Capítulo 3

Introdução à Engenharia Social: da Psicologia Cognitiva aos Ataques Automatizados

Jéferson Campos Nobre (UFRGS), Pamela Carvalho da Silva (UFCSPA), Antônio João Gonçalves de Azambuja (UFRGS), Maurício Ariza (UFRGS), Lisandro Zambenedetti Granville (UFRGS) e Caroline Tozzi Reppold (UFCSPA)

Resumo

A Engenharia Social (ES) é uma disciplina que visa explorar a natureza humana e suas vulnerabilidades psicológicas para obter informações e acessos não autorizados a sistemas, ou então persuadir indivíduos a realizar ações indesejadas. Com base em princípios da Psicologia, a Engenharia Social utiliza uma variedade de estratégias de manipulação com a intenção de explorar aspectos humanos relacionados ao processo de tomada de decisão, bem como às vulnerabilidades nas interações humanas e caraterísticas culturais. Este minicurso tem como objetivo fornecer uma visão abrangente sobre a interseção entre a ES, a Psicologia e a Automatização Computacional, abordando as técnicas de manipulação empregadas. Serão abordados os aspectos psicológicos relacionados aos principais ataques de ES, bem como será observada a crescente preocupação com a Engenharia Social Automatizada . No decorrer do minicurso serão descritos os principais vieses cognitivos explorados nos ataques de Engenharia Social. Em seguida, serão abordados os fundamentos e técnicas aplicados no contexto da Engenharia Social Automatizada, considerando o avanço da Inteligência Artificial e do Aprendizado de Máquina e potenciais ataques em larga escala que utilizam ferramentas de automação. Ao final, espera-se que os participantes compreendam as técnicas de manipulação utilizadas na Engenharia Social, os vieses cognitivos associados e os desafios apresentados pela Engenharia Social Automatizada.

Abstract

Social Engineering (SE) is a discipline that aims to explore human nature and its psychological vulnerabilities to obtain unauthorized information and access to systems,

or therefore persuade individuals to perform unwanted actions. Based on principles of Psychology, ES uses a variety of manipulative strategies with the intention of exploring human aspects related to the decision-making process, such as vulnerabilities in human interactions and cultural characteristics. This mini-course has the objective of providing a comprehensive vision of the intersection between Social Engineering, Psychology and Computational Automation, addressing the manipulation techniques used. The psychological aspects related to the main SE attacks will be addressed, as will be observed the growing concern with Automated Social Engineering (ASE). In the course of the mini-course, the main cognitive processes explored in SE attacks will be described. Next, the fundamentals and techniques applied in the Automated Social Engineering context will be addressed, considering the advancement of Artificial Intelligence and Machine Learning and potential large-scale attacks that use automation tools. In the end, it is hoped that the participants understand the manipulation techniques used in Social Engineering, see the associated cognitive issues and the challenges presented by ASE.

3.1. Introdução

Ataques cibernéticos exploram as vulnerabilidades das estruturas de Tecnologia da Informação e Comunicação. As organizações têm empregado diversas soluções para enfrentar os ataques cibernéticos, tais como *Firewalls*, Sistema de Detecção de Intrusão (*Intrusion Detection System* - IDS), Antivírus, entre outros. No entanto, esses mecanismos de defesa muitas vezes não são suficientes para impedir as ações relacionadas com aspectos humanos no ambiente cibernético. Os ataques têm explorado a interação humana em conjunto com as brechas tecnológicas, enfraquecendo a cadeia de segurança [Salahdine and Kaabouch 2019] [Klimburg-Witjes and Wentland 2021]. As relações de confiança entre humanos no ambiente cibernético têm proporcionado um cenário para a prática de atos ilícitos. Dessa forma, alguns autores apontam que o fator humano é o elo mais fraco na cadeia de Segurança Cibernética [Mitnick and Simon 2003].

A Engenharia Social (ES) é uma disciplina que visa explorar a natureza humana e suas vulnerabilidades para auxiliar em ações como obter informações e acessos não autorizados a sistemas e persuadir indivíduos a realizarem ações indesejadas. Com base em princípios da Psicologia, a ES utiliza uma variedade de estratégias com a intenção de explorar aspectos humanos relacionados ao processo de tomada de decisão, bem como vulnerabilidades nas interações humanas e caraterísticas culturais. Os atacantes que utilizam a ES vêm diversificando os mecanismos para explorar as relações de confiança, tendo como objetivo ampliar o acesso a dados relevantes, assim como potenciais alvos. Neste contexto, a interconectividade das redes sociais e o crescimento da dimensão cognitiva do trabalho estão tornando os recursos humanos como um dos pilares da segurança [Culot et al. 2019] [Greitzer et al. 2019].

Uma das abordagens que pode ser utilizada para compreender aspectos da Es é o viés cognitivo. Este é um conceito oriundo da Psicologia Cognitiva e da Economia Comportamental que estuda as falhas comuns no pensamento humano, resultando em pensamentos distorcidos, imprecisos e incompletos, o que pode levar à tomada de decisões precipitadas e fracas. O viés cognitivo pode ser utilizado e explorado por um atacante para que o alvo do ataque seja induzido a algum erro, como a distorção de determinados julgamentos que o levem a uma tomada de decisão ruim (i.e., permitir um acesso inde-

vido). Informações sobre características psicológicas podem auxiliar em ataques que se utilizam de viés cognitivo. Tais informações podem ser coletadas, por exemplo, em redes sociais.

O crescente aumento do uso das redes sociais para estabelecer relacionamentos pessoais e profissionais abre possibilidades para as ações de ES [Shires 2018] [Klimburg-Witjes and Wentland 2021]. Assim, ao oferecerem serviços, as redes sociais coletam dados pessoais e corporativos formando bases de dados de alto valor, as quais podem ser utilizadas como ferramentas para ataques cibernéticos [Crossler and Bélanger 2014]. Muitas vezes essas bases podem ser empregadas em conjunto com mecanismos automatizados. Os ataques automatizados requerem pouca intervenção humana e podem simular o comportamento humano [Huber et al. 2009] [Shafahi et al. 2016].

Diversas tecnologias podem ser utilizadas para uma Engenharia Social Automatizada (ESA) [Huber et al. 2009]. Um exemplo de tais tecnologias são *bots*, softwares automatizados que são capazes de executar comandos de operação e controle sem a necessidade de participação humana. *Bots* podem ser utilizados para ações positivas, como por exemplo, ajudar o usuário a ter uma melhor Qualidade de Experiência (QoE). Contudo, *bots* têm sido utilizados como ferramenta para ataques de ESA. Como são escaláveis, essa ferramenta permite que um único atacante contate um grande número de potenciais vítimas simultaneamente, por exemplo, na busca de informações confidenciais [Huber et al. 2009][Dewangan and Kaushal 2016].

A ESA tem evoluindo, considerando o avanço da Inteligência Artificial e do Aprendizado de Máquina. Dessa forma, os humanos tendem a confiar nos *bots* [Dickerson et al. 2014], já que os mesmo tem a capacidade de se passar por seres humanos, imitando as atividades dos usuários reais [Shafahi et al. 2016]. Informações oriundas de redes sociais podem ser associadas a conceitos da Psicologias Cognitiva para para tornar os ataques que usam *bots* cada vez mais efetivos.

Na literatura, poucos trabalhos apresentam análises sobre a ESA com o uso de *Bots*. A maioria dos trabalhos estuda a área da Psicologia Social, com foco no comportamento humano diante das ações de ES [Huber et al. 2009]. Os autores [Al-Charchafchi et al. 2019] e [Piovesan et al. 2019] abordam as ameaças à segurança nas redes sociais, decorrentes dos ataques de ES utilizando contas falsas, roubo de identidade e *phishing*. No sentido de influenciar os usuários nas redes sociais, há trabalhos que avaliam as vulnerabilidades das redes sociais com o uso de *bots* para campanhas de convencimento nas redes. Os autores [Freitas et al. 2014] e [Messias et al. 2018] analisam o uso de *bots* no *Twitter* para influenciar os usuários e comprometer a estrutura da rede. Já [Huber et al. 2009] propõem a automação das tarefas de ES por meio de um *bot* no *Facebook*, concluindo que a persuasão é um recurso essencial no processo de ESA.

O presente capítulo apresentar os principais fundamentos de ES, assim como uma visão abrangente sobre sua interseção com a Psicologia e a Automatização Computacional. Além disso, fundamentos e técnicas aplicados no contexto da ESA são discutidos com mais profundidade, assim como questões relacionadas à utilização de redes sociais neste contexto. O objetivo é auxiliar na compreensão das técnicas utilizadas na ES, auxiliando na formação de profissionais com atuação na Segurança da Informação.

O presente capítulo está organizado da seguinte forma. Na Seção 3.2, serão apresentadas definições e conceitos básicos de Engenharia Social. Na Seção 3.3, serão apresentados os principais pontos de interseção entre a Psicologia, a Segurança da Informação e a Engenharia Social. Na Seção 3.4, será apresentada uma contextualização sobre a Engenharia Social Automatizada, bem como uma apresentação das ferramentas e técnicas utilizadas. Na Seção 3.5, serão apresentadas ações para prevenir e mitigar os ataques de Engenharia Social. Na Seção 3.6, serão apresentados estudos de casos relacionados com os temas descritos no capítulo. Na Seção 3.7, serão apresentadas reflexões sobre questões éticas implicadas na prática da Engenharia Social em contextos não maliciosos, como na pesquisa e na execução de testes de invasão. Finalmente, considerações finais e perspectivas de trabalhos futuros são apresentados na Seção 3.8.

3.2. Fundamentos de Engenharia Social

A interconectividade proporcionada pela Internet tem ampliado consideravelmente a superfície de ataques cibernéticos. A crescente dependência da tecnologia da informação (TI) por parte das empresas e organizações, juntamente com a expansão das redes e sistema conectados, resulta em uma maior exposição e ameaças digitais. Essa evolução tecnológica, no que pese tenha gerado benefícios, também abriu novas brechas de segurança que os cibercriminosos têm explorado de forma crescente [Benias and Markopoulos 2017].

Como consequência, a necessidade de uma abordagem abrangente e eficaz em relação à cibersegurança passou a ser uma prioridade para proteger a integridade dos dados, sistemas e infraestruturas das organizações em uma sociedade cada vez mais conectada digitalmente. A ES, aliada a esse cenário, é uma das estratégias utilizadas pelos atacantes para acessar informações confidenciais e comprometer a segurança das organizações [Reep-van den Bergh and Junger 2018]

3.2.1. Definição e Conceitos Básicos

A ES é definida como a arte de explorar as pessoas com o objetivo de obter acesso aos dados e informações de potenciais alvos dos sistemas de informação, independente de usar ou não a tecnologia. É uma técnica de ataque que explora o comportamento humano, por meio da persuasão e manipulação psicológica das pessoas. Na prática, o fator humano é o elo mais fraco da cadeia de cibersegurança. [Mitnick and Simon 2003] [Klimburg-Witjes and Wentland 2021].

A disponibilidade de diversos meios de comunicação de grande alcance gera um cenário propício para os ataques de ES. O avanço da tecnologia tem facilitado a automação e escalabilidade desses ataques, permitindo que os atacantes atingir um número maior de potenciais vítimas em um curto período de tempo [Pinheiro 2020].Compreender e proteger contra os ataques de ES passa, portanto, ser um requisito essencial para garantir a segurança dos dados e sistemas em ambientes conectados e dependentes da tecnologia.

O processo para proteger os dados e informações, visando assegurar a confidencialidade, integridade e disponibilidade está relacionado com a Segurança da Informação (SI). A relação entre ES e SI é reforçada pelo desenvolvimento contínuo da tecnologia tem permitido a automação e escalabilidade dos ataques de ES, tornando-os ainda mais desafiadores de combater [Beal 2005].

Considerando o contexto da ES no campo da SI emergem fatores que impactam o comportamento humano e podem tornar as pessoas mais suscetíveis a manipulação e persuasão por parte dos atacantes. Esses fatores abordam:

1. Conhecimento e conscientização: a falta de conhecimento e conscientização sobre as ameaças de ES e as táticas utilizadas pelos atacantes pode tornar as pessoas menos preparadas para identificar e evitar esses ataques.

2. Confiança: as pessoas tendem a confiar em outras pessoas, principalmente quando essas pessoas parecem ser amigáveis, prestativas ou apresentam uma autoridade aparente. Os atacantes aproveitam essa tendência para criar subterfúgios convincentes e ganhar a confiança de suas vítimas;

3. Curiosidade: a curiosidade humana pode ser explorada por meio de táticas como títulos sensacionalistas ou informações intrigantes, fazendo as pessoas a clicarem em *links* maliciosos ou a interagirem com conteúdos duvidosos, que podem representar riscos à SI;

4. Caos informacional: em um ambiente digital repleto de informações, que cria o caos informacional, as pessoas podem receber um volume grande de informações dificultando a identificação de tentativas de ataques;

5. Conexões de relacionamento: a criação de conexões pessoais ou profissionais pode ser explorada pelos atacantes para obter informações importantes. Eles podem pesquisar sobre as vítimas nas redes sociais para estabelecer conexões usando argumentos convincentes;

6. Emoções: reações à estímulos que, em um processo complexo, geram sentimentos utilizados para influenciar pessoas a revelarem informações ou clicarem em *links* maliciosos; e

7. Rotinas de comportamento: as pessoas podem agir de forma automática, seguindo rotinas estabelecidas sem questionar a legitimidade das ações diárias.

Conscientizar as pessoas sobre esses fatores e promover uma cultura de segurança é essencial para reduzir os riscos associados à ES e fortalecer a SI. Os programas de conscientização devem estar alinhados com as medidas técnicas e processuais relacionadas com a segurança nas organizações.

3.2.2. Importância da Engenharia Social na Segurança da Informação

A ES desempenha um papel fundamental na SI, utilizando a manipulação psicológica para obter acesso não autorizado aos dados, as informações e os sistemas computacionais. A SI é uma área do conhecimento dedicada a proteção dos ativos da informação contra acessos e mudanças não autorizados, falta de disponibilidade dos recursos digitais e quebra da autenticidade [Beal 2005]. Sendo assim, a SI considera os princípios da confidencialidade, integridade, disponibilidade e autenticidade, a saber [Shaabany and Anderl 2018]:

1. Confidencialidade: assegura que os dados, sistemas e as informações serão acessadas exclusivamente pelos usuários autorizadas;
2. Integridade: assegura que os dados e as informações não tenham sido modificadas ao longo do seu ciclo de transmissão entre os usuários com acesso autorizado;
3. Disponibilidade: assegura à capacidade dos dados, sistemas e informações estarem acessíveis e funcionando quando necessário, para os usuários autorizados; e
4. Autenticidade: assegura a origem, a identidade e integridade dos dados e informações, na busca de uma proteção para as transações eletrônicas, os sistemas e as informações em um ambiente digital.

O avanço tecnológico tem possibilitado medidas técnicas para incrementar a capacidade de proteção dos dados, informações e sistemas. No entanto, os atacantes fazem uso de ações de ES para burlar as técnicas de proteção. Essas ações tem como alvo o ser humano, considerado um elo fraco na cadeia de segurança. As vulnerabilidades dos elementos sociais e humanos para acesso aos sistemas e roubo de informações são utilizadas pelos engenheiros sociais [de Souza Pereira et al. 2022].

3.2.3. Tipos de Ataques de Engenharia Social

A Engenharia Social se caracteriza pela manipulação da vítima através de técnicas psicológicas, utilizando a tecnologia como recurso de suporte em diferentes níveis. Essa análise permite entender a ES a partir de duas perspectivas principais, que podemos nos referir como os aspectos psicológicos e os aspectos tecnológicos. Compreender as características e interconexões entre ambos é de vital importância para o entendimento real da ES e, consequentemente, como combater as ameaças da mesma.

O decreto-Lei que instituiu o Código Penal brasileiro em 1940 já incluía no Capítulo VI a tipificação criminal para golpes e fraudes, com destaque ao Artigo 171, que define o crime de Estelionato, descrito como a obtenção ilícita de vantagem sobre outro indivíduo através da indução ao erro, utilizando quaisquer meios fraudulentos para isso [Brasil 1940]. Portanto o engano e manipulação de outros indivíduos é anterior ao advento da internet e da ampliação do acesso à tecnologia.

Do ponto de vista da ES, podemos dizer que a tecnologia trouxe recursos que permitiram que o estelionatário tradicional pudesse aumentar suas chances de sucesso com menor necessidade de auto-exposição, diminuindo assim os seus riscos. Como exemplo, um dos golpes *online* mais popularmente conhecidos é o *e-mail* do Príncipe Nigeriano, cujo ápice ocorreu nos anos 80-90, se trata na verdade de uma variação de outro golpe datado do século 18. A fraude se baseia no pedido de auxílio para a realização de uma transação comercial, prometendo que a vítima será recompensada com uma porcentagem do valor envolvido. Apenas do ponto de vista desse caso, podemos observar algumas características do recurso tecnológico utilizado:

1. O uso de *e-mail* como canal de comunicação permite maior agilidade tanto na produção da mensagem quanto no recebimento da mesma por potenciais vítimas, comparado por exemplo ao envio por carta;

2. Enquanto a relação entre carta/mensagem e destinatário/vítima é de 1:1 por carta, o *e-mail* oferece uma relação de 1:n, oferecendo uma maior escalabilidade sem necessidade de esforço adicional proporcional por parte do atacante;

3. A busca por destinatários/vítimas ou mesmo o envio das mensagens em massa podem ser completamente automatizados, novamente gerando escalabilidade e menor necessidade de esforço.

