking, sendo co-fundadora do hackerspace MateHackers e uma das primeiras mulheres a atuar tecnicamente na cena nacional. Tem participação ativa na promoção de diversidade no ecossistema de segurança, sendo também palestrante desde 2013 nas principais conferências da comunidade de cibersegurança e tecnologia do país, como Roadsec, Hacking na Web Day, BSides, YSTS, DevOpsDays, TDC, entre outras.

Resumo: A Computação Confidencial busca proteger dados sensíveis durante o processamento em ambientes de nuvem, estendendo garantias tradicionalmente limitadas ao armazenamento e à transmissão. Este capítulo traz um panorama sobre as principais tecnologias empregadas em sistemas de Computação Confidencial e Comunicação Anônima sob um modelo de ameaça realista, evidenciando que tais garantias são condicionais e dependem de uma cadeia de confiança baseada em hardware e atestação remota. Também discute a integração com técnicas de Comunicação Anônima para proteção de metadados, destacando sua complementaridade com a Criptografia Ponta-a-Ponta (End-to-End Encryption - E2EE). Como estudo de caso, é analisado o Private Processing da Meta, ilustrando benefícios e limitações práticas desse paradigma.

Palavras-chave: Computação Confidencial. Comunicação Anônima. Ambiente de Execução Confiável. Privacidade. Segurança.

Abstract: Confidential Computing aims to protect sensitive data during processing in cloud environments, extending security guarantees that are traditionally limited to data storage and transmission. This

chapter provides an overview of the main technologies employed in Confidential Computing and Anonymous Communication systems under a realistic threat model, highlighting that such guarantees are inherently conditional and depend on a hardware-based chain of trust and remote attestation mechanisms. It also discusses the integration with Anonymous Communication techniques for metadata protection, emphasizing their complementarity with End-to-End Encryption (E2EE). As a case study, it analyzes Private Processing from Meta, illustrating the practical benefits and limitations of this paradigm.

Keywords: Confidential Computing. Anonymous Communication. Trusted Execution Environment. Privacy. Security.

1. Introdução

O atual cenário de adoção de serviços em nuvem e modelos de IA intensifica os desafios de governança sobre dados críticos. Os métodos tradicionais de proteção são eficazes para dados armazenados ou trafegados, mas falham em isolar o dado durante o ciclo de processamento. Nesse estágio, a exposição na memória principal permite o acesso por parte do sistema operacional e do hypervisor[5], facilitando alguns tipos de ataques, como, por exemplo, aqueles relacionados a supply chain[6]. A Computação Confidencial surge, portanto, como um paradigma para mitigar essa lacuna de segurança em nível de hardware.

5. Hypervisor é a denominação dada ao componente de software ou hardware responsável por criar, rodar e gerenciar máquinas virtuais em uma máquina hospedeira (host machine).
6. Um ataque de supply chain ocorre quando um adversário compromete um componente ou fornecedor confiável para atingir sistemas que dependem dessa confiança.

A Computação Confidencial define-se como um conjunto de técnicas e arquiteturas que permitem a execução de cargas de trabalho em ambientes isolados, oferecendo garantias formais quanto à confidencialidade e à integridade dos dados e do código em execução, mesmo na presença de um sistema operacional ou hypervisor potencialmente não confiável. Tais definições são sistematizadas pelo Confidential Computing Consortium (THE LINUX FOUNDATION PROJECTS, [s.d.]), um consórcio industrial que promove padrões, diretrizes e definições técnicas para a construção de ambientes que oferecem garantias de confidencialidade e integridade dos dados, bem como da integridade do código em execução (CONFIDENTIAL COMPUTING CONSORTIUM, 2022b).

Grandes empresas de tecnologia têm adotado abordagens relacionadas à Computação Confidencial (CONFIDENTIAL COMPUTING CONSORTIUM, 2022c), impulsionadas pela necessidade de operar sobre dados sensíveis, atender requisitos regulatórios e estabelecer garantias técnicas de confidencialidade e integridade. Quando corretamente implementadas, essas abordagens podem ampliar mecanismos de auditabilidade e verificabilidade técnica, contribuindo para o aumento da transparência e garantias formais de segurança do sistema.

Um dos principais componentes da Computação Confiável são os Ambientes de Execução Confiáveis (Trusted Execution Environments - TEEs) ancorados em uma cadeia de confiança com suporte em hardware especializado. Os TEEs impedem o acesso não-autorizado ou modificações em aplicações e dados, mesmo em ambientes de nuvem compartilhados. Assim, há proteção contra usuários com privilégios, como fornecedores de nuvem e administradores de infraestrutura, que normalmente têm acesso ao hardware ou software subjacente. Dessa forma, mesmo diante do comprometimento da infraestrutura, os dados no TEE permanecem protegidos.

Diversas iniciativas têm sido produzidas em computação confidencial, tanto no contexto de infraestrutura em nuvem quanto em aplicações voltadas ao usuário final. O Google, por exemplo, oferece as Confidential Virtual Machines, que utilizam tecnologias de hardware para garantir que os dados de uma organização permaneçam invisíveis durante o processamento, inclusive para o próprio provedor de nuvem.

A Meta introduziu o Private Processing (Processamento Privado), um novo protocolo para o WhatsApp que busca garantir a confidencialidade durante as solicitações à IA da Meta para ações de processamento dos dados do usuário através de IAs generativas, por exemplo, para obter resumos de conversas ou outros recursos de assistência virtual. Já a Apple introduziu o Private Cloud Compute, no contexto da Apple Intelligence, no qual tarefas sensíveis são executadas em servidores isolados baseados em Apple Silicon, sem armazenamento persistente e com possibilidade de auditoria independente.

A adoção de iniciativas em Computação Confidencial possui propriedades interessantes em Cibersegurança, mas levanta questões relevantes. A maioria das soluções de Computação Confidencial concentra-se na atestação remota e na criptografia de memória, como, por exemplo, através do uso de VMs confidenciais do Google (i.e., validação da integridade da VM antes do carregamento de dados sensíveis). Quando essas iniciativas são combinadas com aceleração por GPU, múltiplos intermediários de rede e novos protocolos criptográficos, a complexidade arquitetural aumenta, ampliando também a superfície de ataque.

Diversos relatórios já foram produzidos para discutir aspectos de cibersegurança em soluções de Computação Confidencial. Por exemplo, a Meta contratou análises especializadas, como a do NCC Group[7], para conduzir uma avaliação de segurança e privacidade do Serviço de Resumo de Mensagens do WhatsApp, que faz parte de um sistema mais amplo de Private Processing (NCC GROUP, 2025). Também no contexto do Private Processing, uma análise foi realizada pelo Trail of Bits[8] considerando diferentes aspectos de segurança (HESS et al., 2025). A adoção de auditorias independentes e verificabilidade técnica contribuem para a construção de sistemas mais confiáveis e eticamente responsáveis.

O presente trabalho busca discutir em que medida a Computação Confidencial assegura as propriedades de Cibersegurança sob um modelo de ameaça realista. Como estudo de caso ilustrativo, é analisada a arquitetura de Private Processing apresentada pela Meta no contexto do WhatsApp. A análise considera tanto os mecanismos de isolamento de execução quanto a relação com modelos (e.g, End-to-End Encryption - E2EE), enfatizando que essas tecnologias oferecem garantias complementares e não-substitutivas.

O documento está organizado da seguinte forma. A Seção 2 apresenta os fundamentos de Computação Confidencial e Comunicação Anônima. A Seção 3 discute suas propriedades de segurança e limitações. A Seção 4 analisa o Private Processing do WhatsApp como estudo de caso. Por fim, a Seção 5 traz as considerações finais.

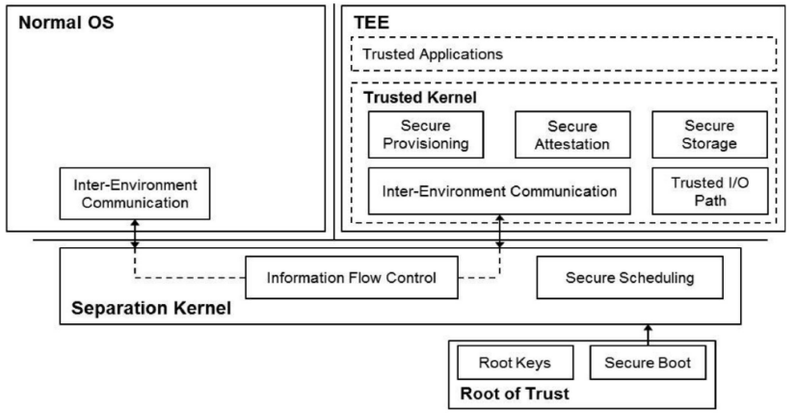
2. Fundamentação teórica

O atual cenário de adoção de serviços em nuvem e modelos de IA intensifica os desafios de governança sobre dados críticos. Os métodos tradicionais de proteção são eficazes para dados armazenados ou trafe-

7. https://www.nccgroup.com.br/pt-br.
8. https://trailofbits.com/.

gados, mas falham em isolar o dado durante o ciclo de processamento. Nesse estágio, a exposição na memória principal permite o acesso por parte do sistema operacional e do hypervisor[9], facilitando alguns tipos de ataques, como, por exemplo, aqueles relacionados a supply chain[10]. A Computação Confidencial surge, portanto, como um paradigma para mitigar essa lacuna de segurança em nível de hardware.

Figura 1. Arquitetura conceitual de um Trusted Execution Environment (TEE), destacando a separação entre o sistema operacional convencional, o ambiente protegido de execução, o Separation Kernel e a raiz de confiança baseada em hardware (SABT; ACHEMLAL; BOUABDALLAH, 2015)



9. Hypervisor é a denominação dada ao componente de software ou hardware responsável por criar, rodar e gerenciar máquinas virtuais em uma máquina hospedeira (host machine).
10. Um ataque de supply chain ocorre quando um adversário compromete um componente ou fornecedor confiável para atingir sistemas que dependem dessa confiança.

O principal objetivo da Computação Confidencial é proteger dados sensíveis durante a execução, isto é, enquanto estão em uso. Diferentemente dos modelos tradicionais de segurança, que se concentram na proteção de dados em repouso ou em trânsito, esse paradigma estende as garantias de segurança para o momento do processamento, considerado historicamente o mais vulnerável (COSTAN; DEVADAS, 2016).

A Comunicação Anônima, por sua vez, tem o objetivo de proteger a identidade das partes envolvidas em uma comunicação, bem como os metadados associados às trocas de mensagens, como endereços IP, padrões temporais e volumes de tráfego. Mesmo quando o conteúdo das comunicações é protegido por criptografia, esses metadados podem revelar informações sensíveis sobre os usuários, suas relações e seus comportamentos, configurando um risco significativo à privacidade. Os dois paradigmas são frequentemente utilizados em conjunto. Finalmente, devido às demandas crescentes de aplicações modernas, sistemas de Computação Confidencial e Anônima estão cada vez mais dependentes de serviços baseados em nuvem para sua operação. Essa mudança de paradigma impõe desafios adicionais quando comparada a cenários tradicionais de comunicação entre pares, pois combinam maior demanda por segurança e privacidade com uma baixa tolerância à latência e problemas de desempenho. Esta seção explora os paradigmas de Computação Confidencial e Comunicação Anônima e sua transição para serviços baseados em nuvem. São abordadas as tecnologias que fundamentam cada sistema, assim como sua integração e suas demandas por segurança.

2.1 Computação Confidencial

Alguns aspectos podem ser destacados como fundamentais na Computação Confidencial. O primeiro deles, a proteção contra amea-

ças internas, exige que dados e códigos sensíveis permaneçam inacessíveis mesmo a usuários com privilégios elevados, como administradores de infraestrutura ou provedores de serviço. Já o isolamento em nível de hardware busca garantir a confidencialidade e a integridade diretamente através da arquitetura do processador, tipicamente utilizando memória criptografada e estados de CPU protegidos. Ambientes de execução confiável atendem ambos aspectos. Outro pilar fundamental, a Atestação Remota (Remote Attestation - RA), permite a um verificador confirmar criptograficamente que um sistema remoto está executando software legítimo e sem modificações dentro de um ambiente seguro antes do envio de dados confidenciais para processamento.

Para garantir essas características, ferramentas de Computação Confidencial costumam contar com o apoio de uma combinação de componentes de hardware e software conforme disponibilizadas pelos fabricantes de seus sistemas.

Uma das principais ferramentas de sistemas de Computação Confidencial é o Trusted Execution Environment (TEE), ou ambiente de execução confiável. Apesar de possuir uma definição ampla na literatura acadêmica (SABT; ACHEMLAL; BOUABDALLAH, 2015), o conceito de TEE foi consolidado recentemente em um padrão da Confidential Computing Consortium (CONFIDENTIAL COMPUTING CONSORTIUM, 2022c) como um ambiente de processamento que proporciona níveis de garantia sobre a integridade e confidencialidade dos dados e integridade do código. Esses ambientes estabelecem um domínio de execução isolado em relação ao sistema operacional convencional e a outros componentes potencialmente não confiáveis. Conforme ilustrado na Figura 1, esse isolamento separa explicitamente Sistema Operacional Normal (Normal OS) e TEE, sendo que, no segundo, aplicações confiáveis são executadas sobre um Núcleo Confiável (Trusted Kernel).