A combinação das interações sociais e tecnológicas caracterizam portanto os ataques de ES. Os ataques normalmente seguem um estrutura de quatro fases, nomeadamente: i) obter informações sobre a vítima para realização da abordagem; ii) estabelecer uma relação de confiança entre o agressor e a vítima; iii) explorar a informação para o desenvolvimento de ações específicas; e iv) executar o ataque para atingir seu(s) objetivo(s) [Mitnick and Simon 2003] [Klimburg-Witjes and Wentland 2021].

Os ataques de ES objetivam comumente a obtenção de informações confidenciais, o acesso não autorizado a sistemas ou persuadir as pessoas a realizar ações indesejadas. Por conta dessas características, o ataque pode visar um objetivo final do atacante, ou servir como etapa para a realização de outros ataques, como a obtenção de informações sensíveis para embasar um ataque direcionado a um usuário privilegiado, ou o roubo de identidade da vítima para realização de transações comerciais fraudulentas.

Considerando diferentes linhas na literatura, os principais tipos de ataques de ES são os seguintes:

1. *Phishing*: é uma forma de ataque que utiliza o envio de mensagens falsas com um *link* de aparência legítima, por *e-mail*, para enganar pessoas a revelarem informações pessoais, como senhas e dados financeiros;

2. *Spear Phishing*: é uma forma mais direcionada de *phishing*, na qual os atacantes personalizam as mensagens de *e-mail* para um alvo específico, possibilitando um taxa de sucesso maior no ataque;

3. *Vishing*: é uma forma de ataque realizada pelos engenheiros sociais, utilizando o telefone para estabelecer uma relação de confiança com o usuário, na busca informações pessoais, corporativas e confidencias;

4. *Whaling*: é uma forma de ataque aos integrantes do alto escalão das organizações, utilizando um *spear phishing* para personalizar as mensagens de *e-mail*;

5. *Smishing*: é uma forma de ataque que utiliza mensagens de texto, como por exemplo SMS, para obter informações pessoais dos usuários alvo;

6. *Impersonation*: é uma forma de ataque, na qual o atacante busca estabelecer um relação de confiança para obter informações, utilizando *e-mails* e/ou mensagens.

7. *Pretexting*: os atacantes criam uma história fictícia para manipular as vítimas a compartilharem informações, geralmente por telefone.

8. *Quid Pro Quo*: os atacantes oferecem algo em troca das informações das vítimas, como suporte técnico falso em troca de senhas;

9. *Water Holing*: é um ataque que os engenheiros sociais buscam comprometer um site frequentado pelas vítimas esperadas, explorando a confiança nas fontes para disseminar *malware*;.

10. *Tailgating*: é uma técnica de ES usada em segurança física e cibernética. Essa abordagem envolve um indivíduo não autorizado aproveitando a entrada de um local seguro ou restrito ao seguir de perto um funcionário autorizado; e

11. *Baiting*: os atacantes oferecem um atrativo, como um *download* gratuito, para convencer as vítimas a realizar ações que comprometam a segurança.

3.3. Psicologia e Engenharia Social

A exploração de fatores humanos representa uma significativa parcela dos incidentes de segurança da informação. Desses incidentes, ataques de ES costumam ocupar papel de destaque nos relatórios de ameaças publicados anualmente e chegam a representar 98% de todos os crimes cibernéticos de phishing e de violação de dados [Martineau et al. 2023]. Tal representatividade pode ser explicada pela maior facilidade na execução dos ataques de ES, uma vez que não exigem o uso de ferramentas complexas ou conhecimentos técnicos prévios, bem como pelo seu potencial de ser bem sucedido.

Enquanto controles e tecnologias aplicadas à segurança da informação são aperfeiçoadas com o passar dos anos, tornando-se mais complexos e mais difíceis de serem comprometidos, aspectos psicológicos tornam-se uma estratégia vantajosa para os atacantes. Isto porque tendem a apresentar a mesma complexidade e são mais fáceis de presumir e de explorar. Descrevendo um contexto no qual o vetor de ataque assume, predominantemente, uma natureza psicológica, contrastando com a abordagem estritamente tecnológica, amplamente prevalente na área [Martineau et al. 2023].

3.3.1. Interface Psicologia e Segurança da Informação

Embora questões relacionadas aos fatores humanos em segurança da informação e segurança cibernética, como conscientização e comportamento, constituam elementos críticos mencionados em pesquisas, em diretrizes e em boas práticas [Robinson 2023] [Collier et al. 2023]. Dimensões culturais e comportamentais tem sido pouco enfatizadas nas abordagens de segurança da informação [Collier et al. 2023].

No contexto maior, segurança da informação tende a ser considerada uma disciplina essencialmente técnica, fortemente atrelada à tecnologia da informação. Essa característica é percebida também na formação dos profissionais, na qual observa-se uma lacuna acerca do estudo dos fatores humanos na área [Nobles 2023]. Promovendo, assim, uma perspectiva que negligencia o caráter estratégico e comportamental, e favorece uma abordagem focada na implementação de controles sobretudo tecnológicos e em políticas que, na maioria das vezes, não consideram elementos comportamentais, culturais e necessidades humanas [Collier et al. 2023].

A efetiva adoção e integração de controles e práticas de segurança da informação pelas pessoas é facilitada quando estes levam em conta o comportamento humano e as necessidades dos usuários. Regras e normas complexas ou insuficientemente justificadas afetam a capacidade de reflexão das pessoas e, frequentemente levam a uma predisposição a não segui-las e a buscar formas de contorna-las [Collier et al. 2023].

A segurança da informação é uma área que se beneficia das pesquisas em psicologia [Ancis 2020]. Reconhecer a influência que os fatores psicológicos exercem no âmbito da segurança da informação, bem como compreender os aspectos da cognição humana explorados nos ataques de ES [Montañez et al. 2020] favorece a elaboração e implementação de estratégias de segurança mais propensas de serem adotadas. Assim, torna-se imprescindível a inclusão ativa da disciplina psicológica, ampliando a perspectiva para além do humano como parte de um problema [Zimmermann and Renaud 2019]. Desta forma, segurança da informação deve igualmente ser vista a partir de uma disciplina comportamental [Martineau et al. 2023], contemplando a complexa interação entre seres humanos e tecnologia, e suas implicações.

Neste sentindo a segurança da informação também deve incorporar os avanços da ciberpsicologia. Disciplina que se dedica a compreender os processos psicológicos e os aspectos e características do comportamento humano na interação com a tecnologia [Attrill-Smith et al. 2019b] [Ancis 2020]. Essa disciplina surge a partir do caráter pervasivo da tecnologia na contemporaneidade. Tem como característica a inter e a transdisciplinaridade, incluí disciplinas como interação humano-computador, ciência da computação, engenharia e psicologia [Attrill-Smith et al. 2019b] [Ancis 2020]. Igualmente, sua aplicação é variada, contemplando áreas como saúde, educação, práticas em psicologia e também segurança [Ancis 2020] [Martineau et al. 2023].

Nessa integração, a visão tradicional de que há uma separação entre os ambientes virtuais e reais é desafiada. A perspectiva contemporânea nos leva a reconhecer que as fronteiras entre essas duas dimensões se tornaram fluidas, revelando a convergência entre as experiências *online* e *offline* sua intrincada relação. Tal como previsto no incio dos anos 2000 por [Castells 2002].

Desfazer a noção de que há distinção entre real e virtual, *online* e *offline*, e fundamentar-se em uma visão integrada das dimensões, permite reduzir a tendência de subestimar riscos e impactos das decisões e ações relacionadas as interações mediadas pela tecnologia. Nas próximas subseções ficará evidente que muitos dos fenômenos observados em ambientes *offline* estão presentes nos ambientes *online* ou mediados, destacando a importância da perspectiva integrada e sua relevância à segurança da informação.

3.3.2. Psicologia Aplicada na Engenharia Social

Apesar da direta relação com fatores humanos, a predominância de um enfoque tecnológico é observada também na compreensão e identificação de ataques de ES [Montañez et al. 2020]. Repetindo, assim, o tradicional modo de abordar segurança da informação e subestimando os aspectos psicológicos associados.

Beneficiando-se de vulnerabilidades humanas do processo de tomada de decisão, que influenciam comportamentos e impactam na motivação, a ES baseia-se, em grande

parte, na exploração da psicologia humana [Montañez et al. 2020]. O processo humano de tomada de decisão é complexo, envolve processos cognitivos como atenção e memória e é afetado por estados emocionais, bem como conhecimentos e experiências prévias.

Para facilitar a tomada de decisão ou a resolução de problemas, devido limitações relacionadas à informações disponíveis e à capacidade de processamento de informações, seres humanos podem recorrer as heurísticas[Korteling and Toet 2022]. Heurísticas são estratégias baseadas na experiência, que podem oferecer uma forma eficiente de encontrar uma solução, devido simplificações e desvios, mas que não garantem um resultado preciso [APA nd].

Quando levam a resultados abaixo do ideal ou incorretos, as heurísticas tornam-se vieses cognitivos, que podem ser facilmente manipulados e explorados. Os vieses cognitivos podem ser definidos como tendências e inclinações que enviesam ou distorcem o processamento de informações [Tversky and Kahneman 1974] [Korteling and Toet 2022]. São numerosos os vieses cognitivos conhecidos [Korteling and Toet 2022], a tabela 3.1 lista alguns dos principais vieses cognitivos que podem ser explorados ou influenciar o desfecho de ataques de ES.

No que tange aspectos psicológicos, para além dos vieses cognitivos, ataques de ES tem por característica estimular e explorar emoções como medo, excitação e surpresa. Estas emoções geram sentimentos que tendem a induzir respostas e ações rápidas, afetando a tomada de decisão e potencializando a probabilidade de sucesso de um ataque de ES.

Restrições de tempo, senso de urgência, intimidação, curiosidade, simpatia, aparente legitimidade, confiança, criação de conexão interpessoal e sobrecarga de informações também são empregados para aumentar a complexidade decisória, influenciar a tomada de decisão e levar aos vieses. Outros fatores cognitivos como carga de trabalho, estresse e vigilância, também se relacionam com ataques de engenharia social podem se relacionar com ataques de ES e influenciar o desfecho destes [Montañez et al. 2020]. Contextos situacionais e ambientes com altos níveis de tensão, tendem a influenciar e prejudicar a tomada de decisão das pessoas.

Abaixo são listados alguns exemplos de como os vieses podem ser utilizados por atacantes na aplicação da ES:

- Viés de ancoragem: dando ênfase inicialmente a uma informação que pode gerar senso de legitimidade, como a referência ao nome de algum individuo que ocupe cargo relevante na organização, o atacante pode gerar a impressão de que conhece e trabalha na organização, encorajando a tomada de decisão rápida;

Tabela 3.1: Principais vieses cognitivos - adaptado de [Korteling and Toet 2022], de [Wilke and Mata 2012] e de [APA nd].

Vies	Descrição
Viés de ancoragem	tendência a utilizar uma informação específica a qual foi exposto previamente na tomada de decisão.
Viés de autoridade	tendência a atribuir maior valor e confiabilidade à opinião de uma figura de autoridade (não relacionada ao seu conteúdo).
Viés de confirmação	tendência de selecionar, interpretar, focar e lembrar informações de uma forma que confirme as própria crenças, pontos de vista e/ou hipóteses, independentemente da veracidade.
Viés de conformidade	tendência de ajustar o pensamento e o comportamento de um indivíduo ao padrão de um grupo. Este viés está na base da dinâmica e do marketing dos influenciadores digitais.
Viés de crença	tendência de ser influenciado pelo conhecimento de alguém ao avaliar conclusões e aceitá-las como verdadeiras porque são críveis e não porque apresentam validade lógica.
Viés de disponibilidade	tendência de julgar a frequência, importância ou probabilidade de uma ocorrência pela facilidade com que exemplos imediatos vêm à mente, priorizando informações facilmente acessíveis.
Viés de enquadramento	tendência de basear decisões na forma como a informação é apresentada - conotações positivas ou negativas - e não nos fatos presentes.
Viés de escassez	tendência de atribuir maior valor subjetivo a itens mais raros, difíceis de obter ou de maior demanda.
Viés de normalidade	tendência de subestimar a probabilidade e as possíveis consequências de eventos e de acreditar que as coisas sempre funcionarão da forma como ocorrem normalmente.
Viés de otimismo	tendência de superestimar a probabilidade de eventos positivos e subestimar a probabilidade de eventos negativos.
Heurística de prioridade	tendência de tomar uma decisão baseada em apenas uma informação dominante.
Viés retrospectivo	uma distorção da memória pela qual, as pessoas tendem a perceberem eventos prévios como mais previsíveis do que foram ou como inevitáveis.
Efeito bandwagon	pode ser considerado uma forma do viés de conformidade, é a tendência em adotar crenças e comportamentos, principalmente porque já foram adotados por outras pessoas.
Reciprocidade	tendência de responder uma ação positiva com outra ação positiva e ter dificuldade em dizer não ou ficar "devendo"à outra pessoa.
Prova social	é a tendência de adaptar ou copiar ações e opiniões de outros, com o intuito de adotar comportamento considerado correto ou esperado em uma determinada situação.

- Viés de autoridade: o ataque é desenhado para simular autoridade, utilizando-se de informações, imagens e outros recursos que possam transmitir autoridade e despertar confiança. Ataques bem elaborados tendem a reunir diversos elementos na tentativa de não levantar suspeitas, incluindo o emprego de linguagem comum a área, como o uso de termos técnicos, o que pode ser facilitado pelo uso de Inteligência Artificial (IA);

- Viés de conformidade: com intuito de gerar um comportamento semelhante, o atacante informa ou da margem para que a pessoa entenda que diversas outras pessoas obtiveram benefícios ao realizar determinada ação ou resposta;

- Viés de disponibilidade: diante de uma solicitação de informação ou para executar uma ação, como desabilitar temporariamente o software de anti-vírus, o indivíduo tende a superestimar a informação de nunca ter experienciado um golpe e subestimar a probabilidade do caráter malicioso da solicitação;
- Viés de escassez: utilizando-se do senso de urgência, de restrições de tempo ou de recursos, ou da ideia de que é uma oportunidade única, atacantes tentam manipular o indivíduo para que realize uma ação;
- Reciprocidade: com a intenção de que a pessoa realize uma ação e ofereça algo em troca, o atacante inicialmente oferece alguma suposta vantagem, prêmio ou favor.

É importante destacar que os vieses cognitivos são intrínsecos à natureza humana e ataques de ES estão intimamente associados á influências situacionais e contextuais. No entanto, ter consciência acerca dos vieses e da forma como são utilizados na ES nos torna cientes da nossa própria vulnerabilidade e nos permite assumir que podemos ser manipulados, influenciados e moldados. Contribuindo para que possamos desenvolver senso crítico ao avaliar informações e interações *online*.

Aprimorando, assim, habilidades para reconhecer e questionar informações ou solicitações suspeitas, bem como facilitando a identificação de potenciais tentativas de manipulação. Além disso, do ponto de vista coletivo, melhora as competências de sensibilização de outras pessoas para os riscos, desempenhando um papel na promoção da cultura de segurança.

Ademais, o processo de tomada de decisão, assim como o comportamento, pode ser influenciado por experiências pregressas e por conhecimento prévio. Desta forma, saber sobre os vieses e como operam, pode antecipar percepções e facilitar a adoção de comportamentos mais cautelosos.

3.3.3. Mídias Sociais, Psicologia e Engenharia Social

Mídias sociais podem ser definidas como canais baseados na internet, com suporte a interações sociais síncronas e assíncronas, que permitem a transmissão de comunicações com públicos amplos e restritos [Bayer et al. 2020] [Carr and Hayes 2015]. A definição também incluí a presença de quatro elementos que caracterizam essa tecnologia [Bayer et al. 2020].

1. Perfil: elemento que permite aos usuários manter conjuntos exclusivos de atributos pessoais criados pelo próprio usuário, pelos usuários de sua rede e/ou pela plataforma [Bayer et al. 2020];
2. Rede: representa conexões sociais e pode ser compreendido como o conjunto de contatos criados por meio de "amizade"mútua ou do "seguir"unilateralmente [Bayer et al. 2020];
3. *Stream*: refere-se ao fluxo de conteúdo, é o elemento de mídia social que permite ao usuário consumir e/ou interagir com *feeds* de conteúdo gerado por outros usuários de sua rede [Bayer et al. 2020];

4. Mensagem: elemento que possibilita aos usuários o envolvimento em interação social direcionada usando texto, vídeo, foto ou qualquer outra mídia suportada pela plataforma utilizada [Bayer et al. 2020].

Em pesquisas psicológicas, uma abordagem a partir dos elementos fundamentais a essas mídias permite aos pesquisadores conceituar mídias sociais sem restringir-se em particularidades relacionadas às plataformas específicas e estabelecer uma base mais duradoura do que o tempo em que uma plataforma se mantém ativa, para assim observar seus efeitos e implicações [Bayer et al. 2020].

Além disso, esses elementos associam-se a efeitos [Bayer et al. 2020], que podem ser vistos como ações e práticas já observadas em ambientes *offline*, mas que também operam no ambiente *online*. No que tange ES, estes efeitos são relevantes porque são utilizados por atacantes, se correlacionam com alguns dos vieses explorados e podem contribuir significativamente para o êxito do ataque de ES.