Esse núcleo é responsável por funções críticas, como provisionamento seguro, atestação, armazenamento protegido e controle de caminhos de entrada e saída confiáveis (Trusted I/O Path). O código em execução não pode ser nem observado nem manipulado por terceiros, considerando outros processos, o sistema operacional, ou pessoas com acesso ao hardware (KNODEL et al., 202

A base do isolamento provido pelo TEE é uma raiz de confiança baseada em hardware, representada na figura pelo bloco Root of Trust, que engloba mecanismos como Secure Boot e uma raiz criptográfica confiável (Root Keys). Essa raiz de confiança busca assegurar que apenas código e dados autenticados não modificados sejam executados no ambiente protegido. Se implementada corretamente, a raiz de confiança estabelece as garantias criptográficas fundamentais que sustentam o modelo de segurança do TEE. Já a proteção de dados em uso é reforçada pela atuação conjunta do Trusted Kernel e do Separation Kernel. O Separation Kernel é responsável por aplicar políticas de isolamento, controle de fluxo de informação (Information Flow Control) e escalonamento seguro (Secure Scheduling), impedindo que dados sensíveis sejam acessados ou modificados por componentes externos ao TEE. Em conjunto com técnicas como criptografia de memória e proteção de estados de CPU, esses mecanismos visam reduzir a superfície de ataque e limitar a necessidade de confiar em elementos privilegiados da infraestrutura.

A atestação remota é um dos pilares centrais da Computação Confidencial, implementada no âmbito do Trusted Kernel. Por meio desse mecanismo, um verificador externo pode confirmar criptograficamente que o sistema está executando um software legítimo e corretamente configurado dentro do TEE antes do envio de dados confidenciais. Esse processo está ilustrado na Figura 1 pelos blocos de Secure Attestation e suas comunicações. Sua função é possibilitar a verificação independente

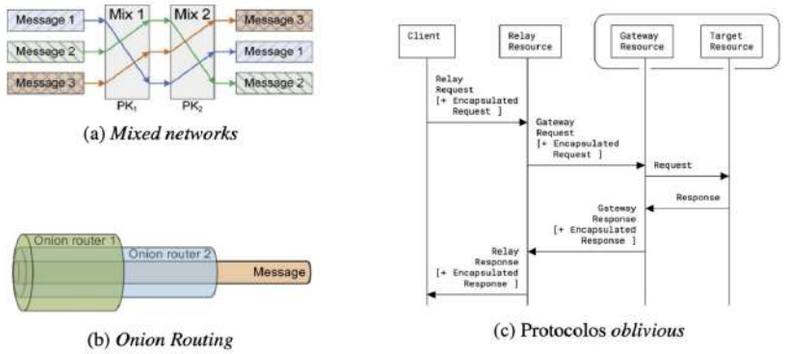


Figura 2 Exemplos de técnicas utilizadas para comunicação anônima: (a) redes de mixagem, (b) roteamento em camadas e (c) comunicação baseada em relays. (THE LINUX FOUNDATION PROJECTS, [s.d.])

das propriedades de segurança do ambiente de execução, reforçando a transparência e a confiabilidade do modelo.

Em conjunto, os princípios da arquitetura estabelecem a base sobre a qual soluções modernas de Computação Confidencial são construídas, permitindo que dados sensíveis sejam processados com maior grau de proteção e controle. Esses princípios são definidos a partir de um modelo de ameaça explícito, no qual se assume que componentes como sistemas operacionais, hipervisors, administradores de infraestrutura e até provedores de nuvem podem estar comprometidos. Entretanto, apesar das garantias oferecidas, nem todos os vetores de ataque são eliminados.

O modelo assume que o hardware e o microcódigo e da cadeia de suprimentos associada (e.g. periféricos) são confiáveis, introduzindo uma confiança residual nos fabricantes e nos processos de atualização de firmware. Além disso, embora mecanismos como criptografia de memória e isolamento de execução reduzam significativamente a superfície de ataque, eles não oferecem proteção completa contra ataques por canais laterais ou falhas microarquiteturais, como aqueles baseados em tempo de execução ou comportamento de cache. A segurança do sistema depende de uma gestão adequada do ciclo de vida de chaves criptográficas e da correta implementação de políticas de falha segura (fail-closed), de modo que qualquer violação das garantias esperadas resulte na interrupção do processamento ou na não aceitação de dados sensíveis. Reconhecer essas limitações é essencial para uma avaliação realista das propriedades de segurança da Computação Confidencial em cenários práticos—mais sobre a segurança dos diferentes aspectos da Computação Confidencial é abordado na Seção 3.

A Computação Confidencial busca garantir a autenticidade e a integridade do código em execução, mas sozinha não dá garantias sobre a anonimidade do usuário ou de seus dispositivos. Para isso, é necessária a integração com técnicas de Comunicação Anônima que ocultam a proveniência dos dados do ambiente de execução. A seguir serão abordados as principais ferramentas que fundamentam a Comunicação Anônima.

2.2 Fundamentos da Comunicação Anônima

O princípio central da comunicação anônima é a separação entre a identidade da fonte da informação e o conteúdo transmitido, de modo que nenhuma entidade isolada tenha acesso simultâneo a ambos. Essa separação reduz a possibilidade de correlação entre usuários e mensagens e limita a capacidade de observação e inferência por adversários. O modelo de ameaça assume que os componentes da infraestrutura de rede podem querer observar e inferir informações sensíveis, atuando como observadores passivos, ou mesmo desempenhar um papel de adversário ativo e potencialmente malicioso. Na prática, a separação entre as identidades da fonte e do conteúdo é implementada por meio

da introdução de intermediários na comunicação, como relays ou gateways, que encaminham mensagens cifradas entre clientes e servidores. No modelo de Comunicação Anônima, diferentes entidades têm acesso apenas a subconjuntos das informações envolvidas na comunicação, estabelecendo uma divisão de responsabilidades que busca impedir a associação direta entre identidade, conteúdo e destino final.

Diversas técnicas operacionalizam o modelo de Comunicação Anônima, conforme ilustrado na Figura 2. (REN; WU, 2010). Embora adotem mecanismos distintos, como, por exemplo, encadeamento de múltiplos saltos, criptografia em camadas ou roteamento probabilístico, todas buscam reduzir a capacidade de um atacante correlacionar a identidade da origem com o conteúdo transmitido ao seu destino final. Esses mecanismos apresentam diferentes trade-offs em termos de desempenho, latência e anonimidade.

Uma classe tradicional de técnicas baseiam-se em mixed networks, nas quais múltiplas mensagens são recebidas, reordenadas e reenviadas por intermediários (mixes). Essa estratégia dificulta a associação entre entradas e saídas e introduz uma incerteza temporal e estrutural no fluxo de mensagens. Com isso, ataques de correlação se tornam significativamente mais complexos, o que contribui para desestimular adversários. Outra abordagem é o roteamento por camadas criptográficas, como no caso das redes de anonimato baseadas em onion routing (GOLDSCHLAG; REED; SYVERSON, 1996).

Nesses sistemas, as mensagens são encapsuladas em múltiplas camadas de criptografia e encaminhadas por uma sequência de relays. Cada relay remove apenas uma camada criptográfica, aprendendo somente o próximo salto, mas nunca a origem e o destino final simultaneamente, o que limita a visibilidade de qualquer nó individual sobre o caminho completo da comunicação.

Protocolos de comunicação oblivious foram desenvolvidos para utilizar intermediários especializados a fim de desacoplar a identidade do cliente do servidor de destino. Nesses protocolos, um relay conhece a identidade do cliente, mas não o conteúdo ou o destino final da requisição, enquanto o servidor de destino processa a solicitação sem acesso à identidade do solicitante. A divisão de responsabilidades reduz a exposição dos metadados, sendo especialmente relevante em cenários de serviços modernos baseados em nuvem. Os protocolos oblivious oferecem anonimização eficiente de metadados em tais cenários. A próxima seção aborda os desafios presentes ao adaptar sistemas de Computação Confidencial e Anônima para ambientes baseados em nuvem, onde os requisitos operacionais de baixa latência muitas vezes prejudicam a atuação de mecanismos de privacidade–como os relays.

2.3 Comunicação Anônima em Serviços Modernos Baseados em Nuvem

Aplicações contemporâneas, tais como serviços de mensagens e plataformas de inteligência artificial, demandam baixa latência, alta disponibilidade e escalabilidade global. Esses requisitos nem sempre são compatíveis com mecanismos clássicos de anonimização, como mixed networks ou onion routing mencionados na seção anterior. Nesse contexto, observa-se uma tendência à adoção de mecanismos de anonimização mais leves, projetados especificamente para arquiteturas cliente–servidor amplamente distribuídas. Em vez de prover anonimato forte contra adversários globais, essas abordagens priorizam a proteção de metadados sensíveis frente a intermediários específicos, como servidores de aplicação, provedores de nuvem ou operadores de infraestrutura de transporte. O objetivo central passa a ser a mitigação de riscos de rastreamento, correlação e perfilamento, mantendo níveis aceitáveis de desempenho e confiabilidade.

Uma característica relevante de ambientes de processamento em nuvem é a separação funcional entre os componentes responsáveis pela terminação da conexão, pelo encaminhamento do tráfego e pelo processamento das requisições. Essa decomposição permite a aplicação prática do princípio de divisão de responsabilidades discutido nos fundamentos da Comunicação Anônima, ao mesmo tempo em que viabiliza integrações com mecanismos de segurança adicionais, como autenticação criptográfica, controle de acesso, monitoramento operacional, e geração de trilhas de auditoria. Além disso, serviços modernos frequentemente combinam técnicas de comunicação anônima com outros mecanismos de proteção, como criptografia de ponta-a-ponta e ambientes de execução confiáveis. Em suma, a comunicação anônima atua na proteção da identidade e dos metadados associados à interação, enquanto técnicas complementares são empregadas para proteger o conteúdo e o processamento dos dados. Essa composição de mecanismos reflete uma abordagem pragmática, na qual diferentes camadas de segurança são utilizadas para enfrentar ameaças distintas ao longo do ciclo de vida da informação.

É importante destacar que as garantias oferecidas por soluções de Comunicação Anônima em ambientes de nuvem dependem do modelo de ameaça adotado e das entidades consideradas confiáveis. A análise dessas soluções deve, portanto, levar em conta não apenas os protocolos empregados, mas também as suposições sobre os papéis desempenhados por intermediários, provedores de infraestrutura e operadores de serviços, bem como os mecanismos de verificação e transparência disponíveis para usuários e auditores. A Seção 3 traz um modelo de ameaças de um sistema de Computação Confidencial, seguido de uma análise das propriedades de segurança oferecidas dentro do contexto.

3. Propriedades de Segurança

Este capítulo apresenta uma análise das propriedades de segurança da Computação Confidencial em arquiteturas baseadas em Trusted Execution Environments (TEEs), incluindo Confidential VMs (CVMs). A análise é conduzida a partir de um modelo de ameaças, seguido do mapeamento entre requisitos de segurança e mecanismos arquiteturais, e, por fim, da identificação da superfície de ataque e de limitações estruturais do modelo.

3.1 Modelo de Ameaça Assumido

O modelo de ameaças usado para a Computação Confidencial assume que (i) o sistema operacional pode estar comprometido, (ii) o hypervisor pode ser malicioso ou vulnerável, (iii) administradores de infraestrutura podem agir como adversários, e que (iv) o provedor de serviços em nuvem pode ter capacidade de observar ou interferir na execução. Esse modelo desloca o perímetro de confiança do software para incluir também o hardware, o que estabelece o TEE como parte do domínio confiável. O modelo também assume a confiança na correção e honestidade do fabricante do processador (e.g., AMD, Intel, NVIDIA), a ausência de vulnerabilidades não documentadas no silício ou no microcódigo, a integridade da cadeia de suprimentos de hardware até o data center, e a correta implementação e gestão do ciclo de vida de chaves provenientes do fabricante (e.g., firmware signing keys, endorsement keys). Também se assume a ausência de vetores microarquiteturais com capacidade prática de extração remota de informação em alta taxa (e.g., variantes exploráveis da família Spectre (KOCHER et al., 2020) ou ataques baseados em modulação dinâmica de frequência, como Hertzbleed (WANG et al., 2022).

3.2 Requisitos de Segurança, Mecanismos e Formalização Condicional

A Tabela 1 apresenta um mapeamento entre os principais requisitos de segurança reivindicados para a Computação Confidencial e os mecanismos técnicos empregados para tentar garanti-los. Embora o mapeamento apresentado sugira uma correspondência direta entre requisitos e mecanismos, as garantias oferecidas são necessariamente condicionais ao modelo de ameaça assumido e à integridade da cadeia de confiança subjacente. Em particular, a eficácia dos mecanismos depende da correta implementação de TEEs, da robustez da atestação remota e da ausência de vulnerabilidades críticas, conforme o modelo de ameaças assumido.

Tabela 1 Requisitos de segurança da Computação Confidencial e mecanismos de suporte correspondentes

Requisito de Segurança	**Mecanismos de Suporte**
Confidential Processing	Execução em TEEs com criptografia de memória; isolamento no nível do hardware; uso de CVMs; proteção contra acesso do hipervisor e sistema operacional; transporte autenticado e protegido para dispositivos externos (e.g., device attestation e estabelecimento de canal seguro).
Enforceable Guarantees	Atestação remota (e.g., RA-TLS); verificação criptográfica do estado do software; políticas fail-closed; validação da cadeia de certificação do fabricante.
Auditable Transparency	Registro de medições, imagens e configurações em logs públicos imutáveis mantidos por terceiros independentes; mecanismos de auditoria externa.
Protection Against Insider Threats	Isolamento criptográfico de dados e código; restrição de acesso privilegiado; independência do hipervisor; proteção contra observabilidade por operadores da infraestrutura.
Non-targetability	Arquitetura uniforme e não personalizada; separação entre identidade do usuário e instância de execução; uso de protocolos oblivious e relays independentes.
Stateless Processing	Ausência de armazenamento persistente; descarte seguro de memória; inexistência de histórico acessível; controle de snapshots e rollback.
Forward Security	Uso de chaves efêmeras; rotação frequente de segredos; derivação segura de chaves; limitação temporal de material criptográfico.

Em particular, cada requisito depende de um conjunto de hipóteses implícitas que, se violadas, podem comprometer a garantia pretendida. A seguir, analisamos criticamente essas dependências e suas implicações.

Confidential Processing. A execução em TEEs, seja por meio de enclaves ou de Confidential Virtual Machines (CVMs), busca assegurar que dados e código permaneçam inacessíveis ao sistema operacional, hypervisor e operadores da infraestrutura. Essa propriedade é implementada por mecanismos como criptografia de memória, isolamento no nível do hardware, proteção de registradores e inicialização

segura (secure boot). Entretanto, sua efetividade depende da robustez microarquitetural do processador, da ausência de vulnerabilidades críticas no silício ou no microcódigo e da inexistência de canais laterais exploráveis sob condições práticas. Quando o processamento é estendido a dispositivos especializados, a confidencialidade passa também a depender da integridade do firmware, da cadeia de inicialização e de mecanismos seguros de autenticação e estabelecimento de canal protegido entre componentes.

Enforceable Guarantees. As garantias enforçáveis decorrem principalmente de mecanismos de atestação remota, que vinculam criptograficamente o estado do software ao ambiente de execução, permitindo que verificadores confirmem que o código executado corresponde a uma imagem legítima e auditada. A atestação constitui, portanto, a base de confiança entre cliente e infraestrutura. Sua eficácia depende da segurança das chaves raiz do fabricante, da integridade do processo de medição, da confiabilidade do serviço de verificação e da correta implementação de políticas fail-closed. Falhas na cadeia de certificação ou na validação das evidências podem invalidar completamente as garantias pretendidas.

Auditable Transparency. A transparência auditável baseia-se no registro de medições, imagens e configurações críticas em logs públicos imutáveis mantidos por terceiros independentes, possibilitando verificação externa. Todavia, essa propriedade depende da governança e independência dos operadores desses logs, da integridade dos mecanismos de publicação e da existência de capacidade técnica efetiva para auditoria por parte da comunidade. Assim, a transparência é condicional não apenas à disponibilização de informações, mas à viabilidade concreta de sua verificação.

Protection Against Insider Threats. A proteção contra ameaças internas decorre do isolamento criptográfico proporcionado pelo TEE, que restringe o acesso de administradores de infraestrutura, operadores de nuvem ou outros usuários privilegiados a dados e código em execução. Contudo, essa propriedade permanece condicionada à integridade do hardware subjacente, à correta configuração do ambiente de virtualização e à inexistência de vetores microarquiteturais ou canais laterais que permitam inferência indireta de informações. Além disso, riscos associados à cadeia de suprimentos e ao firmware permanecem no domínio da confiança residual.

Non-targetability. A non-targetability visa impedir a associação entre usuários específicos e instâncias de processamento, reduzindo a viabilidade de ataques direcionados. Essa propriedade é tipicamente implementada por meio da separação entre identidade de rede e ambiente de execução, arquiteturas uniformes e uso de protocolos oblivious com relays independentes. Sua eficácia depende da correta divisão de papéis entre os componentes da arquitetura, da ausência de retenção indevida de metadados e da inexistência de vetores de correlação temporal ou comportamental.

Stateless Processing. O processamento sem estado busca impedir retenção persistente de dados, reduzindo riscos de comprometimento retroativo. Essa propriedade requer ausência de armazenamento persistente, descarte seguro de memória ao término da sessão e controle rigoroso de mecanismos de snapshot ou rollback. Sua validade depende não apenas de mecanismos técnicos, mas também de práticas operacionais consistentes e auditáveis.

Forward Security. A segurança prospectiva é obtida por meio do uso de chaves efêmeras, rotação frequente de segredos e mecanismos seguros de derivação de chaves, buscando garantir que comprometimentos futuros não afetem dados previamente processados. Contudo, essa propriedade depende de um gerenciamento adequado do ciclo de vida das chaves e da inexistência de exposição inadvertida de material

criptográfico em memória ou por canais laterais.

Os requisitos analisados não são independentes. Por exemplo, forward security depende implicitamente da integridade do mecanismo de atestação, enquanto a nontargetability é reforçada por mecanismos de comunicação anônima externos ao TEE. A busca por stateless processing deve ser balanceada com práticas de verifiable transparency, que podem introduzir retenção de metadados para fins de auditoria. Consequentemente, a arquitetura não elimina riscos, mas os redistribui ao longo da pilha tecnológica, deslocando o núcleo da confiança para componentes de hardware, firmware e serviços de verificação externos.

Cada requisito introduz sua própria superfície de ataque residual. A confidencialidade em uso concentra riscos microarquiteturais; a atestação amplia a dependência em infraestruturas de chave pública do fabricante; e a transparência pública depende da robustez institucional de terceiros. Nesse sentido, a segurança do sistema se torna condicional e sistêmica, fruto da composição de múltiplos mecanismos técnicos e da validade das hipóteses de confiança residual assumidas no modelo de ameaça.

As propriedades de segurança da Computação Confidencial podem ser formalizadas como garantias condicionais: para cada propriedade Pi , a arquitetura provê Pi contra os adversários considerados no modelo de ameaça (Seção 3.1) se, e somente se, um conjunto de pressupostos de confiança residual e condições operacionais Hi,1, . . . , Hi,n permanecer válido. Em outras palavras, as condições H não representam ameaças que o sistema pretende mitigar; elas explicitam o escopo de validade das garantias. A Tabela 2 organiza esses pressupostos de maneira estruturada, evidenciando que a violação de qualquer condição pode comprometer a propriedade correspondente.

Tabela 2 Formalização condicional das propriedades de segurança da Computação Confidencial. Para cada propriedade, a garantia vale contra os adversários do modelo de ameaça assumido se, e somente se, todos os pressupostos de confiança residual e condições operacionais (H) listados permanecerem simultaneamente válidos.

Propriedade	**Pressupostos e condições necessárias (H)**
Confidential Processing	Implementação correta do TEE ∧ ausência de vulnerabilidades exploráveis no silício ∧ inexistência de canais laterais remotos de alta taxa de vazamento ∧ integridade do firmware ∧ canal autenticado e protegido para dispositivos externos.
Enforceable Guarantees	Atestação remota correta ∧ chaves raiz do fabricante não comprometidas ∧ validação adequada das evidências de atestação ∧ política de execução fail-closed.
Auditable Transparency	Logs públicos imutáveis ∧ operadores independentes dos logs ∧ capacidade efetiva de auditoria externa.
Protection Against Insider Threats	Isolamento em relação ao hipervisor e sistema operacional ∧ ausência de canais laterais exploráveis por usuários privilegiados ∧ integridade da cadeia de suprimentos de hardware.
Non-targetability	Instâncias de execução não associáveis a usuários específicos ∧ uso correto de protocolos oblivious ∧ ausência de correlação baseada em metadados.
Stateless Processing	Ausência de armazenamento persistente de dados ∧ descarte seguro de memória ao término da sessão ∧ inexistência de recuperação indevida via snapshots ou rollback.
Forward Security	Geração adequada de chaves efêmeras ∧ rotação correta de chaves ∧ eliminação de segredos expirados.

3.3 Considerações Sistêmicas e Limitações Estruturais

Levando em conta o modelo de ameaças (Seção 3.1) e a sua formalização como requisitos (Seção 3.2), nota-se que as garantias associadas à Computação Confidencial em arquiteturas baseadas em TEEs não eliminam a necessidade de confiança, mas promovem sua redistribuição estrutural ao longo da pilha tecnológica.

Tradicionalmente, a confiança concentrava-se no sistema operacional, no hypervisor e nos operadores da infraestrutura. No modelo baseado em TEEs, esse perímetro é deslocado para o hardware subjacente, para o firmware e para os mecanismos de atestação remota. O núcleo da confiança passa, portanto, a residir na correção do silício, na integridade da cadeia de suprimentos e na segurança das chaves raiz mantidas pelo fabricante.

Esse deslocamento reduz a exposição a adversários tradicionalmente privilegiados, como hypervisors comprometidos ou administradores de nuvem maliciosos. Contudo, introduz dependências estruturais em relação a componentes de hardware e firmware frequentemente proprietários, bem como a infraestruturas de certificação controladas por terceiros.

Consequentemente, vulnerabilidades sistêmicas nesses elementos podem produzir impactos amplificados, dada a centralidade que assumem no novo perímetro de confiança e a migração do cenário de ameaças para uma camada adicional, introduzindo a herança de ataques físicos já conhecidos (e.g., Fault Injection, Side Channel Attacks).

Arquiteturas que estendem o modelo de TEE para além da CPU, incorporando aceleradores externos ou dispositivos especializados (e.g., GPUs), evidenciam o aumento da complexidade sistêmica. Embora mecanismos de transporte autenticado e atestação preservem propriedades de confidencialidade e integridade, a introdução desses componentes amplia a superfície de ataque, eleva a complexidade dos processos de verificação e reforça a dependência de firmware e cadeias de confiança adicionais. Assim, a segurança resultante deve ser analisada sistematicamente, considerando a composição de múltiplos mecanismos interdependentes, não apenas a robustez isolada de primitivas criptográficas.

De modo análogo, a incorporação de protocolos de Comunicação Anônima (e.g., oblivious HTTP) demonstra que a proteção de dados em uso não é suficiente para mitigar riscos associados à exposição de metadados. A dissociação entre identidade do usuário, conteúdo da requisição e ambiente de execução constitui camada complementar de proteção, reforçando propriedades como non-targetability. Ainda assim, tais garantias permanecem condicionais à correta separação de papéis, à ausência de retenção indevida de metadados e à inexistência de vetores de correlação temporal ou comportamental.

É igualmente fundamental distinguir a Computação Confidencial da Criptografia Ponta-a-Ponta (End-to-End Encryption - E2EE). Enquanto a E2EE protege dados em trânsito entre remetente e destinatário, a Computação Confidencial protege dados durante o processamento. Ambos os mecanismos atuam em camadas distintas do ciclo de vida da informação e são conceitualmente complementares, não substitutivos. A adoção de processamento confidencial, portanto, não elimina a necessidade de E2EE, tampouco redefine automaticamente o modelo de confidencialidade percebido pelo usuário final.

Em síntese, a segurança do modelo de Computação Confidencial deve ser compreendida como emergente da composição de múltiplos mecanismos técnicos e da validade das hipóteses de confiança residual explicitadas no modelo de ameaça. Trata-se de um modelo condicional e sistêmico, no qual a violação de qualquer hipótese estrutural pode comprometer propriedades específicas do sistema. Dessa forma, uma avaliação crítica da Computação Confidencial deve examinar não ape-

nas as garantias declaradas, mas também as suposições subjacentes, às dependências institucionais e a superfície de ataque residual associada à crescente complexidade arquitetural.

4. Estudo de Caso

O Private Processing surgiu devido ao impulsionamento de ferramentas do tipo assistente pessoal baseadas em grandes modelos de linguagem (do inglês Large Language Models, LLMs), da parte de grandes fornecedores do mercado, como Google, Microsoft e a própria Meta. Proposto inicialmente para uso no WhatsApp, o objetivo do protocolo é que o usuário possa usar seu assistente pessoal para processar o conteúdo de seus chats privados, individualmente ou de forma agregada. Diferentemente da ferramenta já disponível nos aplicativos da Meta (a Meta AI), que precisa ser adicionada a uma conversa ou acionada em seu próprio chat privado, o Private Processing permitiria o processamento de toda a caixa de mensagens do usuário. Para tanto, a funcionalidade demanda uma capacidade de processamento superior à encontrada nos endpoints dos usuários, o que fez com que a Meta buscasse viabilizar o processamento das mensagens privadas em ambientes de computação em nuvem. Porém, o principal desafio surge quando as mesmas propriedades de segurança fornecidas pelo protocolo de privacidade padrão atual do WhatsApp, a criptografia ponta a ponta (JAN; BRIAN, 2025), precisam ser mantidas nesse novo cenário de processamento.