Abaixo são listados e descritos três destes efeitos e o elemento relacionado, conforme a literatura [Bayer et al. 2020]. Também é realizada e exposta a associação com alguns dos vieses previamente apresentados.

- Autoapresentação: consiste no processo de controlar a percepção de outras pessoas a respeito de alguém [Leary 2019] [Attrill-Smith et al. 2019a], na tentativa de modificar uma resposta com objetivos sociais ou pessoais [Leary and Tangney 2014]. A autoapresentação é um fenômeno presente nas interações sociais, não se restringe as interações mediadas por tecnologias e é considerado também um processo de gerenciamento de impressão [Attrill-Smith et al. 2019a]. Inclui os aspectos deliberados da modificação, mas também aspectos menos conscientes, que podem estar vinculados a normas e expectativas ou componentes culturais, por exemplo. A autoapresentação, não tem necessariamente um caráter de falsidade. Em interações *offline*, indivíduos se comportam de formas diferentes em situações e ambientes variados, o mesmo ocorre nas interações mediadas por tecnologia [Attrill-Smith et al. 2019a].

  São diversas as razões pelas quais as pessoas utilizam a autoapresentação [Attrill-Smith et al. 2019a], diversas também são as motivações de uma pessoa para modificar ou esconder partes ou a totalidade de uma identidade [Leary and Tangney 2014]. Essas alterações podem ocorrer motivadas pela liberdade de expressar-se de uma maneira distinta, por exemplo, adotando uma versão menos introvertida de si [Attrill-Smith et al. 2019a] ou para atingir um objetivo [Leary and Tangney 2014], como adotar um modo de autoapresentação em uma rede profissional com intuito de conquistar uma vaga de emprego [Attrill-Smith et al. 2019a].

  Em mídias sociais é por meio do componente perfil, da sua criação e organização, que a autoapresentação se expressa de maneira direta [Bayer et al. 2020]. Na ES, a autoapresentação estaria motivada pela busca de um objetivo. E em contextos maliciosos, atacantes podem criar e organizar perfis falsos, construindo uma imagem que favoreça a forma como são vistos e gere credibilidade. Sendo capazes de simular e transmitir uma suposta autoridade, explorando assim o viés de autoridade.

- Conexões sociais: no âmbito das mídias sociais, conexão social pode ser compreendida como um efeito ligado as necessidades humanas de relacionamento com outros indivíduos [Ryan and Deci 2000] [Bayer et al. 2020]. Através da conexão social pessoas podem experimentar sentimentos de aceitação e isso favorece a sensação de vínculo. O estabelecimento de conexão e de um bom vínculo aumenta a probabilidade de êxito de ataques de ES.

  Em mídias sociais a conexão social se relaciona com diferentes elementos. Ainda que o elemento de rede seja o seu representante, a conexão social como efeito se manifesta por meio do elemento de comunicação direta, a mensagem.

  Através de contato direto, utilizando-se do recurso para envio de mensagem, o atacante estabelece comunicação com a vítima, se apresenta tanto pelo seu perfil como na comunicação direta e tenta formar um vínculo. Esse elemento pode influenciar o efeito da conexão social, bem como favorecer o viés de ancoragem. O indivíduo tende, então, a dar maior relevância as primeiras informações recebidas, assim como as primeiras percepções e basear nisso as suas decisões e ações subsequentes.

- Mobilização social: também fundamenta-se na necessidade de pertencimento e de se sentir socialmente conectado a outros indivíduos. Refere-se à princípios que podem ser usados para influenciar grupos de indivíduos a participar de determinadas atividades [Rogers et al. 2018]. A presença em mídias sociais como as redes sociais, potencializa e reforça os impactos da mobilização social [Rogers et al. 2018].

  Em nível individual, pesquisas sugerem que a transmissão de conteúdos com natureza de pedidos de ajuda, tendem a receber mais interações e respostas [Bayer et al. 2020]. A nível coletivo, esse fenômeno também pode ser observado nos eventos como a Primavera Árabe em 2011 [Gohn 2014] e os movimentos do Brasil em 2013 [Solano and Rocha 2019].

  Um atacante poderia explorar o efeito de mobilização promovido pelas mídias sociais para sensibilizar, a partir de apelos como pautas políticas e causas sociais, e induzir um individuo a executar determinada ação, como clicar em um link malicioso. Apoiado no elemento rede, com indivíduos reais ou composta por uma série de perfis falsos, o atacante poderia se valer da prova social. Encorajando, assim, um individuo ou grupo, a partir da mobilização de outros.

  Nesse exemplo, é possível identificar uma potencialização da probabilidade de êxito do ataque de ES. Uma vez que a mobilização social tende a influenciar grupos de pessoas para adoção de um comportamento ou execução de uma ação [Rogers et al. 2018].

A lista e a descrição dos efeitos e suas relações com os elementos de mídias sociais, obviamente não foram esgotadas nessa sessão. Considerando o escopo do presente capítulo, foi privilegiada uma apresentação parcial. A aplicação de alguns tópicos cobertos nessa sessão, pode ser observada de forma prática na sessão de estudos de caso.

Por fim, faz-se necessário destacar que plataformas de mídias sociais devem adotar medidas para proteger seus usuários. Implementando controles que dificultem a aplicação de ES e promovendo a conscientização e a sensibilização dos usuários por meio de campanhas e outros recursos informacionais.

3.4. Automação de Ataques de Engenharia Social

Os ataques de ES buscam estabelecer uma posição privilegiada do atacante no fluxo de informações, tendo como objetivo a construção de uma relação de confiança com a vítima, que é obtida por meio da manipulação psicológica. A automação da ES permite que esses ataques sejam realizados de forma mais eficiente e escalável, por meio de sistemas automatizados.

O uso de automação vem sendo uma tendência na área de tecnologia, permitindo que atividades normalmente repetitivas venham a ser executadas sem a necessidade de interferência humana, oferecendo diversas vantagens como a velocidade de execução das tarefas e a menor possibilidade de erros. O tempo normalmente investido na execução passa a ser aplicado no planejamento e definição das regras a serem seguidas, podendo seguir um modelo estrito de instruções passo a passo, como um *script*, ou com maior capacidade de entendimento e reação, como com o uso de inteligência artificial e aprendizado de máquina.

Da mesma forma que a evolução dessas tecnologias permite a melhoria e inovação, as mesmas podem ser também aplicadas a objetivos maliciosos. Da mesma forma que a internet e a computação trouxeram a possibilidade que golpes antes realizados pessoalmente ou através de cartas tivessem a capacidade de atingir um número maior de alvos, gerar mais falsos indícios que ajudem na credibilidade da isca e uma menor exposição do atacante, o uso de automação eleva essa capacidade a outros níveis.

Levando em conta os quatro estágios de um ataque de SE [Mitnick and Simon 2003]:

1. Obtenção de informações sobre a vítima;
2. Construção de uma relação de confiança;
3. Exploração das informações obtidas;
4. Execução do ataque;

O uso da automação permite maior agilidade na execução de tarefas associadas a cada fase, como por exemplo, retornar ao atacante as informações de interesse dos perfis de rede social do alvo. O ganho de tempo que permite maior eficiência de algum profissional traz a mesma vantagem ao agente malicioso.

O uso aprofundado dessas técnicas permite a rotulagem própria dessa forma de ataque, classificada como Engenharia Social Automatizada (ESA), onde o uso da automação se integra completamente à execução dos ataques, sendo peça fundamental para o seu ciclo, ou mesmo em casos extremos podendo rodar o ataque por inteiro sem necessidade de interferência humana, atingindo o máximo de escalabilidade com a menor exposição possível por parte do atacante.

3.4.1. Engenharia Social Automatizada

A Engenharia Social Automatizada (ESA) é uma abordagem que combina técnicas de ES com automação por meio de ferramentas e *scripts* para criar ataques eficazes em escala. Os ataques de ES demandam tempo e recursos para desenvolver uma relação de confiança

do atacante com o usuário. Sendo assim, ao automatizar os aspectos repetitivos e tediosos do processo, os atacantes utilizam a ESA para lançar ataques em larga escala de maneira mais eficiente [Guzman and Lewis 2020].

A comunicação humana tem sido a base para o desenvolvimento de interfaces homem-máquina os atacantes tem utilizado os *Bots* para automatizar os passos para estabelecer uma conexão de confiança com o usuário [Shafahi et al. 2016]. Essa relação de confiança faz uso da manipulação psicológica incentivando os usuários interagirem com ferramentas utilizadas pelos engenheiros sociais no ambiente digital. A combinação de táticas de manipulação psicológica com tecnologia avançada possibilita que atacantes atinjam múltiplos alvos com eficiência surpreendente.

Com o uso diário das redes sociais e o grande volume de dados no ciberespaço, os engenheiros sociais passaram a espalhar *Bots* com comportamento semelhante ao do ser humano para um grande número de usuários. Esses *Bots* simulam conversas humanas, conhecidos como *ChatBots* e, os que atuam nas redes sociais, os *SocialBots* [Shafahi et al. 2016].

Bot é uma ferramenta automatizada para implementar uma série de funções préprogramadas de operação e controle.. Os *Bots* podem ser autênticos para realizar tarefas úteis para os usuários. No entanto, existem *Bots* com foco malicioso, que podem realizar ataques para obter informações relevantes ou manter o controle do dispositivo acessado. *Bots* podem ser utilizados para ações de disseminação de informações falsas (*fake news*), *spam* e *phishing* [Mitnick and Simon 2003]. [Freitas et al. 2015].

SocialBot é uma ferramenta de *software* que simula o comportamento humano para realizar interações automatizadas nas redes sociais . Os *SocialBots* têm a capacidade de comprometer a estrutura das redes sociais, influenciando os usuários e aumentando o número de seguidores, para inflar os índices de popularidade de uma determinada conta de perfil. Essa ferramenta é eficaz para ataques de ES, utilizando-se de informações sensíveis de possíveis vítimas, como o roubo de identidade [Rouse 2013] [Camisani-Calzolari 2012] [Dewangan and Kaushal 2016].

ChatBot é a integração de sistemas, ferramentas e roteiros que promovem conversas por mensagens instantâneas com ou sem a participação de humanos. São desenvolvidos para ajudar usuários humanos em situações de serviços específicos, não sendo exaustivo. Por exemplo: atendimento ao cliente, atendimento por telefone e serviço de educação digital [Grimme et al. 2017]. O uso da linguagem natural nos *ChatBots* é um desafio a ser superado para o desenvolvimento dessa ferramenta [Khan and Das 2018] [Stoeckli et al. 2018].

3.4.2. Ferramentas e Técnicas Utilizadas na Engenharia Social Automatizada

À medida que o avanço tecnológico gera um crescimento exponencial do volume de dados no ciberespaço, tem-se como resultado uma dependência cada vez maior dos recursos tecnológicos. Por decorrência disso, a superfície de ataque para os engenheiros sociais tem aumentado, considerando o uso da automação das ferramentas e técnicas no campo da ES.

Ao explorar o comportamento humano e o uso de sistemas de informações, os

atacantes se beneficiam de uma compreensão profunda dos processos cognitivos, a fim de obter dados e informações relevantes de potenciais alvos. Utilizando ferramentas e técnicas de automação, os engenheiros sociais buscam estabelecer relações de confiança, manipulando psicologicamente as pessoas para que realizem ações específicas.

Algumas das principais ferramentas e técnicas utilizadas nessa abordagem são as seguintes:

1. Perfis falsos: é uma técnica amplamente adotada na Engenharia Social Automatizada é a criação de perfis falsos em plataformas de redes sociais. Esses perfis fictícios são meticulosamente construídos com o objetivo de estabelecer relações de confiança com possíveis alvos, explorando suas fraquezas e vulnerabilidades. Essa abordagem permite aos engenheiros sociais automatizar a criação e a manutenção de múltiplos perfis falsos, ampliando significativamente o alcance de seus ataques.

2. Análise e Mineração de Dados: a ESA depende de uma análise e obtenção eficaz de dados sobre potenciais vítimas. Nesse sentido, técnicas de mineração de dados, análise de redes sociais e obtenção de informações pessoais publicamente disponíveis são aplicadas. Essas técnicas permitem identificar alvos em potencial, compreender suas preferências, hábitos e padrões de comportamento, aumentando assim a eficácia dos ataques.

3. Manipulação Psicológica e Persuasão Automatizada: explorando as fragilidades do ser humano a ESA também abarca o uso de técnicas de manipulação psicológica e persuasão automatizada. Utilizando algoritmos e Inteligência Artificial (IA), é possível personalizar mensagens e interações para se adequarem ao perfil de cada potencial vítima. Essas técnicas visam explorar os mecanismos cognitivos e emocionais dos usuários, persuadindo esses usuários a realizar ações desejadas pelos engenheiros sociais.

4. *Bot* é o termo resumido da palavra da língua inglesa *robot*, que na tradução livre significa robô. É uma ferramenta automatizada que realiza uma série de funções pré-programadas de operação e controle. Os *Bots* podem ser autênticos, que têm como objetivo realizar atividades úteis para os usuários, por outro lado também existem *Bots* de cunho malicioso, que podem realizar ataques para obter informações relevantes ou manter o controle do dispositivo acessado. *Bots* podem ser utilizados para ações de disseminação de informações falsas (*fake news*), *spam* e *phishing* [Freitas et al. 2015].

   No contexto de ESA os cibercriminosos usam os *Bots* maliciosos para simular o comportamento humano, burlando os mecanismos de segurança. Com o crescimento das redes sociais e o grande volume de dados no ciberespaço, os engenheiros sociais passaram a espalhar *Bots* com comportamento semelhante ao do ser humano para um grande número de usuários. Esses *Bots* simulam conversas humanas, conhecidos como *ChatBots* e, os que atuam nas redes sociais, os *SocialBots* [Shafahi et al. 2016].

5. *ChatBot* é a integração de sistemas, ferramentas e roteiros que promovem conversas por mensagens instantâneas com ou sem a participação de humanos [Stoeckli et al. 2018]. São desenvolvidos para ajudar usuários humanos em situações de serviços específicos, não sendo exaustivo. Por exemplo: atendimento ao cliente, atendimento por telefone e serviço de educação digital [Grimme et al. 2017]. O uso da linguagem natural nos *ChatBots* é um desafio a ser superado para o desenvolvimento dessa ferramenta [Khan and Das 2018].

6. *SocialBot* é uma ferramenta de *software* que simula o comportamento humano para realizar interações automatizadas nas redes sociais [Rouse 2013]. Os *SocialBots* têm a capacidade de comprometer a estrutura das redes sociais, influenciando os usuários e aumentando o número de seguidores, para inflar os índices de popularidade de uma determinada conta de perfil [Camisani-Calzolari 2012]. Essa ferramenta é eficaz para ataques de ES, utilizando-se de informações sensíveis de possíveis vítimas, como o roubo de identidade [Dewangan and Kaushal 2016].

Essas ferramentas e técnicas têm sido desenvolvidas com a ajuda de mecanismos de IA que interagem com os usuários [Freitas et al. 2015]. A IA é similar a inteligência humana, desenvolvida com a automatização conforme a necessidade da aplicação [Ferrara et al. 2016]. Na medida que um certo grau de inteligência é incorporado nas ferramentas para simular o comportamento humano, aumenta a capacidade e escalabilidade dos ataques.

3.5. Prevenção e Mitigação de Ataques de Engenharia Social

A prevenção e mitigação de ataques de ES têm sido objeto de estudo e preocupação nas empresas, setor público e acadêmica. Esses ataques exploram a manipulação psicológica e a falta de conscientização das pessoas para obter acesso não autorizado a informações sensíveis ou comprometer sistemas de segurança. Abordagens acadêmicas para combater esse problema incluem a análise de técnicas e táticas empregadas pelos invasores, o desenvolvimento de métodos de detecção e alerta precoce, bem como a conscientização e treinamento dos usuários para identificar e evitar armadilhas de ES.

Como já discutido, a ES se baseia em dois principais pilares, os aspectos tecnológicos e as questões psicológicas humanas. A fim de obter resultados mais concretos na prevenção desses ataques é portanto necessário uma visão completa do ciclo de ataque e controles em ambos os pilares, aumentando não apenas as chances de evitar essas formas de ataque mas também que em caso de falhas em um ativo, que o ataque possa ser detectado ou impedido em outros níveis.

Enquanto a parte humana é normalmente abordada com o uso de treinamento e conscientização, auxiliando os usuários a identificarem e reportarem ações suspeitas, os controles técnicos atuam para evitar que os ataques cheguem nos usuários ou, em caso de falha das vítimas, minimizar os danos. Engenharia Social é muitas vezes utilizada apenas como forma de entrada para execução de outros ataques, portanto não basta que os riscos de ES sejam analisados individualmente, mas devem fazer parte de um programa maior de Segurança da Informação.