Com a manutenção das propriedades de segurança em mente, a Meta anunciou o mecanismo de Private Processing para o WhatsApp em abril de 2025 na tentativa de viabilizar funcionalidades opcionais de IA preservando a promessa de privacidade da plataforma (MARTIN, 2025). O objetivo do protocolo é que mensagens, chamadas e transferências de arquivos, seja em conversas individuais ou em grupo, estejam seguras contra adversários e contra o próprio provedor da plataforma. Contudo, a própria definição de criptografia ponta-a-ponta exige que o conteúdo das mensagens esteja sempre codificado ao deixar o dispositivo do usuário (a ponta), e somente seja decifrado ao atingir o dispositivo do destinatário (a outra ponta). Para que seja possível processar as mensagens dos usuários em ambientes de nuvem, as mesmas garantias de anonimidade dos participantes e confidencialidade e integralidade do conteúdo precisam ser mantidas. Para tanto, a Meta criou um pipeline de diversas tecnologias de computação confidencial para que as solicitações dos usuários sejam encaminhadas de forma anônima até ambientes de execução confiáveis. Esse pipeline inclui o dispositivo do usuário, serviços intermediários de encaminhamento, ambientes isolados baseados em TEEs e componentes especializados de processamento. O objetivo é que nenhum componente individual tenha acesso simultâneo à identidade do usuário, ao conteúdo da requisição e ao ambiente de execução.

A arquitetura do Private Processing proposta pela Meta tem a intenção de permitir o processamento do conteúdo de mensagens privadas de usuários em serviços baseados em nuvem sem que o provedor do serviço, operadores de infraestrutura ou terceiros tenham acesso ao conteúdo processado ou consigam associá-lo a usuários específicos. Para esse objetivo, o sistema combina isolamento de execução, criptografia ponta-a-ponta, se paração funcional de componentes e mecanismos de transparência verificável seguindo os fundamentos previstos para computação confidencial (FENG et al., 2024) e comunicação anônima (REN; WU, 2010) trazidos anteriormente neste capítulo. Essa seção aborda especificamente os componentes de software e hardware usados na implementação do Private Processing segundo o whitepaper divulgado pela Meta (META, 2025), e segue com uma breve análise de segurança do protocolo.

4.1 Implementação do Private Processing

O Private Processing utiliza um conjunto de tecnologias para implementar os paradigmas de computação confidencial e anônima em ambientes de processamento em nuvem discutidos neste capítulo.

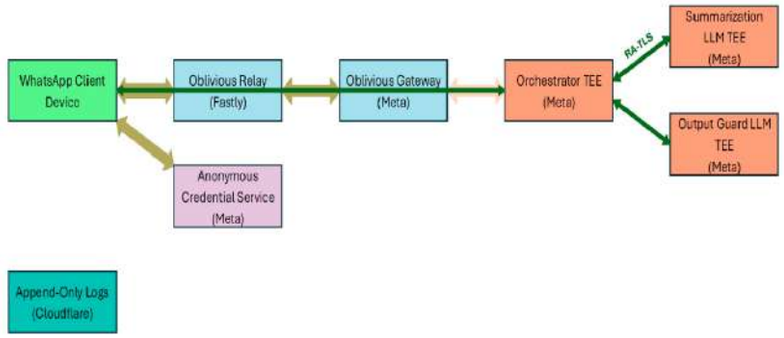
A Figura 3 sintetiza a arquitetura do protocolo e o fluxo de dados associado (NCC GROUP, 2025). O cliente do WhatsApp inicia a requisição e estabelece comunicação com um Oblivious Relay, que atua como intermediário de encaminhamento da solicitação. Em paralelo, o uso de um serviço de credenciais anônimas (ACS) é usado para autenticação funcional sem a revelação da identidade do usuário ao restante da infraestrutura. O Oblivious Gateway direciona as requisições para ambientes de execução confidenciais. O Orchestrator TEE é o responsável por coordenar a execução das tarefas dentro dos ambientes de execução confiáveis (TEEs), após a verificação do ambiente por meio de atestação remota (RA-TLS). Esta seção traz detalhes técnicos sobre cada um desses componentes em sua versão utilizada pela Meta no Private Processing.

Se tratando do ambiente de execução confiável, o Private Processing especifica um TEE como sendo um ambiente de execução onde uma entidade não autorizada não pode visualizar, adicionar ou remover dados ou o código sendo usado. Dentre outras definições existentes, esta se baseia na publicação do Confidential Computing Consortium (CONFIDENTIAL COMPUTING CONSORTIUM, 2022a). Essa definição também inclui técnicas baseadas em hardware para prover as garantias de segurança necessárias, conforme disponibilizado por cada fabricante. O hardware utilizado pela Meta conforme descrição no whitepaper do Private Processing é fornecido pela AMD e NVIDIA.

Da parte da AMD, a tecnologia por trás da computação confidencial é chamada de Secure Encrypted Virtualization - Secure Nested Paging (SEV-SNP), um mecanismo para isolamento e proteção de máquinas virtuais baseado em hardware (AMD, 2020) disponível para os processadores da linha EPYC. Porém, um ponto de atenção diz respeito à integração do processamento com GPUs da NVIDIA. Os dados são transportados para fora da VM rodando na CPU utilizando um túnel SPDM (DMTF, 2023) estabelecido por um driver da NVIDIA. Enquanto algumas definições de TEE estabelecem que os dados não deveriam sair de dentro do ambiente durante o processamento (KNODEL et al., 2024), o hardware da NVIDIA promete garantir a atestação da GPU e a proteção da memória, interface e comunicações.

O Private Processing usa OHTTP para anonimizar as requisições HTTP feitas aos servidores da Meta. Requisições HTTP revelam informações sobre as identidades dos clientes mesmo onde um endereço IP não está diretamente associado a um indivíduo. As requisições feitas a partir de um IP podem ser correlacionadas

Figura 3 Arquitetura simplificada do Private Processing, evidenciando o fluxo requisições do cliente até os ambientes de execução confidenciais (TEEs) (NCC GROUP, 2025)



ao longo do tempo para montar um perfil de comportamento do cliente. O uso de um proxy HTTP é uma estratégia para aprimorar a privacidade do protocolo. Porém, esses proxies devem ser usados com Transport Layer Security (TLS), do contrário o conteúdo da comunicação é revelado ao próprio proxy. Contudo, o TLS demanda que uma nova conexão seja aberta para cada requisição, causando sobrecarga no servidor. O Oblivious HTTP (OHTTP) foi desenvolvido para o encaminhamento de mensagens HTTP criptografadas (THOMSON; WOOD, 2024), evitando as características negativas do uso de proxies HTTP configurados pelo cliente. O OHTTP permite que um cliente faça múltiplas requisições a um servidor sem que este seja capaz de vincular essas requisições ao cliente ou de identificar que as requisições vieram de um mesmo cliente. Nós intermediários com confiança limitada são utilizados para encaminhar mensagens. O OHTTP utiliza um algoritmo de Criptografia de Chave Pública Híbrida (Hybrid Public Key Encryption, HPKE), definido na RFC 9180, (BARNES et al., 2022), para encapsular mensagens HTTP e proteger seu conteúdo. Além disso, há a utilização de um método de encaminhamento em que as requisições trafegam entre o cliente e um gateway por meio de um relay. A combinação de encapsulamento e retransmissão garante que o gateway nunca tenha acesso ao endereço IP do cliente, enquanto o relay jamais visualiza o conteúdo da mensagem em texto plano.

A atestação remota é um componente crucial nesse cenário. O TLS prova quem é o servidor mas não dá garantias sobre a conformidade da computação executada. Um servidor legítimo pode estar comprometido e executando uma versão modificada do código original, introduzindo assim vulnerabilidades em seus resultados. O RA-TLS combina ambas as garantias: o certificado com a identidade do servidor com um relatório criptográfico assinado pelo hardware. Essa atestação prova

matematicamente que o servidor está operando dentro de um TEE, e que o software sendo executado é o código oficial e auditado sem modificações. Na implementação do Private Processing, esse relatório criptográfico cobre o código rodando no sistema, sua versão de firmware, kernel, opções de CPU habilitadas, entre outras. A confiança na atestação tem origem no hardware SEV-SNP da AMD. Um Agente de Atestação é iniciado na máquina virtual operando o TEE, responsável por gerar os certificados que "capturam" o estado atual de execução do ambiente, assiná-los usando uma chave RSA, e encaminhá-los através da conexão TLS.

Não é o objetivo desse capítulo trazer uma explicação e análise extensiva do Private Processing proposto pela Meta, apenas ilustrar de modo geral os componentes e tecnologias utilizados como um caso de uso de Computação Confidencial. O whitepaper original (META, 2025) traz várias camadas extras de detalhes sobre cada componente, assim como diversos outros componentes deixados de fora desse documento por brevidade. A seguir, será feita uma análise das propriedades de segurança oferecidas pelo Private Processing considerando a combinação de componentes descritos até aqui.

4.2 Análise de Segurança

O Private Processing está sujeito às mesmas propriedades de segurança descritas na Seção 3 desse capítulo.

Uma quantidade grande de mecanismos adicionais de segurança aumenta a complexidade do protocolo e dificulta sua auditoria, além de potencialmente introduzir novos pontos de vulnerabilidade a serem explorados por um atacante. Um ponto relevante de atenção na análise de segurança dos TEEs da Meta é o transporte dos dados para fora da VM confidencial fornecida pelo SEV-SNP para o processamento em uma

GPU da NVIDIA. Enquanto diversas outras técnicas são utilizadas para manter a atestação e confidencialidade da computação feita em GPU, o acesso direto à infraestrutura onde os componentes se encontram é um cenário de risco para a possível interceptação de informações antes do seu envio para o ambiente confidencial. Este cenário é possível quando a validação dos atributos das primitivas disponíveis para a atestação da GPU não são correlacionadas à lista de fingerprinting[11] do hardware utilizado no fluxo e validado por auditoria. Avaliando a segurança de componentes do Private Processing mais próximos ao processamento das informações pela LLM, o protocolo também apresenta a possibilidade de sofrer Side Channel Attacks pela forma que caches usualmente são utilizadas no processamento das mensagens enviadas para LLMs durante a etapa de tokenização das informações[12]; e também pela falta de mascaramento da janela de inferência da LLM nas suas respostas. Apesar do escopo das camadas de controle e proteção do protocolo diminuírem a probabilidade deste cenário de extração de dados se tornar real, o risco se mantém presente.

Outra consideração importante sobre o protocolo é o gerenciamento de chaves em seu fluxo, principalmente chaves relacionadas aos componentes Oblivious Relay e o serviço de credenciais anônimas (ACS). Apesar dos serviços segregarem controles absolutos aplicados ao gerenciamento das chaves para a mitigação do risco de desanonimização dos usuários dentro do fluxo, o cenário de risco ainda depende diretamente

11. Fingerprints (ou impressões digitais), nesse contexto, são informações coletadas sobre o software e/ou o hardware de um dispositivo computacional remoto, e podem ser utilizadas para identificar o dispositivo.

12. Aplicações de LLM utilizam cache para evitar reprocessar prompts longos. Um tipo de ataque de canal lateral envia prompts e mede seu tempo de resposta. Uma resposta rápida indica um cachehit, confirmando que certas informações (como documentos confidenciais ou instruções específicas) estão presentes na cache.

do isolamento dos contextos das empresas atuando nestes serviços. Isso significa que a garantia plena de não-rastreabilidade ainda depende da não-colusão entre a Meta e seus provedores externos (e.g., Fastly). Outro aspecto fundamental é a necessidade contínua de transparência e auditoria do código, para garantir que os artefatos e binários previamente auditados são os mesmos sendo empregados durante a execução.

Por fim, analisando o escopo de uso dos TEEs adotados no Private Processing para o WhatsApp, é relevante comparar suas propriedades de segurança com o protocolo atualmente em uso, a criptografia ponta-a-ponta, ou E2EE. Mesmo se implementados corretamente, TEEs fornecem propriedades de segurança ortogonais ao E2EE, ou seja, os dois protocolos não desempenham o mesmo papel e não são substitutivos (KNODEL et al., 2024). Durante o uso do Private Processing, é crucial que a aplicação notifique e conscientize o usuário sobre quaisquer mudanças no protocolo de privacidade utilizado.

5. Comentários Finais

O atual cenário de adoção de serviços em nuvem e modelos de IA intensifica os desafios de governança sobre dados críticos. Os métodos tradicionais de proteção são eficazes para dados armazenados ou trafegados, mas falham em isolar o dado durante o ciclo de processamento. A Computação Confidencial surge para fechar essa lacuna de segurança. O paradigma traz um conjunto de técnicas e arquiteturas que permitem a execução de cargas de trabalho em ambientes isolados, oferecendo garantias formais quanto à confidencialidade e à integridade dos dados e do código em execução. Grandes empresas de tecnologia têm evoluído para a adoção da Computação Confiável, impulsionadas pela necessidade de operar sobre dados sensíveis, cumprir requisitos regulatórios e fortalecer a confiança entre usuários, provedores de serviços e organizações.

O presente documento discute a Computação Confidencial e, em particular, os aspectos relacionados com as propriedades de segurança nas soluções que a implementam. A maioria dessas soluções se concentram na atestação e na criptografia de memória. Como estudo de caso ilustrativo, é analisada o Private Processing, solução apresentada pela Meta no contexto do WhatsApp. Essa solução busca permitir que usuários submetam solicitações a serviços de inteligência artificial para o processamento de mensagens fora de seus dispositivos, como a geração de resumos de conversas, sem que o conteúdo seja armazenado de forma persistente após o término da sessão.