3.5.1. Conscientização e Treinamento dos Usuários

A conscientização e o treinamento dos usuários são elementos essenciais para na prevenção de ataques de ES, considerando que a manipulação psicológica é uma característica intrínseca desses ataques, que tem foco nas emoções humanas como o medo, a curiosidade e a confiança. A conscientização busca capacitar os usuários para reconhecerem as técnicas utilizadas pelos engenheiros sociais e, com isso, o conhecimento oferecido aos usuários permite uma ação defensiva em relação as tentativas de ataques. Um programa de conscientização e treinamento deve incluir:

1. Educação sobre táticas de ataque: os usuários devem ser informados sobre as táticas comuns de ES, como por exemplo: *phishing* e *spear phishing*. O entendimento como essas táticas operam ajuda os usuários a identificar sinais de alerta;

2. Simulações de ataque: simulações controladas de ataques de ES podem ser realizadas para testar a prontidão dos usuários. Com isso, é possível identificar as vulnerabilidades do ambiente computacional, bem como as oportunidades de aprendizado;

3. Desenvolvimento de habilidades críticas: os usuários devem ser capacitados para tomar decisões informadas sobre a divulgação de informações ou o clique em *links*, possibilitando avaliar a legitimidade das demandas e verificar a fonte;

4. Compartilhamento de exemplos reais: estudo de caso reais de ataques bem-sucedidos podem ilustrar as consequências da ES, tornando mais tangíveis os riscos envolvidos;

5. Atualizações regulares: as táticas de ataque são dinâmicas, é importante manter os usuários atualizados sobre as novas ameaças e métodos de defesa; e

6. Cultura de segurança: promover uma cultura de segurança onde os funcionários se sintam encorajados a relatar tentativas de Engenharia Social e outras formas de ataque ou suspeitas ajuda a criar uma abordagem coletiva para a prevenção.

Um importante fator a ser considerado nas ações de conscientização e treinamento é o da aplicação de uma abordagem social, orientada ao coletivo [Collier et al. 2023]. Tal orientação fomenta que os usuários se percebam como integrantes de cenário maior e cujo foco não seja o indivíduo [Zimmermann and Renaud 2019]. Essa perspectiva evita responsabilização excessiva e individualizada, além de promover comportamento pró segurança.

Abordagens individualizantes tendem a se basear no medo, desconsideram comportamento e motivações dos seres humanos e não contribuem para mudanças comportamentais efetivas [Collier et al. 2023]. Além disso não promovem o desenvolvimento de uma cultura de segurança, em contrapartida tendem a prejudicar o contexto e o senso crítico dos indivíduos e interferir nos processos de tomada de decisão, uma vez que elevam preocupações e tensões.

3.5.2. Técnicas de Detecção de Ataques de Engenharia Social Automatizada

Detectar ataques de ES realizados de maneira automatizada é uma abordagem complementar à conscientização e treinamento dos usuários. Esses ataques são conduzidos em grande escala, dificultando a identificação manual. Técnicas de detecção automatizada podem incluir:

1. Análise de Conteúdo: O uso de algoritmos para analisar e identificar padrões em mensagens de phishing ou em links maliciosos pode ajudar a identificar tentativas de Engenharia Social.
2. Análise Comportamental: observar o comportamento do usuário, como cliques rápidos e atípicos ou solicitações incomuns de informações, pode sinalizar atividades suspeitas.
3. Aprendizado de Máquina e Inteligência Artificial: os algoritmos de aprendizado de máquina podem ser treinados para reconhecer padrões de ataques de Engenharia Social com base em dados históricos.
4. Monitoramento de Tráfego: Observar o tráfego de rede em busca de atividades suspeitas, como tentativas de redirecionamento, pode ajudar a identificar ataques.
5. Análise de Metadados: Examinar metadados em emails e links pode revelar informações ocultas, como a verdadeira origem de uma mensagem.
6. Lista de Domínios Maliciosos: Manter uma lista atualizada de domínios conhecidos por hospedar ataques de Engenharia Social pode ser usado para bloqueio preventivo.

3.6. Estudos de Caso

Engenharia Social é definida como a técnica de explorar pessoas com o objetivo de acessar informações sobre potenciais alvos utilizando combinações de interações tecnológicas e sociais [Libicki 2018] [Klimburg-Witjes and Wentland 2021]. Já é um conceito comum a classificação dos usuários/humanos como o "elo fraco"da corrente em Segurança da Informação, visto que ataques explorando as suas falhas tem tido maiores taxas de sucesso e, muitas vezes, necessitando menores habilidades técnicas ou exposição à risco por parte do atacante [Darwish et al. 2012].

3.6.1. Estudo de Caso 1 - LinkedIn

As redes sociais possuem o objetivo de permitir interação virtual entre humanos. Porém além de oferecer um espaço onde as distâncias físicas possam ser ultrapassadas para permitir essas conexões, as redes também trazem para o ambiente virtual muitos dos riscos e ameaças existentes no mundo real. Uma diferença pertinente porém está no fato dos usuários desses espaços de convivência não terem o mesmo nível de conscientização, alerta ou capacidade de reconhecer riscos que teriam no mundo físico, o que aliado a maior capacidade de anonimização e personificação torna esses ambientes um espaço de grande interesse para Engenharia Social [Crossler and Bélanger 2014].

Comparando com outras redes sociais como *Facebook*, *Twitter* e *Instagram*, redes sociais profissionais replicam um ambiente corporativo, focado em conexões profissionais e crescimento de carreira. Esses cenários ligados ao mercado de trabalho criam um sentimento de seriedade, confiança e credibilidade, não tendo um foco em entretenimento ou distração. Esses ambientes atraem recrutadores em busca de candidatos, assim como empresas em busca de potenciais clientes. As relações existentes em redes sociais profissionais já são exploradas por engenheiros sociais, em especial a personificação de recrutadores, utilizando vagas de trabalho atrativas como isca para roubar informações confidenciais de empresas ou dados pessoais das vítimas. [1]

Atualmente o *LinkedIn* é a rede social profissional mais popular, com mais de 930 milhões de usuários em mais de 200 países de acordo com dados da própria rede (Maio de 2023)[2]. As Políticas de Uso do LinkedIn definem na Seção 8.2[3] as ações que não são permitidas aos usuários, destacando-se a proibição do uso de informações falsas ou personificação no perfil, e o uso de *Bots* e automação para realizar ações na plataforma.

Levando em conta as referências de ataques de ES já existentes no próprio LinkedIn, é possível encontrar referências ao uso de perfis e informações falsas para os mais diversos motivos. Analisando especificamente sobre automação, realizando uma busca em sites de repositórios de código é possível encontrar facilmente dezenas de *Bots* e *Scripts* especificamente para uso no *LinkedIn*. Se as políticas da rede social proíbem o uso de dados falsos e automação, porém sendo possível identificar o uso dos mesmos, há indicação de uma potencial falta ou insuficiência na implementação desses controles por parte da plataforma, ou a aplicação das regras é feita apenas baseando-se em reclamações feitas por outros usuários. Além dessas observações, o estudo de caso analisado realizou uma avaliação do capacidade do LinkedIn em detectar e/ou bloquear o uso de automação e informações falsas, sendo estes os requisitos básicos para executar um ataque de ESA contra os usuários da plataforma.

3.6.1.1. Metodologia

O teste foi realizado em um cenário de prova de conceito com 2 *Bots* para avaliar o ataque. O primeiro interagia com a rede social para buscar e contatar alvos - o *Bot* Plataforma. O segundo é um serviço de *ChatBot* que executaria uma entrevista de emprego com as vítimas - o *Bot* Recrutador. Também para dar suporte à execução das ações foi criado um perfil falso no LinkedIn, personificando um recrutador.

Para o *Bot* Plataforma foi desenvolvido um código *Python* para conexão com o LinkedIn. O LinkedIn também oferece uma API[4] para interações via *software*. Entendendo porém que usuários regulares não navegam em uma rede social através de uma API, utilizar esse canal para os testes potencialmente afetaria os resultados buscados. Para melhor reproduzir a mesma interação que um humano, foi utilizada a biblioteca Selenium[5],

[1] https://www.ft.com/content/a8d262f4-5d52-4464-8714-e21a457aab33
[2] https://about.linkedin.com
[3] https://www.linkedin.com/legal/user-agreement#dos
[4] https://developer.linkedin.com/product-catalog
[5] https://www.selenium.dev/

para permitir que o *bot* se comunicasse em formato padrão via navegador.

Para o *Bot* Recrutador, existiam diversas opções disponíveis que poderiam executar as ações necessárias sem que fosse preciso desenvolver uma nova aplicação. Usando um grupo pré-definido de questões padrão de entrevistas de emprego, junto a informações coletadas dos perfis das vítimas no *LinkedIn*, o *Bot* Recrutador basicamente conduz uma falsa emprevista de emprego com a vítima, tendo como objetivo porém obter informações sensíveis. Sendo possível executar tanto questões mais genéricas/padronizadas feitas por times de RH, quanto questões mais técnicas sobre as experiências profissionais atuais e anteriores, o processo inteiro foi executado utilizando o mesmo *bot*. Conforme o objetivo do atacante o *bot* pode incluir questões sobre empresas ou experiências específicas da vítima - coletadas do seu próprio perfil na rede social - buscando obter informações sensíveis sobre projetos, clientes, etc. Também, ao final da entrevista, o atacante poderia "selecionar"a vítima para a vaga, roubando dados pessoais através da assinatura de um falso contrato de trabalho e solicitando documentos como o passaporte, por exemplo, permitindo a realização de ataques posteriores de roubo de identidade.

De forma similar a estrutura de quatro estágios de um ataque de ES [Mitnick and Simon 2003], a proposta do projeto em questão segue um formato de etapas similar: i) Autenticação, ii) Busca, iii) Abordagem e iv) Entrevista. Esse formato permitiu a quebra do ataque em estágios e a verificação individual de cada etapa. A Figure 3.1 indica as fases do ataque testadas por cada um dos *Bots* de prova de conceito propostos.

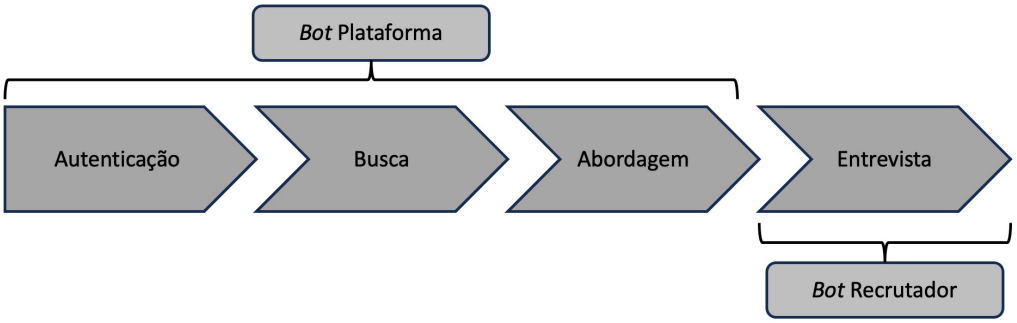

Figura 3.1: Relação entre os *Bots* propostos e as fases de ataque.

1. Autenticação: Como ilustrado na Figura 3.2, o objetivo dessa etapa é verificar se a rede social detecta ou tem diferentes comportamentos quando a autenticação do usuário é realizado através de automação. Para essa avaliação, o *Bot* Plataforma abre a página do *LinkedIn* no navegador, mapeia o código-fonte da página principal para identificar os campos das credenciais, preenche os mesmos com os valores recebidos e então submete as credenciais ao servidor e conclui o processo de autenticação, acessando a página principal do usuário.

2. Busca: Esta etapa busca verificar a detecção de uma busca automatizada de perfis. Similar a primeira etapa, o *Bot* Plataforma mapeia o código-fonte da página, identifica o campo de busca, realiza a busca utilizando os termos recebidos e então armazena temporariamente os perfis retornados. A Figura 3.7 também se refere a essa etapa.

3. Abordagem: O objetivo dessa etapa é iniciar a interação com os perfis de

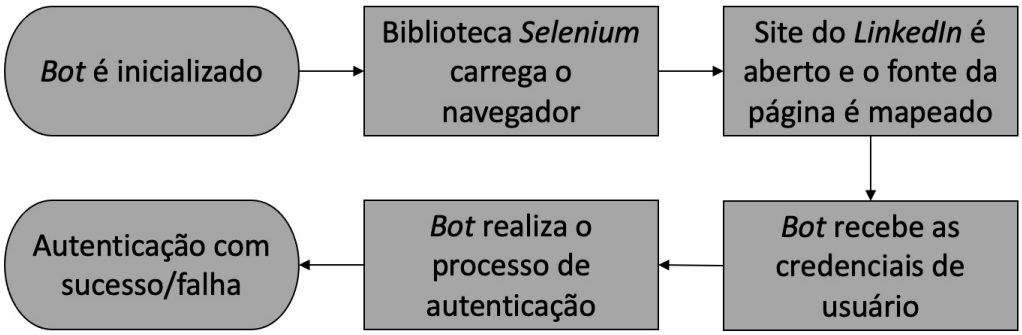

Figura 3.2: Fase de Autenticação.

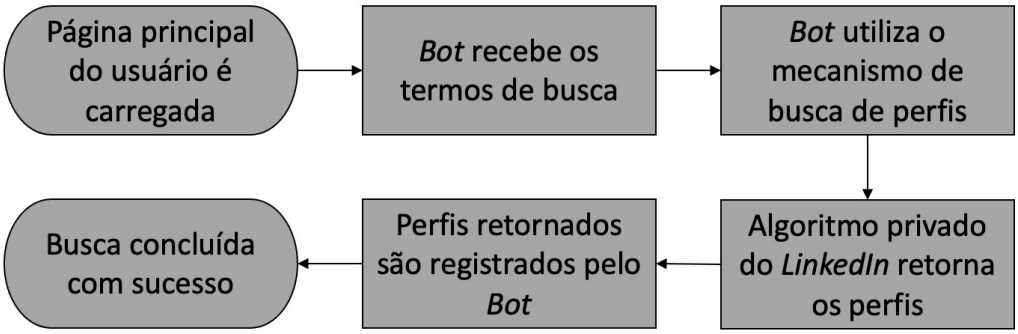

Figura 3.3: Fase de Busca.

potenciais vítimas coletados na etapa anterior. Como visto na Figura 3.4, usando os perfis capturados, o *Bot* Plataforma adiciona as vítimas como contatos e envia uma mensagem customizada, que serve como isca para as interações. Essas ações são realizadas com todos os perfis capturados.

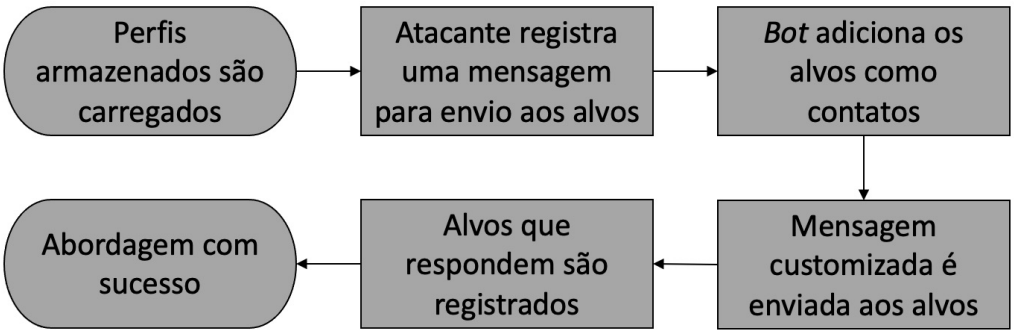

Figura 3.4: Fase de Abordagem.

4. Entrevista: Baseado nos resultados da abordagem realizada na etapa 3, um *script* captura as informações do perfil da vítima no *LinkedIn* para alimentar o banco de dados do *Bot* Recrutador, o qual terá então informações suficientes para executar uma entrevista de emprego com a vítima. A Figura 3.5 também ilustra o ciclo completo desta etapa. Como em geral é comum que um recrutador real aborde indivíduos através do *Linkedin* e então realize a entrevista e outra etapas em diferentes canais de comunicação, é esperado que a isca inclua um convite para realizar uma entrevista em um canal diferente

do próprio sistema de mensagens do *LinkedIn*.

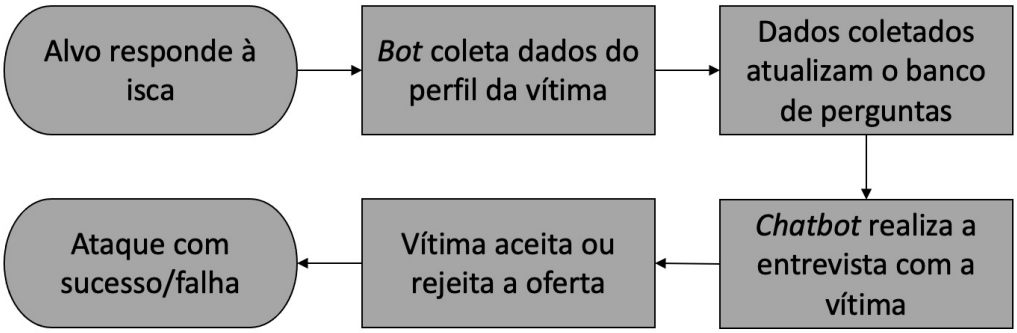

Figura 3.5: Fase de Entrevista.

3.6.1.2. Limitações

A Engenharia Social nasce no campo da psicologia, pois embora utilize da tecnologia como suporte, para atingir seus principais objetivos os atacantes exploram fraquezas humanas e características comportamentais, tópicos de pesquisa das ciências sociais. Para um completo entendimento do impacto de um ataque de ES, seria necessário não apenas validar os aspectos técnicos envolvidos, mas também enganar indivíduos e observar seu comportamento e ações. O campo da psicologia social, especialmente por conta dos diversos cenários envolvendo pesquisas com seres humanos, vem enfrentando desafios e discutindo os limites de ética em pesquisa há um longo tempo. A história nos traz exemplos extremos como os famosos experimentos de Milgran nos anos 70 [Riecken 1974], resultando em grande impacto e trauma nos participantes, situações que a pesquisa moderna entende como anti-éticas.

Expôr pessoas a situações onde elas devem ser enganadas, ter suas vulnerabilidades exploradas sem o seu consentimento (ou ao menos seu completo entendimento da situação) viola diversos dilemas éticos. Como resultado, posteriormente pode-se criar frustração e estresse por conta das expectativas criadas, promessas quebradas ou o sentimento de ter sido enganado ou manipulado. Entender e respeitar esses limites foi um dos guias desse estudo, e mesmo que isso não seja um tópico técnico não é possível executar uma pesquisa no campo de ES ou quaisquer outros tipos de ataques em cibersegurança sem discutir ou levantar questões sobre ética em pesquisa.

O primeiro desafio surgiu em como validar a proposta de ataque respeitando as políticas éticas. Como forma de atingir esses objetivos, foi decidido quebrar o ciclo de ataque em diferentes etapas e testá-las individulamente. Os resultados permitiriam um entendimento mínimo da resposta da aplicação, e cruzando os dados entre as diferentes fases permitiria ter conclusões quanto ao potencial de um ataque completo.

Especialmente a fase de Abordagem foi testada em uma linha tênue entre manter as premissas e violar barreiras éticas. Como seria necessário enviar requisições para usuários reais da plataforma, foi decidido limitar a menor quantidade possível de usuários recebendo a requisição e a mensagem. Essa quantidade foi demarcada pelos perfis

retornados na primeira página de resultados (normalmente entre 15 e 21 perfis), sendo que abordar todos eles simultaneamente ou numa janela de tempo bastante curta já indicaria um mínimo de uso de automação ou envio ativo de *SPAM* ou similares. Como o objetivo não era medir a resposta dos usuários à isca, e entendendo os requisitos para ter pessoas participando da pesquisa, após o envio das requisições e mensagens e validando que não houveram bloqueios ou similares ocorrendo na plataforma, todas as interações eram imediatamente canceladas/excluídas, a fim de evitar a chance que fossem vistas ou respondidas por usuários reais.

Entender a diferença entre quantas requisições por segundo um usuário testando a plataforma poderia fazer, comparado a um usuário apenas navegando de forma normal na rede permite identificar comportamentos que caracterizem automação sem a necessidade de identificar o limite máximo da aplicação ou gerar negação de serviço. Mesmo que potencialmente algum tipo de controle possa ser ativado após centenas ou milhares de requisições serem feitas, isso indicaria muito mais um controle contra negação de serviço ou controle de tráfego do que uma proteção contra automação. Portanto números altos não necessariamente indicam comportamento automatizado.

Para a fase de Entrevista, a avaliação focou na principal funcionalidade do *chatbot* de recrutamento: uma entrevista de emprego. A única diferença entre uma entrevista real e uma maliciosa são os objetivos, pois ao invés de tentar avaliar a capacidade e habilidades de um indivíduo para um certo cargo, o foco do atacante seria na obtenção de dados através de perguntas ou pela assinatura de um contrato de trabalho e apresentação de documentos, numa contratação falsa. O uso de um *chatbot* recrutador já existente permitiu apenas coletar informações sobre uma vítima para alimentar o *bot* e observar se as questões maliciosas propostas eram corretamente distribuídas na entrevista, sem necessidade de validar o *chatbot* em si e sua capacidade.

Considerando todos os mecanismos implementados e as decisões em como testar cada fase, seria possível concluir que suficientes resultados poderiam ser obtidos para validar o potencial do ataque proposta sem necessidade de violar nenhum requisito ético. Essa precaução é um tópico vital para qualquer tipo de pesquisa similar, e uma análise mais profunda do impacto de pesquisa ética, em especial no campo da ES, é um assunto importante para trabalhos futuros.

3.6.1.3. Experimentos

A primeira etapa foi a criação de um perfil falso que deu suporte a realização do ataque. Foi utilizada uma imagem de um banco de imagens gratuito da internet e dados aleatórios para simular experiência de trabalho prévias e formação acadêmica. Foi possível associar o perfil à empresas e universidades reais sem quaisquer verificações ou checagens necessárias. Não foi identificado nenhum impacto por conta do uso de dados falsos, ou pelo fato que desde a criação do perfil todas as interações com a plataforma do *LinkedIn* foram realizadas utilizando alguma forma de automação. A Figura 3.6 mostra algumas informações destacadas do perfil falso criado.

Os experimentos de simulação seguiram as etapas propostas no fluxo de ataque. O *Bot* Plataforma foi executado em uma máquina Windows rodando *Python*, a biblioteca

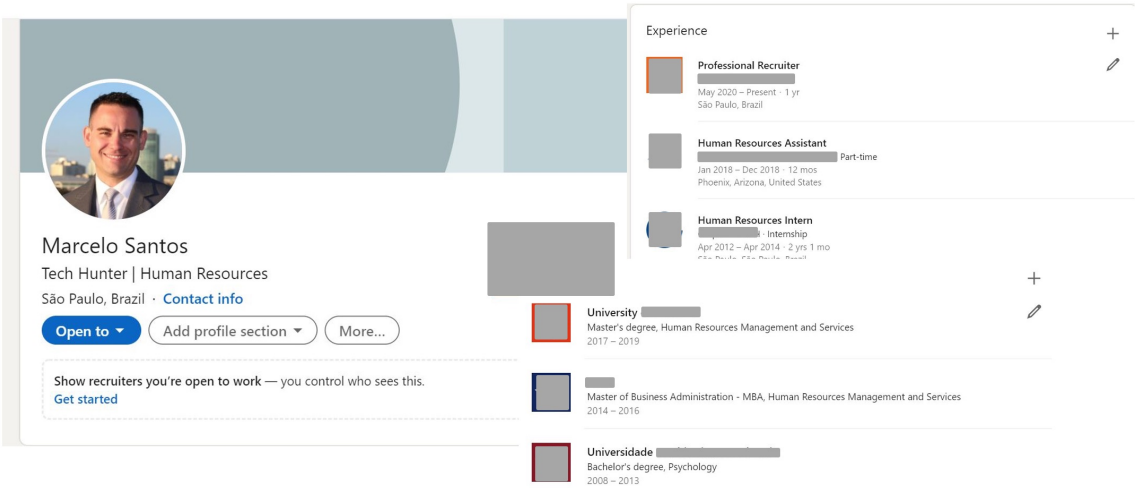

Figura 3.6: Perfil falso de recrutador criado nos testes.

Selenium e o navegador *Google Chrome*. Para o *Bot* Recrutador, foi utilizado a plataforma *SAP Conversational AI*.

Teste Etapa 1 - Autenticação:

O critério de sucesso desta etapa era executar a autenticação na plataforma seguindo diferentes comportamentos e observando se alguma das simulações ativaria controles ou bloqueios na aplicação por conta das características da automação. Como critério comparativo, foram definidos três padrões básicos de comportamento para os testes: (1) Realizar o processo de *logon* 10 vezes simultâneas; (2) Realizar o processo de *logon* 10 vezes com cinco segundos de intervalo entre cada tentativa; e (3) Realizar o processo de *logon* 10 vezes com dez segundos de intervalo entre cada tentativa. Esses padrões tentavam replicar comportamentos não esperados de um usuário humano quanto a quantidade ou velocidade das tentativas, especialmente como a execução ocorre através do navegador. Para os três padrões propostos, foram também testadas as seguintes variações para observar se as mesmas impactavam os resultados de alguma forma:

- receber as credenciais do perfil falso em tempo de execução via *script*;
- ler as credenciais do perfil falso de arquivo;
- uso de credenciais erradas/inválidas;
- uso de um *proxy* público para executar a tentativa de autenticação de um país aleatório, diferente do definido originalmente na criação do perfil.

Analisando os resultados dos testes, não foi possível observar quaisquer diferenças no comportamento da rede social, além de após algumas tentativas com as credenciais incorretas. Os testes de cada formato foram realizados em dias diferentes, de forma a garantir que a execução de um teste não interferisse nos demais. A critério de comparação, foram também executadas 10 tentativas sequenciais de *logon* utilizando as credenciais válidas,

repetindo com credenciais inválidas, porém todos realizados de forma manual (sem uso do *bot*), através do navegador, onde também não foram verificadas diferenças nos resultados.

Quando os testes foram realizados utilizando credenciais inválidas, tanto nos testes automatizados usando o *bot* quando nos manuais, após a sexta tentativas o *LinkedIn* passava a solicitar uma checagem de *puzzle* (similar a verificação *Captcha*) e/ou solicitava uma validação adicional, como um código enviado por email ou mensagem, para prosseguir com o *logon*. Esse padrão indicou que comportamentos de força bruta são identificados e bloqueados, o que não ocorre porém com outras formas de tentativa de acesso automatizadas.

Um ponto de discussão em potencial seria a quantidade de requisições utilizada. Apesar de uma quantidade similar de tentativas em alguns cenários possa ser reproduzida por um ser humano, este padrão ocorreria apenas em caso proposital para testes, não para uso regular da rede social. O processo padrão de entrada envolve inserir as credenciais, realizar a autenticação e então navegar na rede social, com alguns eventuais erros de acesso causados por confusão ou erros de digitação em algumas tentativas. Da mesma forma que após algumas tentativas sem sucesso (seis, no caso específico do *LinkedIn*, um controle adicional (*puzzle/Captcha*) é requisitado pois esse comportamento é considerado suspeito, mesmo que ele possa também ser reproduzido por um ser humano. Da mesma forma, entende-se que valores ao redor de 10 tentativas simultâneas ou em um espaço de tempo muito curso se caracterizariam como suspeitos e potencialmente de caráter automatizado.

Teste Etapa 2 - Busca:

Nessa etapa foi avaliada a capacidade do *Bot* Plataforma de executar consultas na rede social de forma automatizada sem detecção. Para realizar as buscas, foram definidas uma série de palavras-chave. É importante ressaltar que os resultados das buscas, tanto a quantidade de perfis retornados quando a ordem em que aparecem e informações similares tem relação direta com o algoritmo de busca proprietário do *LinkedIn*, e entender ou manipular os seus resultados não fizeram parte do escopo desse trabalho. Os termos de busca utilizados foram apenas uma maneira de avaliar a resposta da rede à consultas automatizadas através do navegador. A Figura 3.7 demonstra o *bot* executando a fase de busca.

Para o teste dessa fase foram listados 10 termos, baseados em alguns conhecimentos comuns na área de computação relacionados apenas para referência. Os termos usados foram os seguintes: *test* ("teste"), *Social Engineering* ("Engenharia Social"), *Bot*, *ChatBots*, *Social Networks* ("Redes Sociais"), *Information Security* ("Segurança da Informação"), *Python*, *Automation* ("Automação"), *GitHub* e *API*. De forma similar a Etapa 1, foram utilizadas as seguintes variações:

- Buscar o mesmo termo 10 vezes de forma simultânea;
- Buscar o mesmo termo 10 vezes sequenciais, com 5 segundos de intervalo entre cada consulta;
- Realizar 10 consultas simultâneas, cada uma utilizando um termo diferente;

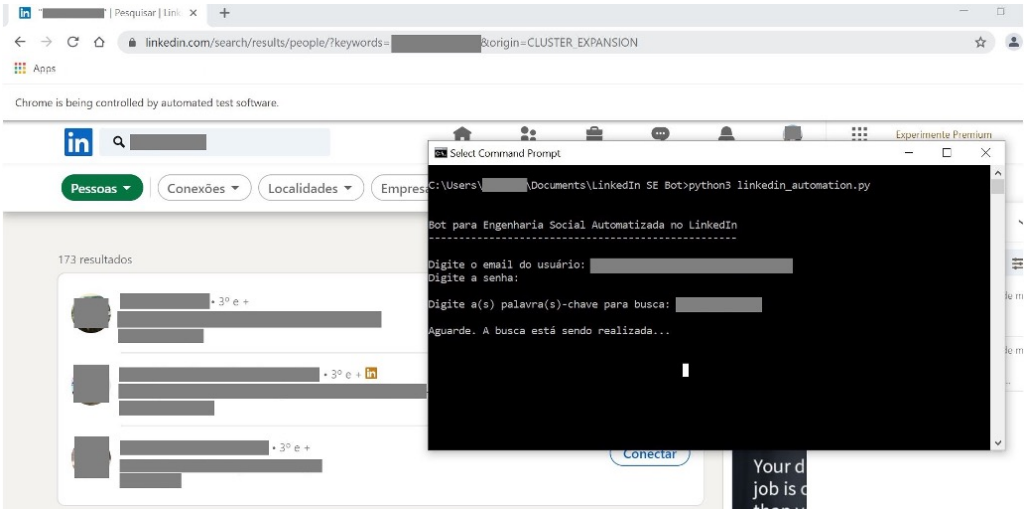

Figura 3.7: Fase de Busca sendo executada pelo *Bot* Plataforma.

- Buscar os 10 diferentes termos na mesma sessão, em sequência, com 5 segundos de intervalo entre cada consulta.

Não foi o objetivo dessa etapa realizar testes de carga ou estresse nem causar negação de serviço na aplicação. Os comportamentos testados analisam a quantidade ou velocidade das consultas em valores acima dos esperados de um usuário humano, especialmente se feitas via navegador. Apesar das diferenças esperadas nos resultados das consultas (considerando os diferentes termos utilizados e o algoritmo da rede social, o qual esteve fora do escopo de análise), não foram observadas diferenças nas variações, e todas as consultas retornaram nos resultados perfis com alguma associação com a palavra-chave utilizada.

Teste Etapa 3 - Abordagem: Na fase anterior, após realizar as consultas, o *Bot* Plataforma armazenava referências aos perfis retornados para que os mesmos forem utilizados nessa próxima fase. Aqui o principal desafio seria não violar os requisitos éticos de pesquisa, interagindo com usuários reais sem autorização ou conhecimento dos mesmos. Buscando limitar o impacto da pesquisa com a menor interferência possível nos resultados, os seguintes controles foram aplicados no *bot*:

- Para cada consulta, ao invés de armazenar as diversas páginas de resultados, foram guardados apenas os perfis da primeira página, normalmente entre 15-21 de acordo com o termo utilizado;
- A abordagem foi realizada em apenas 10 grupos, um por palavra-chave, independente das variações utilizadas nas consultas;
- Após a execução da abordagem, o *bot* registrava o sucesso do envio e então apagava/cancelava todas as ações de interação realizadas com o alvo.

A abordagem ocorria com um pedido de conexão e o envio de uma mensagem customizada, utilizando *tags* para personalizar o uso do nome real do usuário em vez de termos

genéricos. Como já citado, assim que o envio do pedido e da mensagem eram confirmados, a requisição era cancelada e a mensagem apagada, evitando quaisquer interações futuras com os usuários. Novamente nenhum impacto ou ação por parte da rede social pode ser percebida durante os testes. A Figura 3.8 mostra um exemplo de mensagem durante os testes.

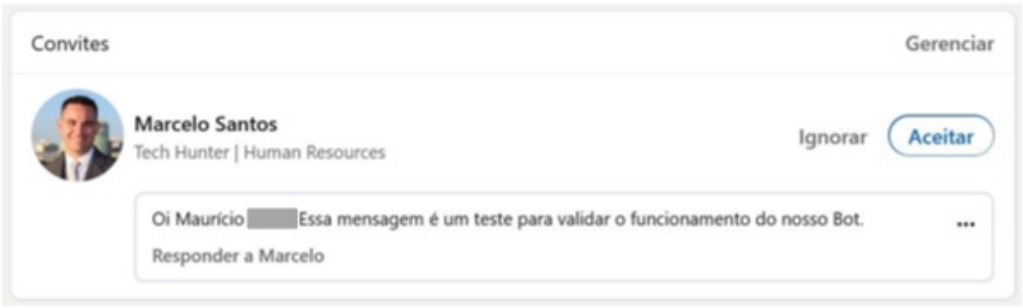

Figura 3.8: Exemplo de mensagem customizada.

Teste Etapa 4 - Entrevista:

O teste dessa etapa seguiu uma direção diferente. Foi utilizado o *SAP Conversational IA*[6], uma plataforma de criação de *chatbots*. Ao invés de criar um próprio, foi utilizado um *chatbot* recrutador já existente na plataforma, chamado *Smart Recruiter*. Utilizando essa abordagem, além de algumas verificações básicas, não seria necessário de validar a capacidade do mesmo em realizar uma entrevista, mas sim apenas de prover os valores maliciosos que seriam de interesse de um atacante e verificar os resultados. Baseado em um banco de dados inicial com questões padrão de uma entrevista de emprego já disponíveis no *chatbot*, foi utilizado um *script* para puxar os dados do perfil da vítima e utilizar os mesmos na criação de perguntas adicionais, permitindo questões mais específicas como "Conte sobre sua experiência na E=empresa X?", "Você poderia falar mais sobre suas habilidades na tecnologia Y?"ou mesmo "Você poderia mencionar os principais clientes e projetos nos quais você teve um papel chave enquanto esteve na empresa X?".