Embora a Computação Confidencial ofereça avanços em relação a alternativas conhecidas, sua implementação enfrenta desafios significativos. O isolamento de dados e aplicações introduz sobrecarga de processamento devido a mecanismos criptográficos, o que pode reduzir a velocidade de processamento. Além disso, existe uma barreira de desenvolvimento já que é necessário adaptar softwares legados para rodar dentro desses ambientes. Finalmente, a fragmentação do ecossistema dificulta a adoção em massa, já que diferentes provedores de nuvem e fabricantes de chips utilizam padrões distintos. Assim, as cargas de trabalho não são portáveis sem comprometer as propriedades de segurança.

Referências

AMD. AMD SEV-SNP: Strengthening VM Isolation with Integrity Protection and More. 2020. Disponível em: https://docs.amd.com/v/u/en-US/SEV-SNP-strengthening-vm-isolation-with-integrity-protection-and-more. Acesso em: 15 jan. 2026.

BARNES, R.; BHARGAVAN, K.; LIPP, B.; WOOD, C. A. Hybrid Public Key Encryption. **RFC 9180**, fev. 2022. Disponível em: https://www.rfc-editor.org/info/rfc9180.

CONFIDENTIAL COMPUTING CONSORTIUM. A technical analysis of confidential computing. 2022a. Disponível em: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf. Acesso em: 15 jan. 2026.

CONFIDENTIAL COMPUTING CONSORTIUM. Common terminology for confidential computing. 2022b. Disponível em: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/Common-Terminology-for-Confidential--Computing.pdf. Acesso em: 3 mar. 2026.

CONFIDENTIAL COMPUTING CONSORTIUM. Confidential computing: hardware-based trusted execution for applications and data. 2022c. Disponível em: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC_outreach_whitepaper_updated_November_2022.pdf. Acesso em: 3 mar. 2026.

COSTAN, V.; DEVADAS, S. Intel SGX explained. **Cryptology ePrint Archive**, Paper 2016/086, 2016. Disponível em: https://eprint.iacr.org/2016/086.

DMTF. Security Protocol and Data Model (SPDM) Specification, version 1.3.0. **DSP0274**. 2023.

FENG, D.; QIN, Y.; FENG, W.; LI, W.; SHANG, K.; MA, H. Survey of research on confidential computing. **IET Communications**, v. 18, n. 9, p. 535–556, 2024.

GOLDSCHLAG, D. M.; REED, M. G.; SYVERSON, P. F. Hiding routing information. **Information Hiding**, v. 1174, p. 137–150, 1996.

HESS, T.; HUSSAIN, S.; KAOUDIS, K.; MILLER, J.; SMITH, C.; CAO, A. Meta WhatsApp Private Processing: Security Assessment with Fix Review. **Trail of Bits**, 2025. Disponível em: https://github.com/trailofbits/publications/blob/

master/reviews/2025-08-meta-whatsapp-privateprocessing-securityreview.pdf. Acesso em: 3 mar. 2026.

JAN; BRIAN. End-to-end encryption. 2025. Disponível em: https://blog.whatsapp.com/end-to-end-encryption. Acesso em: 3 mar. 2026.

KNODEL, M.; FÁBREGA, A.; FERRARI, D.; LEIKEN, J.; LI HOU, B.; YEN, D.; DE ALFARO, S.; CHO, K.; PARK, S. How to think about end-to-end encryption and AI: Training, processing, disclosure, and consent. **Cryptology ePrint Archive**, Paper 2024/2086, 2024. Disponível em: https://eprint.iacr.org/2024/2086.

MARTIN, A. WhatsApp says in-app AI tools will still keep messages secret. **The Record (Recorded Future News)**, abr. 2025. Disponível em: https://therecord.media/whatsapp-in-app-tools-secret-messages. Acesso em: 3 mar. 2026.

META. Private Processing for WhatsApp Overview. Technical White Paper and Security Guide. jun. 2025. Version 1. Disponível em: https://web.archive.org/web/20260226124239/https://ai.meta.com/static-resource/private-processing-technical-whitepaper. Acesso em: 3 mar. 2026.

NCC GROUP. Public Report: Meta WhatsApp Message Summarization Service. 2025. Disponível em: https://www.nccgroup.com/research-blog/public-report-meta-whatsapp-message-summarization-service/. Acesso em: 12 jan. 2026.

REN, J.; WU, J. Survey on anonymous communications in computer networks. **Computer Communications**, v. 33, n. 4, p. 420–431, 2010.

SABT, M.; ACHEMLAL, M.; BOUABDALLAH, A. Trusted execution environment: What it is, and what it is not. In: **IEEE TRUSTCOM/BIGDATASE/ISPA**, 2015. Anais [...]. v. 1, p. 57–64, 2015. DOI: 10.1109/Trustcom.2015.357.

THE LINUX FOUNDATION PROJECTS. Confidential Computing Consortium. [s.d.]. Disponível em: https://confidentialcomputing.io/. Acesso em: 5 jan. 2026.

THOMSON, M.; WOOD, C. A. Oblivious HTTP. **RFC 9458**, jan. 2024. Disponível em: https://www.rfc-editor.org/info/rfc9458.

Capítulo V
Análise Sociotécnica

Criptografia, confiança e governança: uma análise sociotécnica da moderação de conteúdo em sistemas de comunicação fim a fim

Rodolfo Silva Avelino[1]

Resumo: A criptografia de ponta a ponta (end-to-end encryption – E2EE) tem sido amplamente adotada como mecanismo técnico de proteção da privacidade e da segurança das comunicações digitais. Contudo, sua implementação em larga escala produz efeitos que extrapolam o domínio estritamente técnico, reconfigurando relações de poder, confiança e responsabilidade nas infraestruturas comunicacionais contemporâneas. Este artigo propõe uma análise sociotécnica da moderação de conteúdo em sistemas de comunicação protegidos por E2EE, compreendendo-a como um experimento institucional que tensiona os limites da governança digital centralizada. Argumenta-se que a criptografia não deve ser entendida apenas como matemática aplicada, mas como uma tecnologia política que redistribui capacidades de controle, desloca a autoridade observacional e fragmenta a responsabilidade entre plataformas, Estados, usuários e arquiteturas técnicas. A partir da aná-

1. Professor do Insper, doutor em Ciências Humanas e Sociais pela UFABC, com atuação na interseção entre tecnologia, poder e sociedade, com ênfase em cibersegurança, governança digital e soberania tecnológica. Autor de Colonialismo Digital, membro do CGI.br e do CITDigital, com mais de 30 anos de experiência em segurança da informação, infraestrutura tecnológica e políticas públicas digitais.

lise dos fundamentos técnicos da E2EE, dos mecanismos de moderação disponíveis em ambientes criptografados e das propostas institucionais de reintrodução de observabilidade (em especial a varredura no lado do cliente) o artigo demonstra que a moderação de conteúdo em sistemas E2EE não é eliminada, mas transformada em um processo fragmentado, reativo e distribuído. Por fim, discute-se como arranjos alternativos de governança, como implementações baseadas em Software Livre e plataformas federadas, evidenciam que os dilemas entre privacidade, segurança pública e responsabilização decorrem de escolhas arquiteturais e institucionais, e não de limitações técnicas da criptografia.

Palavras-chave: Criptografia de ponta a ponta; Moderação de conteúdo; Confiança algorítmica; Governança digital; Análise sociotécnica.

Abstract: End-to-end encryption (E2EE) has been widely adopted as a technical mechanism to protect privacy and security in digital communications. However, its large-scale deployment produces effects that extend beyond purely technical concerns, reshaping relations of power, trust, and responsibility within contemporary communication infrastructures. This article presents a sociotechnical analysis of content moderation in E2EE-based communication systems, framing it as an institutional experiment that exposes the limits of centralized digital governance. It argues that cryptography should not be understood merely as applied mathematics, but as a political technology that redistributes control capabilities, displaces observational authority, and fragments responsibility among platforms, states, users, and technical architectures. By examining the technical foundations of E2EE, the available moderation mechanisms in encrypted environments, and institutional proposals aimed at reintroducing observability (particularly client-side scanning) the article demonstrates that content moderation in E2EE systems is not eliminated, but transformed into a fragmented, reactive, and distributed process. Finally, it discusses how alternative governance arrangements, such as open-source implementations and federated platforms, highlight that the tensions between privacy, public security, and accountability stem from architectural and institutional choices rather than from technical shortcomings of cryptography.

Keywords: end-to-end encryption; content moderation; algorithmic trust; digital governance; sociotechnical analysis.

1. Introdução aos conceitos fundamentais

O desenvolvimento dos protocolos criptográficos constituiu uma força fundamental na construção dos sistemas de comunicação segura, evoluindo por meio de uma série de marcos significativos impulsionados pela crescente demanda por privacidade e proteção de dados. Os

Figura 1 Esquema ilustrativo do funcionamento da criptografia de ponta a ponta (E2EE).

primeiros esforços no campo da criptografia concentraram-se predominantemente em cifras clássicas, como a cifra de César e a máquina Enigma, que estabeleceram as bases para a compreensão da criptografia como um meio de proteger informações contra acessos não autorizados. No entanto, esses métodos iniciais apresentavam limitações significativas quanto ao escopo e ao nível de segurança, sendo utilizados principalmente em comunicações de natureza militar e diplomática.

O surgimento da criptografia de chave pública, na década de 1970, representou uma mudança paradigmática no campo. Protocolos como o RSA (Rivest–Shamir–Adleman) introduziram o conceito de criptografia assimétrica, possibilitando a troca segura de chaves por canais inseguros sem a necessidade de um segredo previamente compartilhado. Esse avanço foi decisivo para o desenvolvimento de protocolos de comunicação digital segura e sustentou a consolidação da internet como uma infraestrutura voltada a transações confidenciais e à circulação protegida de informações (SOOD; KAUR, 2023).

Posteriormente, a introdução de algoritmos de criptografia simétrica, como o AES (Advanced Encryption Standard), ofereceu meios robustos e eficientes para a proteção de grandes volumes de dados, conciliando elevados níveis de segurança com desempenho computacional adequado. A integração dessas técnicas criptográficas a protocolos de comunicação amplamente utilizados, como SSL/TLS, viabilizou a navegação segura na web e o comércio eletrônico, contribuindo de forma decisiva para o fortalecimento da confiança dos usuários em ambientes digitais (HAZRA et al., 2024; REDDY et al., 2024).

A consolidação dos sistemas de criptografia de ponta a ponta (end-to-end encryption – E2EE) representou um avanço adicional no campo da privacidade digital, ao garantir que apenas os usuários diretamente envolvidos na comunicação sejam capazes de descriptografar o conteúdo das mensagens. Ao excluir intermediários (incluindo provedores de serviços) do acesso ao conteúdo transmitido, a E2EE reduz significativamente a superfície de vigilância e interceptação. Protocolos como o Signal Protocol, amplamente adotados em aplicações de mensagens como WhatsApp e Signal, exemplificam essa abordagem ao combinar sigilo futuro (forward secrecy), mecanismos de negação plausível e autenticação robusta para proteger as comunicações contra múltiplos vetores de ataque.

As motivações para a adoção desses protocolos criptográficos são multifacetadas. As preocupações com a privacidade intensificaram-se paralelamente à expansão das comunicações digitais, levando usuários a buscar proteção contra vigilância em massa, vazamentos de dados e práticas de coleta não autorizada de informações. Pressões regulatórias e escândalos de grande repercussão envolvendo dados pessoais reforçaram a necessidade de mecanismos criptográficos robustos capazes de salvaguardar informações sensíveis.

Simultaneamente, considerações de segurança impulsionaram o desenvolvimento contínuo de protocolos voltados ao enfrentamento de ameaças cibernéticas em constante evolução, como interceptação, adulteração de mensagens e falsificação de identidade. Nesse contexto, os protocolos criptográficos operam como ferramentas fundamentais para assegurar confidencialidade, integridade e autenticidade nos sistemas de comunicação, viabilizando interações digitais seguras que se tornaram essenciais para o comércio, a governança e as dinâmicas sociais contemporâneas.

Ao assegurar privacidade e autonomia individual, a criptografia não transforma apenas as práticas de proteção de dados, mas também a própria estrutura das relações sociais e políticas mediadas por tecnologias digitais. Ao promover a confiança algorítmica, reduz-se a dependência de formas tradicionais de confiança depositadas em instituições estatais ou corporativas, inaugurando um paradigma no qual a validação das

interações ocorre por meio de protocolos técnicos rigorosos. Esse deslocamento implica que plataformas digitais e interfaces sociais, historicamente centrais na mediação da comunicação, percam parte de seu controle exclusivo sobre os fluxos informacionais, abrindo espaço para sistemas mais distribuídos, ainda que sujeitos a desafios persistentes de governança e regulação.

Essa reconfiguração, contudo, introduz tensões significativas para os Estados, cuja capacidade de monitorar e regular o ambiente digital é enfraquecida, especialmente em contextos de crise, nos quais demandas por segurança pública exigem respostas rápidas e eficazes. Emergiriam, assim, sociedades criptográficas híbridas, nas quais a autoridade deixa de estar concentrada em agentes humanos específicos e passa a ser distribuída por mecanismos algorítmicos responsáveis por assegurar a integridade e a transparência das comunicações e transações. Tal fenômeno evidencia uma dualidade complexa: enquanto ocorre a descentralização do poder técnico, intensifica-se a necessidade de novos arranjos institucionais capazes de equilibrar liberdade individual, segurança coletiva e responsabilidade social.

As plataformas de comunicação que adotam criptografia de ponta a ponta permanecem formalmente responsáveis pelo conteúdo que circula em seus sistemas, ainda que não tenham acesso direto às mensagens trocadas entre usuários. O exercício dessa responsabilidade (central nos regimes contemporâneos de moderação de conteúdo) torna-se substancialmente mais restrito em ambientes protegidos por E2EE, sem, contudo, ser completamente eliminado. Nesses contextos, a moderação deixa de operar como uma prática centralizada e preventiva, passando a assumir formas fragmentadas, reativas e distribuídas entre diferentes atores humanos e técnicos. A moderação de conteúdo em sistemas de comunicação protegidos por E2EE pode, assim, ser compreendida como um experimento sociotécnico que tensiona as fronteiras entre

criptografia, segurança pública e liberdade de expressão (BARTUSEK et al., 2023; DIAS OLIVA, 2020; SCHEFFLER; MAYER, 2023).

Do ponto de vista técnico, a criptografia de ponta a ponta baseia-se em um conjunto de procedimentos criptográficos que garantem que apenas os dispositivos terminais envolvidos em uma comunicação sejam capazes de acessar o conteúdo das mensagens trocadas. Em sistemas E2EE, o processo inicia-se com mecanismos de autenticação mútua, nos quais os usuários estabelecem e validam identidades criptográficas por meio da troca e verificação de chaves públicas ou certificados digitais. Em seguida, ocorre um processo autenticado de troca de chaves, no qual os participantes negociam diretamente (e sem mediação de terceiros) os segredos criptográficos necessários à comunicação. Essas chaves são, em geral, efêmeras, válidas apenas para sessões ou mensagens específicas, o que limita a reutilização de credenciais e reduz a superfície de ataque. Uma vez estabelecidas, as chaves permitem que o conteúdo seja protegido por algoritmos simétricos robustos, operando em modos de criptografia autenticada que asseguram simultaneamente confidencialidade e integridade das mensagens. Essa arquitetura técnica, ao excluir intermediários do acesso ao conteúdo, não apenas fortalece a segurança das comunicações, mas também fundamenta a noção de ausência de autoridade intermediária que sustenta os debates contemporâneos sobre confiança, responsabilidade e moderação em ambientes criptografados.

Este artigo adota uma abordagem qualitativa de natureza teórica e conceitual, orientada por uma perspectiva sociotécnica inspirada nos estudos de ciência, tecnologia e sociedade. A análise parte da compreensão de que sistemas técnicos, como a criptografia de ponta a ponta, não podem ser dissociados dos arranjos institucionais e das relações de poder nas quais se inserem. Nesse sentido, o trabalho não se propõe a realizar investigação empírica, mas a desenvolver uma análise crítica

baseada em revisão bibliográfica interdisciplinar, articulando literatura técnica, jurídica e sociológica.

A estratégia analítica consiste em examinar como diferentes arquiteturas comunicacionais (em especial sistemas E2EE, mecanismos de varredura no lado do cliente e modelos alternativos de governança) reconfiguram a distribuição de autoridade, responsabilidade e visibilidade, produzindo efeitos institucionais que tensionam os regimes contemporâneos de moderação de conteúdo.

2. Fundamentos da criptografia como tecnologia de poder

A história da criptografia indica que práticas sistemáticas de cifragem antecedem em muito a modernidade. A cifra tradicionalmente atribuída a Júlio César é descrita por Suetônio em As vidas dos doze Césares, constituindo um dos exemplos mais conhecidos de criptografia clássica no Ocidente (SUETÔNIO, 2012). No entanto, evidências históricas sugerem que o uso da criptografia como instrumento político e militar remonta ao menos ao século III a.C. (KAHN, 1996). De modo paralelo, registros históricos indicam que, desde o início da Era Comum, práticas de comunicação cifrada eram empregadas na China imperial como forma de proteger informações estratégicas contra interceptação, especialmente em contextos administrativos e militares (NEEDHAM, 1959). Nesse sentido, a criptografia pode ser compreendida como um dispositivo sociotécnico de proteção informacional, funcionando como um reduto que assegura confidencialidade comunicacional em contextos específicos. Ao longo da história, esse reduto foi alternadamente interpretado como um recurso legítimo de segurança, uma ameaça à autoridade constituída ou, frequentemente, como ambos simultaneamente, evidenciando o caráter ambíguo da criptografia enquanto tecnologia de poder.

Os sistemas de comunicação contemporâneos incorporam a criptografia de ponta a ponta como um protocolo que permite que somente os comunicantes tenham acesso ao conteúdo das mensagens. A E2EE realoca parte do poder de moderação de conteúdo das plataformas para os usuários, embora nem toda forma de moderação se torne impossível, e a dificuldade de moderação em diversos cenários não dispense a responsabilidade das plataformas. A adoção da E2EE reforça, assim, o argumento de que a moderação de conteúdo é uma operação inerentemente complexa, que não pode ser realizada exclusivamente por meios automatizados nem delegada integralmente a algoritmos.

Além de um instrumento técnico, a criptografia pode ser compreendida como uma infraestrutura política que reorganiza relações de poder entre Estados, corporações e indivíduos. Conforme argumenta David Lyon (2018), tecnologias de comunicação não são neutras, mas moldam regimes de visibilidade e invisibilidade que afetam diretamente práticas de vigilância, governança e controle social. Nesse contexto, a criptografia de ponta a ponta atua como um mecanismo de resistência estrutural à vigilância, ao limitar a capacidade de Estados e plataformas de acessar conteúdos comunicacionais, deslocando o eixo do poder informacional. Esse deslocamento, contudo, não elimina conflitos políticos, mas os reconfigura, gerando novas disputas em torno da legitimidade do acesso estatal, da segurança pública e da proteção de direitos fundamentais, como a privacidade e a liberdade de expressão.

A reação institucional à disseminação da E2EE (visível em propostas como o escaneamento no lado do cliente, o enfraquecimento criptográfico ou a criação de "acessos excepcionais") revela que o debate não se restringe à eficácia técnica da criptografia, mas envolve

a definição de quem deve decidir sobre os limites da visibilidade comunicacional. Como apontam Abelson et al. (2015) e Green e Smith (2016), tentativas de conciliar criptografia forte com mecanismos de acesso governamental tendem a introduzir vulnerabilidades sistêmicas, afetando de forma desproporcional usuários comuns e grupos politicamente vulneráveis. Assim, a criptografia emerge como uma tecnologia de poder ambígua: ao mesmo tempo em que protege indivíduos contra formas de vigilância arbitrária, torna-se alvo de disputas políticas que buscam reinscrever, por meios técnicos, formas historicamente consolidadas de controle e assimetria informacional.

A perspectiva apresentada no manifesto e nas conversas reunidas em Cypherpunks: Freedom and the Future of the Internet reforça essa compreensão da criptografia como uma tecnologia de poder deliberadamente política. Para Julian Assange e os autores associados ao movimento cypherpunk, a criptografia não é apenas um mecanismo de proteção da privacidade individual, mas uma ferramenta de redistribuição assimétrica de poder, capaz de enfraquecer estruturas centralizadas de vigilância e coerção estatal (ASSANGE et al., 2012). Nesse enquadramento, a criptografia forte atua como um meio técnico de limitar o alcance do poder institucional, ao tornar a vigilância em larga escala economicamente e operacionalmente inviável. A oposição sistemática de Estados e grandes plataformas à adoção irrestrita da criptografia de ponta a ponta pode, portanto, ser interpretada não apenas como uma preocupação com a segurança pública, mas como resistência à perda de controle informacional. O embate em torno da E2EE revela, assim, uma tensão fundamental entre projetos concorrentes de ordem política: de um lado, a defesa da autonomia comunicacional e da autodeterminação informacional; de outro, a manutenção de regimes de monitoramento que historicamente sustentaram formas de governança, disciplina e poder.

3. Moderação de conteúdo em sistemas E2EE: Conceitos e implicações

Sistemas de comunicação protegidos por criptografia de ponta a ponta (E2EE), nos quais o conteúdo das mensagens é inacessível a intermediários, são frequentemente apresentados como um obstáculo estrutural à moderação de conteúdo. Essa percepção decorre da suposição de que a moderação (compreendida como um conjunto de práticas destinadas a reduzir danos, garantir segurança e proteger direitos dos usuários) depende necessariamente do acesso centralizado ao conteúdo em trânsito. Sob essa ótica, a ausência de uma autoridade capaz de inspecionar mensagens seria interpretada como uma falha de projeto. No entanto, compreender as implicações da E2EE para a moderação de conteúdo exige uma análise mais ampla, que considere não apenas os mecanismos técnicos disponíveis, mas também os arranjos institucionais, jurídicos e sociotécnicos nos quais a moderação se insere.

Nos sistemas de comunicação convencionais, a moderação de conteúdo é exercida majoritariamente pelas plataformas, que operam como entidades privadas responsáveis por regular as interações entre usuários e por mitigar riscos jurídicos por meio de termos de uso e contratos de adesão. Esse modelo implica a privatização e a monetização da moderação, mas não assegura que a responsabilidade formal atribuída às plataformas coincida com a responsabilidade efetiva pela produção e circulação de danos. Em sistemas E2EE, a moderação não deixa de existir; ela deixa de ser uma função centralizada e preventiva, passando a assumir formas distribuídas, reativas e contextuais, desempenhadas por diferentes atores (usuários, autoridades estatais, instituições judiciais e mecanismos técnicos complementares) em momentos distintos do ciclo comunicacional.

É nesse contexto que emerge a proposta de varredura no lado do cliente (client-side scanning - CSS), apresentada por seus defensores como uma solução intermediária entre criptografia forte e demandas por moderação e aplicação da lei. Em vez de enfraquecer a criptografia ou introduzir portas de acesso excepcionais, a CSS propõe a análise automatizada de conteúdos diretamente nos dispositivos dos usuários, antes da criptografia ou após a descriptografia local. Caso seja detectado conteúdo considerado alvo, apenas metadados ou alertas seriam encaminhados a servidores centrais ou às autoridades competentes. Essa abordagem foi amplamente analisada e criticada por Abelson et al. (2021), que argumentam que a CSS não constitui uma forma limitada de moderação, mas sim um mecanismo de vigilância distribuída em escala populacional.

Os sistemas de varredura no lado do cliente baseiam-se fundamentalmente em dois paradigmas técnicos: hashing perceptual e métodos de classificação baseados em aprendizado de máquina, podendo empregar arquiteturas híbridas que combinam ambos. Ainda assim, essas variações não ampliam o espaço de projeto, uma vez que todo sistema de CSS depende, em última instância, da comparação com conjuntos previamente definidos ou da inferência estatística sobre classes de conteúdo. O hashing perceptual permite gerar impressões digitais resistentes a pequenas modificações em imagens, possibilitando a comparação entre conteúdos locais e bancos de dados de material previamente classificado, como no caso de imagens de abuso sexual infantil (ABELSON et al., 2021). Já os sistemas baseados em aprendizado de máquina são treinados para identificar padrões associados a determinadas classes de conteúdo, incluindo imagens inéditas, vídeos ou textos. Embora distintos em funcionamento, ambos os métodos compartilham características críticas: requerem acesso a conteúdo não criptografado, apresentam taxas inevitáveis de falsos positivos e falsos negativos e dependem de listas ou modelos cuja curadoria é opaca para os usuários.

A introdução da CSS desloca de forma significativa as fronteiras tradicionais da moderação de conteúdo. Diferentemente da moderação realizada em servidores, que incide apenas sobre conteúdos compartilhados, a CSS permite a inspeção de dados armazenados localmente nos dispositivos, incluindo informações que nunca seriam transmitidas ou tornadas públicas. Conforme argumentam Abelson et al. (2021), essa característica transforma a CSS em uma forma de interceptação em massa distribuída, com riscos substanciais à privacidade, à segurança e à liberdade de expressão. Além disso, a arquitetura da CSS cria incentivos para a expansão progressiva de seu escopo, tanto em termos dos tipos de conteúdo analisados quanto das finalidades políticas e jurídicas da varredura.

Nesse sentido, a CSS não resolve as tensões entre E2EE e moderação de conteúdo, mas as reconfigura em um plano mais profundo. Ao transferir práticas de controle do nível da plataforma para o nível do dispositivo, a CSS redefine a própria noção de responsabilidade, deslocando-a do espaço público da governança das plataformas para a esfera privada dos usuários. Como resultado, a moderação em ambientes E2EE deixa de ser apenas um problema técnico ou jurídico e passa a constituir um experimento sociotécnico de alto risco, no qual se tensionam valores fundamentais como privacidade, segurança coletiva, proporcionalidade e autonomia comunicacional. A análise crítica desses sistemas torna-se, portanto, indispensável para avaliar se as soluções propostas preservam (ou comprometem) os princípios que sustentam a criptografia de ponta a ponta como tecnologia de proteção e de poder.