Baseando-se na observação da capacidade do *chatbot* em realizar entrevistas utilizando informações de diferentes perfis selecionados aleatoriamente da etapa anterior, o mesmo teve capacidade de realizar a entrevista completa sem necessidade de intervenção. Esses resultados permitem demonstrar a capacidade do mesmo em ser utilizado no ataque proposto sem a necessidade de abordar pessoas reais, podendo afetar os resultados buscados ou violar princípios éticos de pesquisa. A Figura 3.9 mostra parte da interação com o usuário durante uma entrevista. Apesar da tela padrão de conversa do *chatbot* ter sido utilizada durante os testes, a plataforma oferecia ferramentas que permitiram a execução da entrevista através de diferentes canais de comunicação, utilizando APIs ou *webhooks*. Seria possível inclusive ao atacante criar um *website* de uma falsa empresa e integrar o *chatbot* a mesma para criar um cenário que inspire ainda mais confiança da vítima.

[6]https://cai.tools.sap

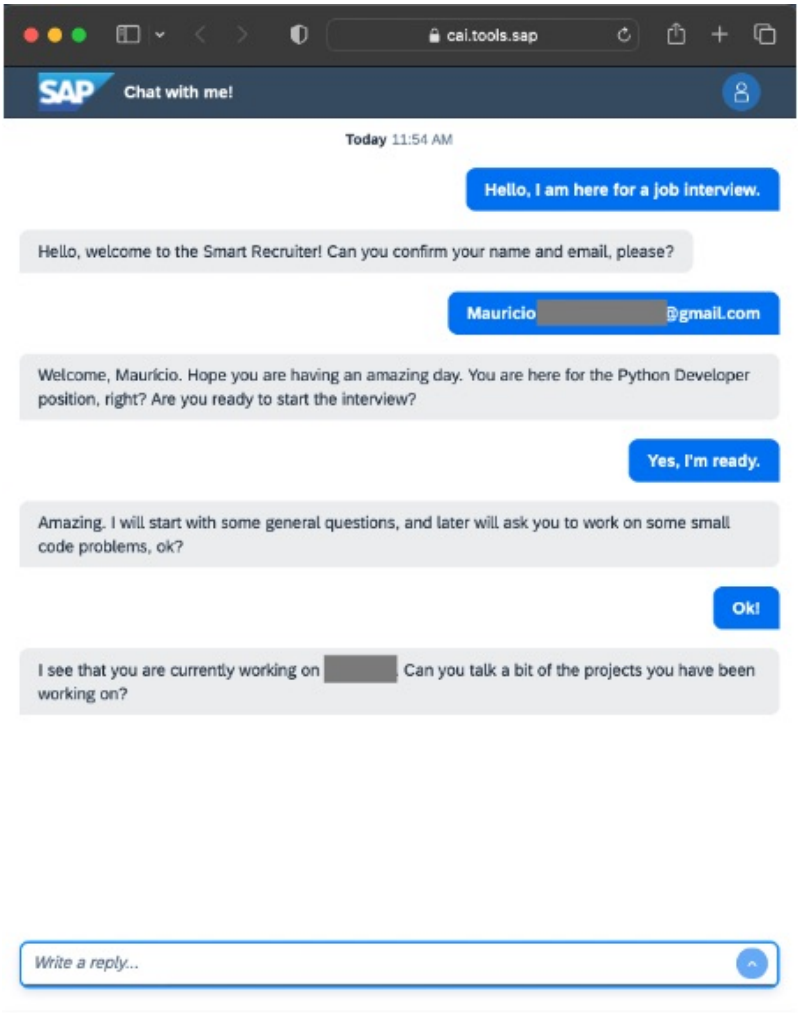

Figura 3.9: *Bot* Recrutador realizando uma entrevista.

3.6.1.4. Análise dos Resultados e Recomendações

O principal objetivo desse caso foi validar a hipótese da falta ou insuficiência de controles implementados nas redes sociais. Mesmo que esses resultados fossem esperados por algumas pessoas do campo da tecnologia, não foram encontratos pesquisas oficiais ou referências acadêmicas que pudessem embasar cientificamente essas conclusões. Não foi o objetivo desse teste específico avaliar o fator humano, que é normalmente o foco em boa parte dos trabalhos no campo de ES, mas de observar como canais tecnológicos permitem ou não oferecem suficientes barreiras para evitar ou diminuir os riscos aos seus usuários, especialmente pois em casos como o analisado o atacante pode atingir grande escalabilidade.

Certamente, o principal impacto de implementar controles rígidos em uma rede social seria a própria experiência do usuário, que como resultado por tornar a navegação menos fácil e fluída, causando potencial migração dos usuários para serviços concorrentes e causar efeitos negativos nos indicadores da plataforma. É, porém, possível encontrar um equilíbrio entre o aumento dos níveis de segurança com pouco ou nenhum impacto à

usabilidade.

Considerando que os Termos de Uso atuais do *LinkedIn* já proíbem o uso de automação, controles que permitam a detecção de comportamento automatizado poderiam ser aplicados. Eles não apenas ajudariam a evitar ESA, mas também outros padrões já proibidos. Esse controle porém poderia ser implementado com um nível de flexibilidade. Um usuário regular, conectado via navegador, possui algumas limitações humanas na sua velocidade e quantidade de requisições - acima de um certo limite, não apenas há caracterização de uso de automação, como há altas chances de *SPAM* e outras formas de interação não-solicitadas.

Baseando-se nos resultados desse estudo de caso, alguns exemplos de controles simples para identificar comportamente automatizado seriam:

- Mais de um *logon* simultâneo do mesmo usuário (com algumas variações, como considerar o caso de um usuário conectado através do computador e do *smartphone* ao mesmo tempo), ou uma grande quantidade de *logons* feita num curto espaço de tempo;
- Diversas requisições simultâneas e/ou contínuas (não apenas de busca, mas para qualquer ação) em uma quantidade ou faixa de tempo acima da capacidade normal de um ser humano;
- Diversos pedidos de conexão e/ou envio de mensagens para diferentes usuários simultâneamente ou numa curta faixa de tempo (potencialmente também indicando *SPAM*.

A aplicação dos controles propostos não precisa causar o bloqueio da requisição ou penalidades ao usuário. Exigir dados adicionais como uma validação do tipo *Captcha*, similar ao que já é implementado para evitar ataques de força bruta, já seria uma excelente forma de evitar a automação, já que o mesmo seria necessário apenas em certos cenários, não afetando a grande maioria dos usuários em seu uso regular.

Indo um pouco além, é possível imaginar que alguns usuários ou em certos cenários algum nível de automação seja útil ou necessário - para recrutadores reais, por exemplo, a aplicação desses controles pode ser mais rígida nas conexões via navegador (onde usuários humanos são esperados, não automação), e mais flexível nas conexões via API, por exemplo. Isso permitiria um melhor monitoramento e controle por parte da plataforma, podendo ser inclusive uma oportunidade de negócios oferecer uma certa quantidade de requisições para clientes pequenos (como recrutadores independentes), ou soluções mais robustas vendidas como serviço pela plataforma, como o já existente serviço *LinkedIn Recruiter*.

Os controles propostos permitiram resolver a questão da automação, aplicando políticas que já existem com o menor impacto possível aos usuários. Um desafio diferente porém é a questão dos perfis e dados falsos, problema que não afeta apenas o *LinkedIn* mas também todas as demais redes sociais nos dias de hoje. É possível estabelecer uma base de interações e contatos que crie uma sensação de legitimidade, organicamente através de usuários reais ou mesmo através de uma rede de perfis falsos.

Verificação de usuários envolve validações mais complexas. Porém o impacto dos perfis falsos tem crescido tão rápido que discussões sobre validação obrigatória de usuários já surgiram até mesmo em outras redes como o Twitter[7]. O projeto de pesquisa analisado nesse estudo de caso vem recomendando alguma funcionalidade de validação desde o seu início, e recentemente o *LinkedIn* anunciou seu programa de verificação[8]. O programa ainda tem suas limitações, como a validação através do email corporativo para algumas empresas registradas ou através de documentos de identidade (por enquanto disponível apenas nos EUA), mas certamente é um grande avanço. Considerando o objetivo do *LinkedIn* como rede social profissional, credibilidade e veracidade devem ser um ponto de interesse de todos os seus usuários. Potencialmente mesmo sem obrigatoriedade muitos usuários devem buscar a validação como forma de reconhecer seu trabalho e responsabilidade - ou mesmo alguns grupos-alvo podem ser priorizados, como recrutadores. Existem diversas oportunidades, cada com seus prós e contras, mas certamente algum controle nesse sentido é necessário para ao menos dificultar os ataques de personificação.

3.6.2. Estudo de Caso 2 - Testes Padronizados de ES

Esse estudo de caso propôs uma metodologia para realização de testes de *phishing* e executou testes com grupos de participantes em diferentes cenários para observar sua resposta e comportamentos. Um dos objetivos do trabalho foi oferecer guia para a produção de futuros testes, devendo os passos serem acessíveis, podendo ser facilmente reproduzidos, e a infraestrutura necessária com o menor custo possível, para que os testes não dependam de grandes orçamentos.

O processo se divide em cinco etapas: autorização, planejamento, montagem da infraestrutura *web*, montagem e envio dos *e-mails* e coleta e análise dos resultados. Antes de qualquer teste de Engenharia Social, por envolver informações e os colaboradores da empresa, deve-se iniciar pela autorização explícita de um gestor ou pessoa responsável, sendo uma boa prática que o cargo do mesmo seja superior ao do alvo com maior nível de hierarquia. Jamais realize um teste do tipo sem autorização formal. A etapa de planejamento então envolve a análise dos alvos e a escolha do serviço/sistema a ser clonado, diretrizes iniciais que irão nortear o restante do teste. A escolha do serviço ou sistema a ser clonado deve dar preferência à páginas que solicitem credenciais de acesso, pois assim é possível diferenciar usuários que apenas acessaram a página dos que realmente foram vítimas da fraude.

A etapa de montagem da infraestrutura *web* é a que envolve mais passos técnicos. Em primeiro lugar é necessário uma infraestrutura de servidor para hospedagem do *phishing*. Nos testes executados foi criada uma máquina virtual com acesso IP público na nuvem do serviço *Amazon Web Services* (AWS). A máquina utiliza sistema operacional Ubuntu Linux Server, seguindo a configuração básica disponível, classificada dentro do plano de *Free Tier*, o qual, dentro de determinados critérios de utilização não possui custos, apresentando-se então como uma plataforma acessível para criação de testes. Foi possível realizar os dois experimentos práticos deste trabalho sem ultrapassar os limites estabelecidos, portanto sem nenhuma cobrança. Nesse servidor foi então instalado o

[7]https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

[8]https://www.linkedin.com/help/linkedin/answer/a1359065/verifications-on-your-linkedin-profile

Social Engineering Toolkit (SET) para clonagem dos *websites* válidos.

Um fator importante para maior credibilidade do teste é o registro de um domínio. No primeiro teste foi utilizado um domínio *.br devidamente registrado, com grafia semelhante ao original, ao custo de 40 reais anuais, um valor acessível. Para o segundo teste utilizou-se um registro gratuito, tendo, porém, uma URL mais chamativa para identificação do *phishing*. A escolha por um domínio registrado ou gratuito pode ser equilibrada com o nível de dificuldade para identificação da fraude nos demais aspectos do teste. A não utilização de um domínio pode causar problemas com filtros *anti-SPAM* na etapa de envio dos *e-mails*. Os domínios devem ser configurados no arquivo de *VirtualHosts* do Apache no servidor, procedimento bastante simples e amplamente documentado na internet.

Através do SET é feita então a clonagem dos *websites* verdadeiros, através da ferramenta *Web Cloner*, disponível no serviço de *Credential Harvester*. É importante no momento da criação do clone o endereço de retorno ser apontado para o domínio criado anteriormente. Para garantia de que não ocorra cruzamento dos dados que pudessem comprometer a identificação dos resultados da pesquisa, utiliza-se arquivos individuais para cada alvo do teste. A Figura 3.10 apresenta a estrutura de arquivos utilizada.

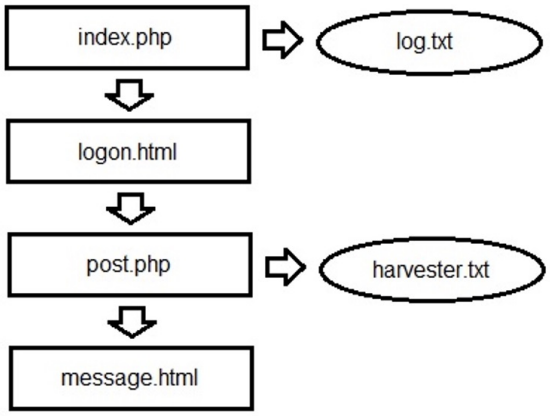

Figura 3.10: Estrutura de arquivos do *phishing*.

O primeiro arquivo, *index.php*, captura o IP do usuário, a data e a hora do acesso, o endereço de origem (*HTTP Referer*, para garantir a origem do usuário) e as informações do navegador utilizado. As informações são então salvas no arquivo *log.txt* e o usuário é automaticamente redirecionado para o arquivo *logon.html*, que é a página clonada com o auxílio do SET. A Figura 3.11 demonstra o conteúdo do arquivo *index.php*.

O arquivo *logon.html* é o *phishing* em si, a página que deve capturar as credenciais do usuário. O meio de validar que um usuário realmente foi capturado seria a coleta das credenciais, e o acesso ao sistema/serviço verdadeiro com as mesmas. Esse tipo de teste, porém, entra em conflito com boas práticas de segurança, não sendo sugerida a sua realização. O teste de Engenharia Social visa uma visão da resposta dos colaboradores à um ataque, não podendo ser um vetor de incidentes, devendo a segurança das informações do usuário ser uma prioridade. Em virtude disso, um importante fator a ser considerado é o fato do *phishing* não utilizar criptografia (HTTPS), portanto as credenciais do usuá-

```php
<?php

$arquivo = 'log.txt';
$novalinha = "\n";

$timestamp = date("d/m : H:i");
$ip = $_SERVER['REMOTE_ADDR'];
$referer = $_SERVER['HTTP_REFERER'];
$browser = $_SERVER['HTTP_USER_AGENT'];

file_put_contents($arquivo, $timestamp.$data.PHP_EOL, FILE_APPEND);
file_put_contents($arquivo, $ip.$novalinha, FILE_APPEND);
file_put_contents($arquivo, $referer.$novalinha, FILE_APPEND);
file_put_contents($arquivo, $browser.$novalinha.$novalinha, FILE_APPEND);

?>

<meta http-equiv="refresh" content="0; url=http://www.site.com/logon.html" />
```

Figura 3.11: Arquivo *index.php*.

rio que for vítima serão enviadas ao servidor sem qualquer tipo de proteção. Para evitar esse tipo de problema, uma boa prática é adicionar na própria página do *phishing* um pequeno código que apague a senha antes mesmo dela ser enviada, de forma completamente transparente para o usuário, capturando apenas o nome de usuário. Para isso é necessário abrir o código da página e identificar o ID relacionado ao campo de senha, normalmente *password*. Depois localizar a linha do formulário, a responsável pelo envio das informações, iniciada pelo código *<form name=*. Verificar se a linha possui o parâmetro *onsubmit=*, se sim, o mesmo deve ser substituído pelo código, caso não, basta adicionar o mesmo código, lembrando de substituir *id_senha* pelo ID identificado no corpo da página: *onsubmit="document.getElementById('id_senha').value='';"*. A utilização desse código elimina a senha antes dos dados serem tratados pela página, portanto garantindo a mínima segurança para as informações do usuário.

Uma vez submetidas as credenciais, o próximo arquivo chamado é o *post.php*, responsável pela captura das informações. No momento que a página é clonada através do SET é gerado automaticamente um arquivo *post.php*, porém o mesmo não captura todas as informações necessárias e redireciona o usuário vitimado para a página verdadeira, fugindo do objetivo do teste. A Figura 3.12 apresenta o código utilizado no *post.php*. Além das credenciais, ele captura as mesmas informações que o arquivo *index.php*, de forma que seja possível comparar as mesmas e garantir que é o mesmo usuário que acessou e forneceu as informações.

As informações do usuário são então enviadas para o arquivo *harvester.txt*, e o usuário é automaticamente redirecionado para o arquivo *message.html*. Esse arquivo contém uma mensagem para o usuário que falhar no teste. Recomenda-se iniciar com logo ou cabeçalho oficial da instituição. Em seguida, solicitar ao usuário que evite comentar com seus colegas a respeito do teste, evitando assim que seja gerado um alerta que comprometa os resultados. O texto deve então reiterar que a senha do usuário não foi comprometida, e apresentar os objetivos do teste, que são avaliar o nível de capacidade dos colaboradores para identificação de fraudes e, baseado nisso, montar futuras estratégias de treinamento e capacitação. Alertar o usuário que o teste não visa punir quem for vítima da fraude é uma prática extremamente recomendável, visto que o mesmo pode se caracterizar como assédio moral dentro de determinadas circunstâncias e a prática do medo de punição tem

```php
<?php

$file = 'harvester.txt';
$quebra = "\n";

$dados = $_POST;
$date = date("d/m : H:i");
$ip = $_SERVER['REMOTE_ADDR'];
$referer = $_SERVER['HTTP_REFERER'];
$navegador = $_SERVER['HTTP_USER_AGENT'];

file_put_contents($file, $dados, FILE_APPEND);
file_put_contents($file, $quebra.$date.$data.PHP_EOL, FILE_APPEND);
file_put_contents($file, $ip.$quebra, FILE_APPEND);
file_put_contents($file, $referer.$quebra, FILE_APPEND);
file_put_contents($file, $navegador.$quebra.$quebra, FILE_APPEND);
?>

<meta http-equiv="refresh" content="0; url=http://www.site.com/message.html" />
```

Figura 3.12: Arquivo *post.php*.

um efeito contrário aos objetivos buscados com essa forma de teste.

O próximo passo na mensagem é auxiliar o usuário a perceber as falhas da fraude enviada. Revelar os exemplos que possam ser observados no próprio teste criado, como *e-mail* de origem de um domínio inválido, a URL com grafia incorreta, a ausência de criptografia na página, etc. Outros exemplos relevante podem também ser adicionados. Esse é o mais importante item do teste, pois é a ferramenta que vai auxiliar os usuários a fixar os conhecimentos necessários para identificar um ataque verdadeiro. É importante também apontar ao usuário como proceder em caso de suspeita de ES, e quais canais devem ser utilizados para notificação. A mensagem pode terminar oferecendo *links* e outras informações que auxiliem os usuários a descobrir mais sobre o assunto, aprofundando o conhecimento gerado, como treinamentos, alguma política ou diretriz interna, e afins.

A fim de criar os arquivos individuais, sugere-se a utilização de alguma forma de codificação nos nomes dos arquivos, para garantir que cada usuário seja direcionado para um fluxo separado e não consiga alterar os resultados dos demais usuários. É importante lembrar que além de criar os arquivos é necessário alterar o conteúdo dos mesmos, pois os mesmos possuem *links* para o próximo arquivo. A codificação errada no conteúdo irá misturar os dados capturados e afetar os resultados do teste. É importante também conferir as permissões de acesso dos arquivos *log.txt* e *harvester.txt* para garantir que seja possível gravar as informações capturadas neles.

A próxima etapa é a criação do e-mail. Tendo um domínio, é possível criar um *e-mail* no mesmo domínio gratuitamente durante uma fase de testes através do serviço *Google Workspace*, o que foi utilizado no primeiro teste realizado. Para o segundo teste foi criado um *e-mail* simples no serviço *Gmail*, o qual não tem custos. Para a criação do corpo do *e-mail*, é ideal seguir um modelo de mensagem verdadeiro, lembrando-se de utilizar a URL correta para cada usuário. Uma boa prática é realizar alguns testes de envio do *e-mail* antes de encaminhar aos alvos, a fim de evitar que o mesmo seja identificado por filtros *anti-SPAM* e bloqueado.

Uma vez enviados os e-mails, os resultados capturados podem ser verificados conferindo o conteúdo dos arquivos *log.txt* e *harvester.txt*. Deve-se aguardar um período suficiente para que todos os usuários possam ter visto a mensagem, sendo sugerido cerca

de 48 horas. Os usuários que forem vitimados recebem o material informativo no mesmo momento que caírem na fraude, porém a informação deve ser repassada mesmo aos demais envolvidos. Quando o teste se der por encerrado, recomenda-se enviar um *e-mail* de um endereço válido da instituição para todos os participantes no escopo do teste, detalhando que no período foi realizado um teste de Engenharia Social, e apresentando as informações presentes na mensagem vista pelos usuários vitimados. É importante mais uma vez frisar que o objetivo do teste não deve ser o teste em si, mas sim o treinamento e conscientização dos usuários para que sua capacidade de reconhecer fraudes aumente.

3.6.2.1. Experimentação Prática

O primeiro teste foi realizado em três ondas diferentes, envolvendo um total de 179 usuários com variados níveis de conhecimento. O ambiente envolvia colaboradores do setor de TI de uma instituição de nível superior, sendo o teste devidamente autorizado junto aos responsáveis e gestores do setor. A proposta de *phishing* envolveu um clone da página de acesso do serviço de *email*, sendo então enviada aos mesmos uma mensagem de correio eletrônico simulando as mensagens reais do serviço enviadas quando a cota de armazenamento da conta está próxima do limite. A Figura 3.13 mostra o modelo de e-mail enviado aos usuários, sendo que o link *Log in here* direcionava o usuário para uma página individual. As informações que permitam identificar a instituição foram suprimidas por questões de privacidade.

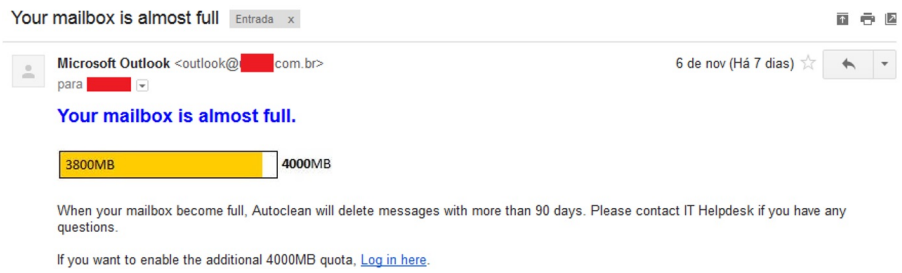

Figura 3.13: Modelo de *e-mail* utilizado no primeiro teste.

A primeira onda foi enviada para um grupo pré-selecionado de 05 usuários, todos de nível técnico mais avançado. O objetivo era ter uma ideia inicial do teste, e obter sugestões desses usuários sobre a qualidade do teste. A segunda onda por sua vez ocorreu para um primeiro grupo de 13 usuários, envolvendo diretores, responsáveis de área, líderes de equipe e outros cargos de gestão. Esse teste serviu como validação para que a direção da instituição autorizasse a última onda de testes. A terceira e última onda então foi enviada para os demais 161 usuários do setor, sendo a grande maioria destes usuários regulares com nível de conhecimento básico a respeito de *phishing*.

O segundo teste foi realizado em uma única onda, envolvendo um total de 47 usuários, sendo a grande maioria pessoal técnico. O ambiente envolvia colaboradores dos setores de TI, comercial e administrativo de uma empresa privada da área de tecnologia, sendo o teste devidamente autorizado junto aos responsáveis e gestores do setor.

A proposta de *phishing* envolveu um clone da página de serviço de calendário da companhia, utilizado para marcação de reuniões e compromissos, hospedado na plataforma *WebCalendar*, sendo então enviada aos colaboradores uma mensagem de correio eletrônico simulando as mensagens reais de agendamento de reunião enviadas pelo serviço.

A Figura 3.14 apresenta o modelo de *e-mail* enviado aos colaboradores. Como o modelo original apresentava uma URL completa, quando se configurava a mesma para direcionar para outra URL, porém mostrando a original, os filtros *anti-SPAM* identificavam a mensagem como fraude. O recurso utilizado foi substituir a URL no corpo do *e-mail* por uma imagem da mesma, com o link para o *phishing* associado à imagem, sendo assim possível burlar os filtros.

Figura 3.14: Modelo de *e-mail* utilizado no segundo teste.

3.6.2.2. Análise dos Resultados

O primeiro teste realizado envolveu 179 usuários, sendo destes cinco usuários técnicos e capacitados, 13 usuários gerenciais e 161 usuários com treinamento básico. Do total do grupo, apenas 12 acessaram a página do *phishing*, quatro usuários técnicos, três gerenciais e cinco usuários padrão. Nenhum usuário foi vitimado, sendo que sete entraram com dados apenas de teste. Também foi verificado que 13 usuários reportaram a tentativa de fraude ao setor responsável. Depois do evento, através de uma amostra de entrevistas individuais, houve a confirmação de que o teste de fato ampliou o nível de preocupação dos usuários, que reportaram estar mais atentos a partir daquele momento.

O segundo teste envolveu 47 usuários, sendo destes 42 usuários técnicos e capacitados, e cinco usuários padrão, porém com treinamentos na área de segurança. Do total do grupo, 14 acessaram a página do *phishing*, sendo todos com nível técnico de conhecimento. Desses, dez foram vitimados pelo teste, incluindo dois líderes de equipe técnica. Nenhum usuário utilizou os canais formais da empresa para notificar o incidente, mesmo os que perceberam a fraude. A equipe de segurança da informação elogiou a montagem e análise dos resultados do teste, sendo que os testes passarão a ser realizados periodicamente como parte das campanhas de segurança da empresa da companhia, e o teste e treinamentos específicos serão adicionados aos treinamentos periódicos já realizados com

as equipes.

Uma situação observada nos testes chamou a atenção durante a análise dos resultados. Um usuário que foi vítima da fraude cerca de cinco minutos após ter submetido suas credenciais submeteu em sequência os nomes de usuário verdadeiros de outros sete colaboradores, inclusive de gestores e executivos da empresa, possivelmente temendo os resultados do teste. Essa atitude traz à tona a importância de conscientizar os usuários quanto aos objetivos do teste, que não envolvem a punição, e sim o treinamento. Também foi assim validada a importância da utilização de arquivos individuais para cada alvo do teste, sendo possível diferenciar a identidade e informações fornecidas por cada usuário, com menores riscos de cruzamento de informações ou tentativas de manipular os resultados do teste por parte dos usuários.

Após a realização de dois testes em ambientes reais foi possível verificar que a estrutura proposta para a criação do *phishing* e *e-mail* atende aos requisitos de ser acessível, de baixo custo e simples de ser reproduzida, atingindo portanto os objetivos esperados. A montagem da estrutura não apresentou dificuldades, e o retorno recebido dos envolvidos foi positivo.

Uma das propostas iniciais foi a modificação do código do SET para automatizar a montagem da estrutura de arquivos. Após longa análise do código e diversas tentativas sem sucesso foi concluído que a abordagem não atenderia aos objetivos do trabalho. Optou-se então por utilizar apenas o mecanismo de clonagem de *websites*, o qual funcionou com sucesso em todos os testes realizados. Outra situação identificada neste trabalho foi um problema na captura credenciais quando o *phishing* padrão criado pelo SET era utilizado. Em alguns *websites* testados, ao submeter as credenciais, as mesmas não eram capturadas. Além de não fornecer todas as informações da vítima necessárias, em alguns casos o arquivo preparado pelo SET nem mesmo capturava as credenciais, devendo portanto sua utilização em configuração padrão ser feita com cautela, sendo sugerida a modificação manual dos arquivos.

Um possível problema verificado no primeiro teste, onde nenhum usuário foi vítima do teste, é o correto planejamento da montagem da página e do *e-mail*. No caso, foi opção de comum acordo com os responsáveis o envio das mensagens falsas em inglês, enquanto o sistema de *e-mail* da instituição estava configurado em português. Não há como garantir que esse tenha sido o real motivo da ausência de vítimas, porém a utilização de fraudes excessivamente fáceis de serem percebidas irão prejudicar os resultados esperados no trabalho.

A configuração da captura das informações dos usuários que tanto apenas acessavam a página quanto dos que submetiam as credenciais em arquivos individuais permitiu a correta identificação dos usuários, sem cruzamento de informações. Foi possível também diferenciar o foco de futuros treinamentos a partir das situações onde a maioria dos usuários caiu na fraude, o que seria o pior cenário, e os usuários que apenas acessaram a URL, situação também não recomendável em muitas situações.

O teste também mostrou a importância de campanhas que incentivem os usuários a reportarem mensagens suspeitas. A ocorrência de um ataque direcionado à companhia ou instituição deve gerar um alerta imediato nas equipes e acionamento dos procedimentos

de resposta a incidentes.

3.7. Ética e Engenharia Social

Em segurança cibernética, o conceito de *hacking* ético refere-se ao uso de conhecimentos, técnicas, ferramentas e outros recursos para identificar, analisar e apoiar a mitigação de falhas e vulnerabilidades utilizadas por atacantes [Hatfield 2019] [Peake 2003]. Tem como objetivo proteger contra ataques e aprimorar a segurança [NIST 2019].

No *hacking* ético a execução das ações para explorar falhas e vulnerabilidades ocorre após consentimento [Hatfield 2019] e respeita diretrizes estabelecidas e aprovadas previamente [Peake 2003]. Adicionalmente, o processo completo é registrado, detalhando os procedimentos adotados, para que possam ser reproduzidos caso seja necessário [Peake 2003].

A utilização de ES também é aplicada por *hackers* éticos, entretanto, por operar nos fatores humanos da segurança, na tentativa de manipular indivíduos através de outros indivíduos ou de maneira automatizada, levanta questões éticas importantes [Hatfield 2019]. Apesar disso, observa-se a ausência de qualquer formalização, ainda que mínima, relacionada ao impacto ético de um ataque de ES em contextos não maliciosos [Mouton et al. 2015], como nas pesquisas ou na prática de *hacking* ético.

Na pesquisa e na prática em segurança da informação, a ética no campo da ES é pouco estudada. Ainda que aspectos éticos relacionados a execução de testes de invasão e técnicas de *hacking* sejam abordados, discussões específicas acerca da ética da ES e ESA não são facilmente identificadas [Hatfield 2019]. Segundo [Hatfield 2019], além de serem escassas pesquisas que se dedicam explicitamente à ética da ES, estas apresentam limitações relacionadas a abordagem do tema, que tende utilizar abordagens teóricas reduzidas.

Dentre os fatores que podem contribuir para a pouca produção no tema, merece destaque as complexidades ligadas à ética coletiva e individualizada e suas respectivas restrições [Mouton et al. 2015]. E a necessidade de abordagens plurais e de um diálogo interdisciplinar, no qual pesquisadores e especialistas em psicologia, ética e segurança da informação possam colaborar de maneira conjunta para o debate.

Dado o cenário de escassa produção acadêmica sobre o tema, bem como a ausência de formalizações, esta seção não tem objetivo apresentar um consenso sobre a ética e a ES em contextos não maliciosos. No entanto, busca colaborar para abertura do debate acerca da ética em ES, apresentando possíveis contribuições à discussão, a partir da ética em pesquisa com seres humanos, da ética em pesquisas na internet e das orientações e códigos de ética profissionais do campo da tecnologia e da internet. Áreas nas quais já há conceituações, formalizações e cujo debate já está amplamente difundindo.

Ainda que as técnicas de ES possam ser aplicadas fora do contexto de mediação tecnológica, é consenso que a ES costuma utilizar predominantemente tecnologias baseadas em internet. Principalmente pelo poder amplificador destas. Desta forma, justifica-se a inclusão no debate sobre ética, das considerações sobre a condução de pesquisas éticas baseadas em internet. Além disso, a investigação baseada em internet deixou de ser uma metodologia e passou a uma prática quase onipresente, exigindo consideração cui-

dadosa no que diz respeito aos conceitos tradicionais da pesquisa com seres humanos [Buchanan and Zimmer 2021][.

Em pesquisas que relacionam-se com tecnologia e internet, a ética compreende questões relacionadas a consentimento dos participantes, a privacidade, confidencialidade e integridade de dados, ao anonimato, à propriedade intelectual, a aspectos coletivos e códigos e condutas profissionais [Buchanan and Zimmer 2021]. É definida como sendo a análise de questões éticas e a aplicação de princípios de ética em pesquisa no que se refere à pesquisas conduzidas ou mediadas pela internet [Buchanan and Zimmer 2021]. Abrange, entre outras, quaisquer pesquisas e estudos que coletem dados ou realizem a análise de atividades a partir ou em ambientes *online*, bem como pesquisas relacionadas com os efeitos da mediação pela internet em comportamentos de indivíduos [Buchanan and Zimmer 2021].

Independente do cenário de pesquisa incluir intervenção tecnológica ou mediada, qualquer investigação deve considerar e ser orientada por princípios éticos fundamentais [Buchanan and Zimmer 2021]. Na pesquisa com seres humanos, há princípios bem estabelecidos. Uma pesquisa ética é aquela que respeita o participante, levando em conta a sua vulnerabilidade e ponderando diligentemente sobre os riscos e benefícios, tanto os estabelecidos quanto os possíveis. Considerando impactos individuais e coletivos. Além de comprometer-se em não causar dano, em maximizar possíveis benefícios e em minimizar possíveis danos e prejuízos. Essas características são resumidas pelos princípios da autonomia, beneficência, não maleficência e justiça.

No que se refere a ética profissional, códigos como o IEEE Código de ética [IEEE 2020] apontam , entre outras orientações: (a) priorizar a segurança, a saúde e o bem-estar dos indivíduos, esforçar-se para cumprir o design ético e práticas de desenvolvimento sustentável, proteger a privacidade e divulgar fatores que possam gerar riscos à indivíduos ao ao meio ambiente; (b) melhorar a compreensão, dos indivíduos e da sociedade, sobre as capacidades e as implicações sociais das tecnologias convencionais e emergentes, incluindo os sistemas inteligentes; (c) evitar condutas ilícitas nas atividades profissionais e rejeitar o suborno em todas as suas formas; (d) reconhecer e corrigir erros, ser honesto e realista ao declarar afirmações ou estimativas baseadas em dados disponíveis; (e) evitar prejudicar terceiros, a sua propriedade, reputação ou emprego através de ações falsas ou maliciosas, rumores ou quaisquer outros abusos verbais ou físicos; (f)tratar todas as pessoas de forma justa e respeitosa e não praticar discriminação; e (g) esforçar-se para garantir que o código seja respeitado[IEEE 2020].