4. Redistribuição da autoridade e fragmentação da responsabilidade

A adoção generalizada da criptografia de ponta a ponta (E2EE) altera de modo estruturante quem detém poder sobre a visibilidade dos conteúdos comunicacionais. Em regimes não-encriptados, as plataformas concentram a capacidade de observação dos fluxos informacionais, o que lhes confere capacidade técnica e institucional de aplicar políticas de moderação de forma centralizada. A E2EE desloca essa capacidade para os terminais dos participantes, reduzindo o papel operativo das plataformas enquanto agentes observadores. Dessa mudança decorre um deslocamento da autoridade observacional: a quem antes podia "ver" e intervir, sobra-lhe agora responsabilidade formal, mas sem a mesma capacidade de intervenção técnica (LYON, 2018; CELESTE et al., 2023). Essa dissociação produz consequências institucionais que vão além da técnica: redesenha incentivos, responsabilidades jurídicas e canais de prestação de contas.

A fragmentação da responsabilidade é uma consequência íntima dessa redistribuição. Em contextos E2EE, a moderação deixa de ser uma única função executada por um ator com acesso privilegiado; passa a ser uma função social distribuída entre vários agentes - usuários e comunidades, plataformas, atores estatais e novos mecanismos técnicos (por exemplo, varredura no lado do cliente). Essa multiplicidade cria diferentes tipos de responsabilidade: a formal (o que a lei atribui às plataformas), a operacional (o que efetivamente se consegue fazer) e a moral (o que a sociedade espera). Frequentemente, tais dimensões não coincidem, gerando lacunas onde danos podem ocorrer sem um responsável efetivo identificável (REIDENBERG, 2009; CELESTE et al., 2023). Nesse sentido, "responsabilização" torna-se um problema de arranjo institucional, não apenas de engenharia.

A proposta de varredura no lado do cliente (client-side scanning - CSS) ilustra bem essa tentativa de reatribuição de autoridade. Em termos funcionais, a CSS busca restaurar parte da observabilidade perdida, movendo algoritmos de detecção para os dispositivos dos usuários e reportando apenas eventos selecionados à infraestrutura central (ABELSON et al., 2021). Entretanto, a adoção da CSS implica transformar o dispositivo pessoal em uma extensão das capacidades de fiscalização das instituições, com efeitos ambíguos: por um lado, promete atender exigências de investigação e proteção; por outro, amplia a superfície de ataque, cria vetores de abuso e altera radicalmente o espaço privado dos indivíduos (EFF, 2021; OHCHR, 2022). A CSS, portanto, é menos uma solução técnica isolada do que uma medida institucional - uma tentativa de deslocar de volta à esfera pública (ou estatal) o poder de observação, agora apoiada em dispositivos particulares.

A literatura empírica e analítica destaca riscos políticos e sociais associados a essa reconfiguração. Relatórios de organizações de direitos e relatórios técnicos sublinham que mecanismos instalados em larga escala tendem a sofrer expansões de escopo e a serem reutilizados para fins distintos dos originalmente declarados (HUMAN RIGHTS WATCH, 2021; EFF, 2021). Do ponto de vista das ciências sociais, autores que estudam governança tecnológica também mostram que soluções técnicas incorporadas por instituições reproduzem, frequentemente, relações de poder pré-existentes, em vez de resolvê-las (CELESTE et al., 2023; ZUBOFF, 2019). Assim, a tentativa de "restaurar" autoridade observacional por meios técnicos tem efeitos institucionais paradoxais: ao mesmo tempo em que responde a demandas por controle, pode corroer normas de privacidade e confiança que legitimam as próprias instituições.

Exemplos práticos ajudam a concretizar essa dinâmica. Quando a Apple apresentou, em 2021, um esquema de detecção local para ima-

gens de abuso infantil, a reação de organizações civis e instâncias internacionais evidenciou preocupações distintas, ainda que convergentes, quanto aos riscos da proposta. A Electronic Frontier Foundation enfatizou a dificuldade de limitar tecnicamente a detecção a categorias estritas de conteúdo, bem como os riscos de expansão progressiva de escopo (mission creep) e as vulnerabilidades introduzidas pela análise de dados diretamente nos dispositivos dos usuários, que ampliam a superfície de ataque e criam vetores de abuso (EFF, 2021). Relatores Especiais das Nações Unidas, por sua vez, destacaram o risco de instrumentalização dessas tecnologias por regimes autoritários, alertando para os impactos estruturais sobre direitos fundamentais, em especial a privacidade e a liberdade de expressão, quando capacidades de vigilância são incorporadas à infraestrutura cotidiana de comunicação (UNITED NATIONS, 2018; OHCHR, 2022).

Em outros contextos, iniciativas menos centralizadas (por exemplo, práticas comunitárias de denúncia e moderação em grupos fechados) mostram que a redistribuição de autoridade pode gerar modelos de responsabilização mais próximos das comunidades, embora com capacidade limitada para enfrentar criminalidades organizadas ou abusos em grande escala (BARTUSEK et al., 2023; SCHEFFLER; MAYER, 2023).

Do ponto de vista da governança das infraestruturas comunicacionais, a controvérsia em torno da moderação em sistemas E2EE (e, em particular, das propostas de rastreabilidade e varredura pelo lado do cliente) não pode ser adequadamente reduzida a uma escolha binária entre "privacidade" e "segurança". O que está em disputa é como e onde capacidades de observação e intervenção são reintroduzidas em comunicações privadas, e quais efeitos institucionais decorrem desse deslocamento. Como argumenta Udbhav Tiwari, a varredura pelo lado do cliente representa uma mudança qualitativa no regime de vigilância, ao transformar dispositivos pessoais em pontos permanentes de inspeção

e incorporar práticas de monitoramento diretamente à infraestrutura técnica da comunicação (TIWARI, 2025). Essa incorporação altera a relação entre usuários, plataformas e autoridades, ao normalizar formas de observação preventiva que antes dependiam de intervenções pontuais e contextuais. Nesse cenário, mecanismos como a CSS deixam de ser apenas soluções técnicas e passam a operar como instrumentos institucionais de controle, com potencial de expansão progressiva de escopo e impactos significativos sobre a privacidade, a confiança e a liberdade de expressão. Assim, a moderação em ambientes E2EE não se apresenta como um problema a ser resolvido por ajustes incrementais de engenharia, mas como uma questão estrutural sobre os limites e as consequências das formas contemporâneas de governança digital.

Em síntese, a redistribuição da autoridade promovida pela E2EE e a consequente fragmentação da responsabilidade exigem repensar modelos de governança digital. O desafio consiste em articular arranjos institucionais que preservem os benefícios da criptografia (privacidade, autonomia) ao mesmo tempo em que viabilizem mecanismos legítimos e proporcionais de proteção coletiva. Chamar essa tarefa de mera "solução técnica" é insuficiente; trata-se de reformular quem decide, sob quais critérios e mediante que processos de legitimidade e responsabilização. É nesse contexto que propostas de reintrodução de observabilidade, como a varredura no lado do cliente, passam a ocupar lugar central no debate institucional.

5. A varredura no lado do cliente como resposta institucional aos limites da E2EE

A varredura no lado do cliente surge, como desdobramento direto desse rearranjo institucional, como uma resposta aos limites impostos pela criptografia de ponta a ponta à observabilidade dos conteúdos. Di-

ferentemente de propostas anteriores de enfraquecimento da criptografia, a CSS preserva formalmente a E2EE em trânsito, mas reintroduz a capacidade de inspeção no nível do dispositivo, antes ou depois da criptografia. Essa estratégia busca conciliar, ao menos retoricamente, a proteção da privacidade com demandas de investigação criminal e moderação de conteúdo (ABELSON et al., 2021).

Do ponto de vista sociotécnico, a CSS representa uma reconfiguração profunda da arquitetura de controle: o dispositivo pessoal deixa de ser apenas um terminal privado e passa a funcionar como ponto ativo de fiscalização. Como demonstrado no Capítulo 4, esse deslocamento altera o regime de vigilância, antecipando práticas de detecção e tornando-as permanentes, distribuídas e menos visíveis ao escrutínio público. A literatura técnica aponta que tais sistemas dependem invariavelmente de hashing perceptual ou aprendizado de máquina, métodos que apresentam taxas inevitáveis de erro e são suscetíveis a abuso, evasão e expansão de escopo (ABELSON et al., 2021; TIWARI, 2025).

Uma forma de concretizar a discussão sobre client-side scanning e suas variantes é observar implementações atuais de mecanismos de varredura no lado do cliente que realizam análise de conteúdo diretamente no dispositivo, antes ou sem o envio de dados a servidores centralizados. Embora essas iniciativas não configurem um modelo abrangente de CSS baseado em listas externas de conteúdo ilegal, elas demonstram a viabilidade técnica de processamento local voltado à mitigação de riscos específicos.

No ecossistema da Apple, os recursos agrupados sob a denominação Communication Safety utilizam aprendizado de máquina no próprio dispositivo para analisar imagens e vídeos recebidos ou prestes a ser enviados por menores, com o objetivo de identificar possível nudez e apresentar avisos ou desfocar o conteúdo antes de sua visualização. De acordo com a documentação oficial disponibilizada pela empresa, toda a análise ocorre localmente, sem que as imagens sejam transmitidas à Apple ou a terceiros, preservando a confidencialidade das comunicações ao mesmo tempo em que introduz um mecanismo de intervenção preventiva (APPLE, [s.d.]).

De modo semelhante, o aplicativo Google Messages incorporou a funcionalidade de Sensitive Content Warnings, que emprega processamento on-device para detectar e desfocar imagens potencialmente explícitas em conversas protegidas por criptografia de ponta a ponta. Conforme descrito na documentação técnica oficial, a classificação das imagens ocorre inteiramente no dispositivo do usuário, sem envio de dados aos servidores do Google, caracterizando uma abordagem de análise local restrita e orientada à proteção de usuários vulneráveis (GOOGLE, [s.d.]).

Esses exemplos evidenciam que mecanismos de detecção local já estão em uso em larga escala, ainda que limitados em escopo e finalidade. Ao mesmo tempo, eles ilustram a distinção fundamental entre tais implementações restritas e propostas mais amplas de CSS voltadas à rastreabilidade generalizada de conteúdos, reforçando as preocupações institucionais e normativas discutidas na literatura crítica sobre a expansão dessas tecnologias.

Além das fragilidades técnicas, a CSS levanta problemas institucionais centrais. Ao operar de forma automatizada e preventiva, ela desloca decisões normativas (sobre o que deve ser detectado, reportado ou punido) para a infraestrutura técnica, reduzindo a transparência e a contestabilidade dessas decisões. Relatórios da Electronic Frontier Foundation e de Relatores Especiais da ONU alertam que, uma vez implementada, a CSS cria incentivos fortes para ampliação de seu escopo, inclusive para fins alheios à proteção inicialmente declarada, como controle político ou censura indireta (EFF, 2021; OHCHR, 2022).

Nesse sentido, a CSS não resolve o dilema da moderação em E2EE, mas o desloca: em vez de uma ausência de observabilidade, institui-se

uma observabilidade difusa e estrutural, com riscos ampliados à privacidade e à confiança. O problema deixa de ser "como moderar" e passa a ser "quem controla a infraestrutura de controle", sob quais critérios e com quais salvaguardas democráticas.

6. Governança digital, confiança e os limites das soluções técnicas

A análise dos capítulos anteriores permite afirmar que a moderação de conteúdo em sistemas protegidos por criptografia de ponta a ponta não constitui um problema técnico a ser simplesmente "resolvido", mas um desafio estrutural de governança digital. A E2EE redistribui confiança e autoridade, deslocando parte significativa do poder de controle das plataformas e do Estado para os próprios usuários e para a arquitetura técnica dos sistemas. Esse deslocamento gera tensões reais (especialmente no campo da responsabilização e da segurança pública), mas também expressa uma escolha política legítima em favor da autonomia comunicacional e da limitação da vigilância centralizada.

Nesse contexto, a noção de confiança algorítmica assume papel central. Em sistemas E2EE, a confiança deixa de se apoiar primariamente em instituições humanas ou organizacionais e passa a ser mediada por protocolos criptográficos e implementações técnicas. No entanto, essa confiança não é homogênea nem automática: ela depende fortemente de como os sistemas são implementados e quem controla sua infraestrutura. A distinção entre softwares proprietários e Software Livre torna-se, assim, politicamente relevante. Em sistemas de código aberto, a possibilidade de auditoria independente, verificação comunitária e contestação pública do funcionamento dos algoritmos reduz assimetrias informacionais e amplia a capacidade de escrutínio democrático. Embora o Software Livre não elimine conflitos nem garanta automaticamente práticas justas de moderação, ele altera o regime de confiança ao tornar visíveis decisões técnicas que, em sistemas proprietários, permanecem opacas e concentradas nas mãos de poucos atores.

Um exemplo ilustrativo dessa dinâmica pode ser observado em implementações amplamente utilizadas como o Signal Protocol, cujo código-fonte e especificações são publicamente auditáveis. A abertura do código permite que pesquisadores independentes verifiquem a integridade das implementações criptográficas, identifiquem vulnerabilidades e questionem decisões técnicas, reduzindo a dependência de confiança exclusiva em provedores privados. Essa característica não elimina os dilemas da moderação de conteúdo, mas altera significativamente o regime de confiança, ao tornar a infraestrutura técnica passível de escrutínio público e contestação informada.

A importância do Software Livre para ferramentas de E2EE não reside apenas em considerações técnicas de segurança, mas em sua dimensão institucional. Protocolos e implementações abertas permitem que a confiança seja distribuída entre comunidades técnicas, pesquisadores, usuários e organizações da sociedade civil, em vez de depender exclusivamente de promessas corporativas ou garantias estatais. Esse modelo não resolve o dilema da moderação, mas desloca o debate para um plano mais transparente, no qual decisões sobre arquitetura, coleta de dados e eventuais mecanismos de controle podem ser discutidas, criticadas e, em certos casos, modificadas coletivamente. Nesse sentido, o Software Livre reforça a tese central deste artigo: os dilemas entre privacidade, segurança e responsabilidade decorrem de escolhas arquiteturais e institucionais, e não de limitações matemáticas da criptografia.

A comparação com plataformas federadas, como aquelas que compõem o fediverso (por exemplo, Mastodon), ajuda a evidenciar ainda mais o caráter institucional da moderação de conteúdo. Diferentemente de sistemas centralizados ou estritamente baseados em E2EE, plata-

formas federadas adotam um modelo de governança distribuída, no qual múltiplas instâncias autônomas definem suas próprias regras de moderação, políticas de convivência e critérios de exclusão ou interconexão. A prática da defederação (defederation) (isto é, a decisão de interromper relações com instâncias que violem normas locais) ilustra um mecanismo de moderação que é explicitamente político e coletivo, não algorítmico nem centralizado.

No caso do fediverso, essa reconfiguração institucional torna-se ainda mais evidente. Plataformas como Mastodon operam por meio de instâncias autônomas que definem suas próprias regras de moderação e podem decidir, por exemplo, bloquear ou "defederar" outras instâncias consideradas incompatíveis com seus padrões normativos. Esse mecanismo permite que comunidades estabeleçam formas locais de governança e controle, sem depender de uma autoridade central. Ao mesmo tempo, essa estrutura implica fragmentação e potencial inconsistência entre regras, evidenciando que a moderação em ambientes distribuídos não desaparece, mas assume formas explicitamente políticas e situadas.

Embora o fediverso não utilize criptografia de ponta a ponta como princípio estruturante, ele demonstra que a moderação de conteúdo pode ser organizada fora do paradigma da vigilância central ou da automação em larga escala. A governança federada torna visíveis os conflitos normativos inerentes à moderação, obrigando comunidades a deliberar sobre valores, riscos e limites de tolerância. Em contraste com propostas como a varredura no lado do cliente, que deslocam decisões normativas para a infraestrutura técnica, a moderação em ambientes federados mantém essas decisões no plano institucional e social, ainda que ao custo de fragmentação, inconsistência e conflitos entre comunidades.

A análise comparativa entre E2EE, Software Livre e plataformas federadas permite compreender que não existe um modelo único ou tecnicamente ótimo de moderação de conteúdo. Cada arranjo redistribui autoridade, visibilidade e responsabilidade de maneira distinta, produzindo diferentes equilíbrios entre autonomia, segurança e governança. O erro recorrente nos debates contemporâneos consiste em tratar essas escolhas como problemas de engenharia, quando são, na realidade, decisões políticas sobre quem pode ver, decidir e intervir nas comunicações digitais.

Dessa forma, reconhecer os limites das soluções técnicas (sejam elas a criptografia forte, a automação algorítmica ou a varredura preventiva) é condição necessária para uma governança democrática das infraestruturas comunicacionais. A contribuição do Software Livre e dos modelos federados não está em oferecer respostas definitivas à moderação de conteúdo, mas em ampliar o espaço de experimentação institucional, tornando explícitas as escolhas e os valores embutidos nas arquiteturas técnicas. É nesse horizonte que a moderação em ambientes digitais deve ser compreendida: não como um problema a ser eliminado, mas como um campo permanente de negociação entre tecnologia, poder e confiança.

7. Conclusão

Este artigo analisou a moderação de conteúdo em sistemas de comunicação protegidos por criptografia de ponta a ponta a partir de uma perspectiva sociotécnica, compreendendo a E2EE não apenas como um conjunto de técnicas criptográficas, mas como uma tecnologia política que redistribui autoridade, confiança e responsabilidade nas infraestruturas comunicacionais contemporâneas. Argumentou-se que a adoção ampla da E2EE não elimina a moderação de conteúdo, mas a transforma profundamente, deslocando-a de um modelo centralizado e preventivo para um arranjo fragmentado, reativo e distribuído entre múltiplos atores humanos e técnicos.

A análise demonstrou que as tensões associadas à moderação em ambientes E2EE não decorrem de falhas matemáticas ou limitações técnicas da criptografia, mas de escolhas arquiteturais e institucionais que expressam valores sociais e políticos. Nesse contexto, propostas como a varredura no lado do cliente surgem como tentativas de reintroduzir observabilidade e controle em comunicações privadas, mas o fazem à custa de uma reconfiguração profunda do regime de vigilância, deslocando práticas tradicionalmente excepcionais e juridicamente delimitadas para o nível permanente da infraestrutura técnica. Tal deslocamento acarreta riscos significativos à privacidade, à confiabilidade epistêmica e à legitimidade democrática das práticas de investigação e moderação.

Ao examinar a redistribuição da autoridade observacional promovida pela E2EE, o artigo evidenciou a fragmentação da responsabilidade como um fenômeno central da governança digital contemporânea. Plataformas permanecem formalmente responsáveis por conteúdos que não podem acessar; Estados enfrentam limites técnicos legítimos à vigilância; usuários e comunidades passam a ocupar posições mais relevantes na regulação cotidiana das interações. Esse cenário não configura um vazio normativo, mas um campo de disputa institucional no qual se renegociam continuamente os limites entre autonomia comunicacional, proteção coletiva e responsabilização.

A discussão sobre Software Livre e plataformas federadas reforçou a tese de que os dilemas da moderação não podem ser resolvidos exclusivamente por soluções técnicas centralizadas. Implementações abertas e auditáveis de E2EE alteram o regime de confiança algorítmica ao reduzir assimetrias informacionais e ampliar a contestabilidade pública das decisões técnicas. De modo semelhante, modelos federados de governança, como os observados no fediverso, demonstram que a moderação pode ser organizada como prática explicitamente política e comunitária, ainda que marcada por fragmentação e conflitos normativos. Esses arranjos não oferecem respostas definitivas, mas tornam visíveis as escolhas e os valores embutidos nas arquiteturas comunicacionais.

Dessa forma, a principal contribuição deste trabalho consiste em deslocar o debate sobre moderação de conteúdo em sistemas E2EE do plano estritamente técnico para o plano institucional e normativo, onde ele de fato se situa. Reconhecer os limites das soluções algorítmicas e da vigilância automatizada não implica abdicar da proteção de direitos ou da segurança pública, mas assumir que tais objetivos só podem ser perseguidos de maneira legítima por meio de arranjos democráticos, transparentes e juridicamente controlados. A moderação de conteúdo em ambientes criptografados deve, portanto, ser compreendida como um experimento sociotécnico permanente, no qual tecnologia, poder e confiança permanecem em negociação contínua.

Em última instância, insistir na restauração de modelos de controle total em infraestruturas comunicacionais protegidas por criptografia forte representa um erro de categoria: trata-se de exigir da engenharia aquilo que apenas a política, o direito e a governança democrática podem oferecer. A tarefa que se impõe não é "corrigir" a criptografia, mas deliberar coletivamente sobre os limites aceitáveis de observação, intervenção e responsabilização em sociedades digitais pluralistas.

Referências

ABELSON, Hal et al. Keys under doormats: mandating insecurity by requiring government access to all data and communications. **Journal of Cybersecurity**, v. 1, n. 1, p. 1–17, 2015.

ABELSON, Hal et al. Bugs in our pockets: the risks of client-side scanning. **Cybersecurity**, v. 10, n. 1, 2021. Disponível em: https://arxiv.org/abs/2110.07450.

Acesso em: 20 jan. 2026.

APPLE. Child Safety. Disponível em: https://www.apple.com/child-safety/. Acesso em: 15 jan. 2026.

APPLE. Communication Safety parental controls and on-device sensitive content analysis features. Disponível em: https://support.apple.com/en-us/105069. Acesso em: 15 jan. 2026.

ASSANGE, Julian et al. Cypherpunks: freedom and the future of the internet. New York: **OR Books**, 2012.

BARTUSEK, Jan; GARG, S.; JAIN, A.; POLICHARLA, G.-V. End-to-end secure messaging with traceability only for illegal content. In: [Obra coletiva]. Cham: **Springer Nature Switzerland**, 2023. p. 35–66. DOI: https://doi.org/10.1007/978-3-031-30589-4_2.

CELESTE, Edoardo et al. The content governance dilemma: digital constitutionalism, social media and the search for a global standard. Cham: **Palgrave Macmillan**, 2023.

DIAS OLIVA, Thiago. Content moderation technologies: applying human rights standards to protect freedom of expression. **Human Rights Law Review**, v. 20, n. 4, p. 607–640, 2020. DOI: https://doi.org/10.1093/hrlr/ngaa032.

EFF. In 2021, we told Apple: Don't Scan Our Phones. **Deeplinks** (blog). By Joe Mullin. 28 dez. 2021. Disponível em: https://www.eff.org/am/deeplinks/2021/12/2021-we-told-apple-dont-scan-our-phones. Acesso em: 15 jan. 2026.

GOOGLE. Understand sensitive content warnings in Google Messages. **Google Messages Help Center**. Disponível em: https://support.google.com/messages/answer/15724426. Acesso em: 15 jan. 2026.

GREEN, Matthew; SMITH, Matthew. The cryptopals crypto challenges: teaching cryptography through programming. **Cryptology ePrint Archive**, 2016. Disponível em: https://eprint.iacr.org/2016/1006. Acesso em: 20 jan. 2026.

HAZRA, R.; CHATTERJEE, P.; SINGH, Y.; PODDER, G.; DAS, T. Data encryption and secure communication protocols. In: [Obra coletiva]. Hershey: **IGI Global**, 2024. p. 546–570. DOI: https://doi.org/10.4018/979-8-3693-6557-1.ch022.

HUMAN RIGHTS WATCH. World report 2021: events of 2020. New York: **Human Rights Watch**, 2021. Disponível em: https://www.hrw.org/world-report/2021. Acesso em: 12 jan. 2026.

KAHN, David. The codebreakers: the comprehensive history of secret communication from ancient times to the internet. New York: **Scribner**, 1996.

LYON, David. The culture of surveillance: watching as a way of life. Cambridge: **Polity Press**, 2018.

NEEDHAM, Joseph. Science and Civilisation in China: Volume 3: Mathematics and the Sciences of the Heavens and the Earth. Cambridge: **Cambridge University Press**, 1959.

OFFICE OF THE HIGH COMMISSIONER FOR HUMAN RIGHTS (OHCHR). Spyware and surveillance: Threats to privacy and human rights growing, UN report warns. **Press release**, 16 set. 2022. Disponível em: https://www.ohchr.org/en/press-releases/2022/09/spyware-and-surveillance-threats-privacy-and-

-human-rights-growing-un-report. Acesso em: 10 jan. 2026.

REDDY, K. K.; CHADHA, A. R.; NIKHIL, P. S.; SOWNTHARRAJAN, S. Hybrid cryptography techniques for data security in cloud computing. In: Proceedings of the IC2PCT 2024. [S.l.]: **IEEE**, 2024. p. 1836–1842. DOI: https://doi.org/10.1109/IC2PCT60090.2024.10486794.

REIDENBERG, Joel R. Risk and regulation in information security. **Journal of Information Technology & Politics**, v. 6, n. 3–4, p. 1–16, 2009.

SCHEFFLER, Samuel; MAYER, Jonathan. SoK: content moderation for end-to-end encryption. **Proceedings on Privacy Enhancing Technologies**, v. 2023, n. 2, p. 403–429, 2023. DOI: https://doi.org/10.56553/popets-2023-0060.

SOOD, R.; KAUR, H. A literature review on RSA, DES and AES encryption algorithms. In: Proceedings of the Soft Computing Research Society. [S.l.]: **Soft Computing Research Society**, 2023. p. 57–63. DOI: https://doi.org/10.56155/978-81-955020-3-5-07.

SUETÔNIO. As vidas dos doze Césares: Júlio César, Augusto, Tibério, Calígula, Cláudio, Nero, Galba, Óton, Vitélio, Vespasiano, Tito, Domiciano. Brasília: Senado Federal, **Conselho Editorial**, 2012.

TIWARI, Udbhav. Da detecção à censura: os perigos da implementação da varredura pelo lado do cliente em comunicações privadas. In: BRITO CRUZ, Francisco; SIMÃO, Bárbara; HOUANG, André (org.). Direitos fundamentais e processo penal na era digital: doutrina e prática em debate. v. 7. São Paulo: **InternetLab**, 2025.

UNITED NATIONS. Human Rights Council. Report of the Special Rapporteur on the right to privacy: note by the Secretariat (A/HRC/37/62). Geneva: **United Nations**, 25 out. 2018. Disponível em: https://www.ohchr.org/en/documents/thematic-reports/ahrc3762-report-special-rapporteur-right-privacy. Acesso em: 11 jan. 2026.

ZUBOFF, Shoshana. The age of surveillance capitalism: the fight for a human future at the new frontier of power. New York: **PublicAffairs**, 2019.