A partir de tais ponderações e de alguns pontos de convergência, podemos considerar que as práticas de ES em contextos não maliciosos devem analisar e avaliar questões éticas relacionadas principalmente: a informação e consentimento dos indivíduos, à privacidade, á confidencialidade e à integridade de dados, à anonimização de dados, a minimização de riscos e prejuízos aos indivíduos e a não geração de danos. Adicionalmente, códigos de ética e condutas profissionais específicas do contexto devem ser seguidos.

No âmbito da informação e do consentimento, dada as características de um teste de invasão e do *hacking* ético, um forma de garantir que as pessoas estejam cientes da possibilidade de participarem de ações que poderão submetê-las a práticas de ES, são comunicados coletivos e gerais. Informando que eventualmente podem ser realizados

pela organização testes ou rotinas para checar a eficácia dos controles de segurança da informação implementados. Sem individualizar os testes e sem especificar o momento no qual irá ocorrer, o que poderia gerar viés nos resultados.

Essa já é uma prática comum em organizações com áreas de segurança da informação estabelecidas e com rotinas para testar a probabilidade de seus colaboradores caírem em ataques de *phishing*. Nestes comunicados, devem estar explícitos que os indivíduos não sofrerão sanções de qualquer natureza baseado nos resultados ou nas informações coletadas. Além de informar que poderão ocorrer ações do gênero, devem ficar claros também os objetivos da ação. Sugere-se ainda que os participantes sejam informados acerca da natureza dos dados coletados, bem como do compromisso e das medidas adotadas para garantir privacidade.

No que tange a privacidade e confidencialidade em ES de caráter não malicioso, os dados que permitem identificar participantes, podendo gerar constrangimento, exposição ou qualquer tipo de prejuízo podem e devem ser considerados de acesso restrito. No caso de organizações, restritos a equipe interna ou contratada para execução dos testes e no caso de pesquisas acadêmicas, restritos aos pesquisadores envolvidos.

Os resultados devem ser tratados de maneira coletiva, fazendo referência a organização inteira, a setores ou áreas de uma organização. E sempre que possível, sugere-se a anonimização dos dados ou descarte de dados que possam identificar pessoas. Em pesquisas acadêmicas, os dados devem ter sua utilização restrita a pesquisa e as produções científicas relacionadas. Quando não necessária a identificação dos participantes não recomenda-se a coleta de dados que a permitam. Nos casos em que se faz necessária, estes devem ser anonimizados. A guarda dos dados deve obedecer as orientações e regulamentações dos comitês de ética em pesquisa e das legislações vigentes.

Uma abordagem coletiva e ampla, reduz a probabilidade de individualizar a responsabilidade e de gerar sobrecarga. Isso, por sua vez, reduz a autorresponsabilização e os efeitos negativos que a tomada de consciência sobre a própria susceptibilidade a um ataque de ES poderia gerar em uma pessoa ou a um pequeno grupo. Minimizando possíveis prejuízos aos quais a pessoa estaria vulnerável.

É importante lembrar que qualquer pessoa é susceptível a ataques de ES. Assim, a responsabilização de indivíduos, para além de promover mal-estar e condutas antiéticas, não tem potencial para promover uma cultura de segurança e não se relaciona com práticas efetivas em segurança da informação.

Em práticas éticas, seja qual for sua natureza, debater questões relativas aos objetivos, à eficácia e os benefícios de determinadas ações é imprescindível. Não seria diferente na execução de ES em cenários não maliciosos.

O objetivo de qualquer ação ou pesquisa em ES é garantir e promover segurança da informação, protegendo assim usuários e organizações. Dessa forma, tais práticas não podem perder de vista tal objetivo. O fator humano e suas vulnerabilidades são explorados, entretanto, o foco das práticas são os controles implementados ou necessários e as possíveis melhorias nas tecnologias.

O debate é extenso e merece atenção e produções futuras que incluam novas defi-

nições e consensos a partir de regulamentações e de orientações às pesquisas que utilizam tecnologias. Suas particularidades também podem exigir novas orientações regulatórias e profissionais. Por fim, a exigência do debate sobre ética em cibersegurança torna mais evidente o quanto segurança da informação se inscreve para além do campo tecnológico, abrangendo campos sociais e comportamentais.

3.8. Considerações Finais

Neste capítulo, exploramos a interseção entre a ES, a Psicologia Cognitiva e a ESA. O conteúdo apresentado busca entender os impactos da ES no comportamento humano e suas vulnerabilidades psicológicas para obter informações pessoais e corporativas, bem como formas de persuasão dos usuários para realizarem ações indesejadas.

É indiscutivelmente desafiador abordar de maneira completa todos os ângulos da ciência cognitiva, da psicologia e de suas intersecções com a segurança da informação em apenas um capítulo, aprofundar-se nesses campos certamente proporciona uma perspectiva mais abrangente sobre os ataques de ES. Ao invés de buscar uma compreensão exaustivamente técnica, permite direcionar esforços para compreender e, consequentemente, proteger vulnerabilidades inerentes à condição humana.

A relevância dos fatores humanos, já intrínseca no pilar pessoas da segurança da informação, ganha proeminência no contexto da ES. A compreensão da interação entre seres humanos e tecnologia, juntamente com a análise dos elementos psicológicos envolvidos, também deve ser incorporada no âmbito da segurança da informação.

Com o estudo realizado emerge o impacto da Psicologia relacionadas com as ações de ES. A manipulação de vieses cognitivos e a compreensão das características humanas envolvidas no processo de tomada de decisão têm sido a base para os ataques de ES. Com a análise realizada fica evidente que as estratégias de manipulação dos usuários no contexto da ES, estão ganhando escala no mundo digital com a ESA.

A ESA utiliza *Bots* e técnicas automatizadas no ecossistema digital, impulsionando exponencialmente a capacidade de ataques em larga escala. Essa transformação reforça a necessidade de abordar a segurança não apenas com medidas tecnológicas. A segurança deve considerar os aspectos psicológicos envolvidos.

À medida que as redes sociais e outras plataformas digitais continuam a moldar nossa interação *online* e *offline*, se faz necessário reconhecer os desafios e riscos associados à ESA. A confiança nas interações cibernéticas pode ser utilizada por atores maliciosos, comprometendo a segurança e a privacidade dos usuários.

Por fim, este capítulo destaca a importância de uma abordagem multidisciplinar para enfrentar os desafios da ES. A colaboração entre especialistas em Psicologia e SI é fundamental para desenvolver estratégias abrangentes de defesa contra os ataques de ES e ESA. Essa integração busca implementar o uso ético da automação para interação das ferramentas automatizadas com os usuários.

Referências

[APA nd] (n.d.). Apa dictionary of psychology - american psychological association.

[Al-Charchafchi et al. 2019] Al-Charchafchi, A., Manickam, S., and Alqattan, Z. N. (2019). Threats against information privacy and security in social networks: A review. In *International Conference on Advances in Cyber Security*, pages 358–372. Springer.

[Ancis 2020] Ancis, J. R. (2020). The Age of Cyberpsychology: An Overview. *Technology, Mind, and Behavior*, 1(1). https://tmb.apaopen.org/pub/2yn6jhyv.

[Attrill-Smith et al. 2019a] Attrill-Smith, A., Fullwood, C., Keep, M., and Kuss, D. J. (2019a). The online self. In *The Oxford Handbook of Cyberpsychology*, pages 17–34. Oxford University Press.

[Attrill-Smith et al. 2019b] Attrill-Smith, A., Fullwood, C., Keep, M., and Kuss, D. J. (2019b). *The Oxford Handbook of Cyberpsychology*. Oxford University Press.

[Bayer et al. 2020] Bayer, J. B., Triêu, P., and Ellison, N. B. (2020). Social media elements, ecologies, and effects. *Annual Review of Psychology*, 71(1):471–497.

[Beal 2005] Beal, A. (2005). Segurança da informação: Princípios e melhores práticas para a proteção dos ativos de informação nas organizações. *Atlas*.

[Benias and Markopoulos 2017] Benias, N. and Markopoulos, A. P. (2017). A review on the readiness level and cyber-security challenges in industry 4.0. In *2017 South Eastern European Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–5. IEEE.

[Brasil 1940] Brasil (1940). Lei nº 2.848, de 7 de dezembro de 1940. *Diário Oficial [da] República Federativa do Brasil*.

[Buchanan and Zimmer 2021] Buchanan, E. A. and Zimmer, M. (2021). Internet research ethics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2021 edition.

[Camisani-Calzolari 2012] Camisani-Calzolari, M. (2012). Analysis of twitter followers of the us presidential election candidates: Barack obama and mitt romney. *Online). http://digitalevaluations. com*.

[Carr and Hayes 2015] Carr, C. T. and Hayes, R. A. (2015). Social media: defining, developing, and divining. *Atlantic Journal of Communication*, 23(1):46–65.

[Castells 2002] Castells, M. (2002). *A sociedade em rede*. Editora Paz e Terra.

[Collier et al. 2023] Collier, H., Morton, C., Alharthi, D., and Kleiner, J. (2023). Cultural influences on information security. In *European Conference on Cyber Warfare and Security*, volume 22, pages 143–150.

[Crossler and Bélanger 2014] Crossler, R. and Bélanger, F. (2014). An extended perspective on individual security behaviors. *ACM SIGMIS Database*, 45(4):51–71.

[Culot et al. 2019] Culot, G., Fattori, F., Podrecca, M., and Sartor, M. (2019). Addressing industry 4.0 cybersecurity challenges. *IEEE Engineering Management Review*, 47(3):79–86.

[Darwish et al. 2012] Darwish, A., Zarka, A. E., and Aloul, F. (2012). Towards understanding phishing victims' profile. In *2012 International Conference on Computer Systems and Industrial Informatics*, pages 1–5.

[de Souza Pereira et al. 2022] de Souza Pereira, L. A., Vicentine, A. L., and Rizo, A. C. (2022). Impactos da engenharia social na segurança da informação. *Revista Brasileira em Tecnologia da Informação*, 4(1):48–58.

[Dewangan and Kaushal 2016] Dewangan, M. and Kaushal, R. (2016). Socialbot: Behavioral analysis and detection. In *International Symposium on Security in Computing and Communication*, pages 450–460. Springer.

[Dickerson et al. 2014] Dickerson, J. P., Kagan, V., and Subrahmanian, V. (2014). Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 620–627.

[Ferrara et al. 2016] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

[Freitas et al. 2015] Freitas, C., Benevenuto, F., Ghosh, S., and Veloso, A. (2015). Reverse engineering socialbot infiltration strategies in twitter. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 25–32. IEEE.

[Freitas et al. 2014] Freitas, C., Benevenuto, F., and Veloso, A. (2014). Socialbots: Implicações na segurança e na credibilidade de serviços baseados no twitter. *SBRC, Santa Catarina, Brasil*, pages 603–616.

[Gohn 2014] Gohn, M. d. G. M. (2014). *Sociologia dos movimentos sociais*. Number 47 in Questões da nossa época Sociologia. Cortez Editora, São Paulo, 2. ed edition.

[Greitzer et al. 2019] Greitzer, F. L., Purl, J., Leong, Y. M., and Sticha, P. J. (2019). Positioning your organization to respond to insider threats. *IEEE Engineering Management Review*, 47(2):75–83.

[Grimme et al. 2017] Grimme, C., Preuss, M., Adam, L., and Trautmann, H. (2017). Social bots: Human-like by means of human control? *Big data*, 5(4):279–293.

[Guzman and Lewis 2020] Guzman, A. L. and Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1):70–86.

[Hatfield 2019] Hatfield, J. M. (2019). Virtuous human hacking: The ethics of social engineering in penetration-testing. *Computers & Security*, 83:354–366.

[Huber et al. 2009] Huber, M., Kowalski, S., Nohlberg, M., and Tjoa, S. (2009). Towards automating social engineering using social networking sites. In *2009 International Conference on Computational Science and Engineering*, volume 3, pages 117–124. IEEE.

[IEEE 2020] IEEE (2020). *Code of Ethics*. IEEE - The Institute of Electrical and Electronics Engineers, Inc.

[Khan and Das 2018] Khan, R. and Das, A. (2018). Build better chatbots. *A complete guide to getting started with chatbots*.

[Klimburg-Witjes and Wentland 2021] Klimburg-Witjes, N. and Wentland, A. (2021). Hacking humans? social engineering and the construction of the "deficient user" in cybersecurity discourses. *Science, Technology, & Human Values*, 46(6):1316–1339.

[Korteling and Toet 2022] Korteling, J. and Toet, A. (2022). Cognitive Biases. In *Encyclopedia of Behavioral Neuroscience, 2nd edition*, pages 610–619. Elsevier.

[Leary and Tangney 2014] Leary, M. and Tangney, J. P., editors (2014). *Handbook of self and identity*. Guilford Press, New York London, second edition edition.

[Leary 2019] Leary, M. R. (2019). *Self-Presentation: Impression Management and Interpersonal Behavior*. Routledge, 1 edition.

[Libicki 2018] Libicki, M. (2018). Could the issue of dprk hacking benefit from benign neglect? *Georgetown Journal of International Affairs*, 19:83–89.

[Martineau et al. 2023] Martineau, M., Spiridon, E., and Aiken, M. (2023). A comprehensive framework for cyber behavioral analysis based on a systematic review of cyber profiling literature. *Forensic Sciences*, 3(3):452–477.

[Messias et al. 2018] Messias, J., Benevenuto, F., and Oliveira, R. (2018). Bots sociais: Como robôs podem se tornar pessoas influentes no twitter? *Revista Eletrônica de Iniciação Científica em Computação*, 16(1).

[Mitnick and Simon 2003] Mitnick, K. D. and Simon, W. L. (2003). *The art of deception: Controlling the human element of security*. John Wiley & Sons.

[Montañez et al. 2020] Montañez, R., Golob, E., and Xu, S. (2020). Human cognition through the lens of social engineering cyberattacks. *Frontiers in Psychology*, 11.

[Mouton et al. 2015] Mouton, F., Malan, M. M., K., K. K., and Venter (2015). Necessity for ethics in social engineering research. *Computers Security*, 55:114–127.

[NIST 2019] NIST (2019). Glossary. *National Institute of Standards and Technology*.

[Nobles 2023] Nobles, C. (2023). Human factors in cybersecurity: academia's missed opportunity. *MWAIS 2023 Proceedings*.

[Peake 2003] Peake, C. (2003). Red teaming: the art of ethical hacking | sans institute.

[Pinheiro 2020] Pinheiro, P. P. (2020). Segurança digital: Proteção de dados nas empresas. *1ª edição. São Paulo, SP: Grupo GEN*.

[Piovesan et al. 2019] Piovesan, L. G., Silva, E. R. C., de Sousa, J. F., and Turibus, S. N. (2019). Engenharia social: Uma abordagem sobre phishing. *REVISTA CIENTÍFICA DA FACULDADE DE BALSAS*, 10(1):45–59.

[Reep-van den Bergh and Junger 2018] Reep-van den Bergh, C. M. and Junger, M. (2018). Victims of cybercrime in europe: a review of victim surveys. *Crime science*, 7(1):1–15.

[Riecken 1974] Riecken, H. W. (1974). Obedience to authority. an experimental view. stanley milgram. harper and row, new york, 1974. xx, 224 pp., illus. 10. *Science*, 184(4137):667–669.

[Robinson 2023] Robinson, N. (2023). Human factors security engineering: The future of cybersecurity teams. *EDPACS*, 67(5):1–17.

[Rogers et al. 2018] Rogers, T., Goldstein, N. J., and Fox, C. R. (2018). Social Mobilization. *Annual Review of Psychology*, 69(1):357–381.

[Rouse 2013] Rouse, M. (2013). What is socialbot? *WhatIs.com*.

[Ryan and Deci 2000] Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.

[Salahdine and Kaabouch 2019] Salahdine, F. and Kaabouch, N. (2019). Social engineering attacks: a survey. *Future Internet*, 11(4):89.

[Shaabany and Anderl 2018] Shaabany, G. and Anderl, R. (2018). Security by design as an approach to design a secure industry 4.0-capable machine enabling online-trading of technology data. In *2018 International Conference on System Science and Engineering (ICSSE)*, pages 1–5. IEEE.

[Shafahi et al. 2016] Shafahi, M., Kempers, L., and Afsarmanesh, H. (2016). Phishing through social bots on twitter. In *2016 IEEE International Conference on Big Data*, pages 3703–3712. IEEE.

[Shires 2018] Shires, J. (2018). Enacting expertise: Ritual and risk in cybersecurity. *Politics and Governance*, 6(2):31–40.

[Solano and Rocha 2019] Solano, E. and Rocha, C., editors (2019). *As direitas nas redes e nas ruas: a crise politica no Brasil*. Expressao Popular, Sao Paulo, 1a edicao edition. OCLC: on1126542066.

[Stoeckli et al. 2018] Stoeckli, E., Uebernickel, F., and Brenner, W. (2018). Exploring affordances of slack integrations and their actualization within enterprises-towards an understanding of how chatbots create value. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[Tversky and Kahneman 1974] Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131.

[Wilke and Mata 2012] Wilke, A. and Mata, R. (2012). Cognitive Bias. In *Encyclopedia of Human Behavior*, pages 531–535. Elsevier.

[Zimmermann and Renaud 2019] Zimmermann, V. and Renaud, K. (2019). Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies*, 131:169–187. 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies.